

Learning from High-Dimensional Data using Local Descriptive Models

Rui Miguel Carrasqueiro Henriques

Supervisor: Doctor Sara Alexandre Cordeiro Madeira

Thesis approved in public session to obtain the PhD Degree in
Information Systems and Computer Engineering

Jury final classification: **Pass with Distinction and Honor**

Jury

Chairperson: Chairman of the Scientific Board

Members of the Committee: Doctor Mário Alexandre Teles de Figueiredo
Doctor Joaquín Dopazo Blásquez
Doctor Miguel Francisco de Almeida Pereira da Rocha
Doctor Francisco João Duarte Cordeiro Correia dos Santos
Doctor Sara Alexandre Cordeiro Madeira

Abstract

Models learned from high-dimensional data, where the high number of features usually exceeds the number of observations, have higher propensity to either overfit or underfit data. In this context, it is thus important to focus the learning on regions of interest, such as subsets of features, guaranteeing that these regions are both informative and statistically significant. Although a composition of relevant regions can be learned under specific assumptions to offer these guarantees, the state-of-the-art learning methods place restrictive constraints on the allowed structure, coherency and quality of regions. This has prevented the understanding of how the properties of the selected regions affect the performance of descriptive and classification methods in both tabular and structured data contexts.

In this work, we propose robust, flexible and statistically significant local descriptive models and study their relevance to improve (associative) classification in high-dimensional data contexts. This task is tackled in three major steps. *First*, we propose new local descriptive models from tabular and structured data with robustness and flexibility guarantees. In the presence of matrices and network data, the focus is placed on learning biclustering models able to tackle existing challenges: learn from regions with flexible coherency (additive, symmetric, plaid and order-preserving models); guarantee scalable searches; robustness to varying forms and degree of noise; model regions from sparse data; and effectively incorporate background knowledge. In the presence of structured data, possibly given by multivariate time series or multi-sets of events, the focus is placed on new deterministic and generative methods to learn local descriptive models given by cascades of modules or arrangements of informative events. *Second*, we propose principles to both assess and guarantee the statistical significance of these descriptive models. *Third*, the previous contributions are extended towards labeled data contexts, and new training and testing functions are proposed to learn associative classification models. In this context, we assess the impact of varying structures, coherencies and quality of local descriptive models on the performance of classifiers, and combine statistical significance and accuracy views to study and revise their behavior. Finally, we extend these contributions for data with structured classes to adequately answer predictive tasks.

The proposed contributions were applied to tackle a wide-set of real-world tasks in biomedical and social domains, including the learning of descriptive and predictive models from gene expression data, repositories of health records, clinical data, collaborative filtering data, and (biological and social) networks.

Keywords:

High-Dimensional Data
Structured Data
Biclustering
Local Descriptive Models
Associative Classification
Biomedical Data Analysis
Statistical Significance
Multivariate Time Series
Multi-Sets of Events
Sparse Data

Resumo

A aprendizagem de modelos a partir de dados com elevada dimensionalidade, onde o número de atributos pode exceder o número de observações, é propensa aos riscos de sobre- e sub-ajustamento. Neste contexto, é importante focar a aprendizagem em regiões de interesse, como subconjuntos de atributos, garantindo que estas regiões são informativas e estatisticamente significativas. No entanto, o estado-da-arte em aprendizagem coloca estritas restrições na estrutura, coerência e qualidade destas regiões. Isto previne a compreensão de como as propriedades das regiões seleccionadas afectam a performance de métodos para a descrição e classificação de dados tabulares e estruturados.

Para responder a este problema, este trabalho propõe modelos descritivos locais com garantias de robustez, flexibilidade e significância estatística, e estuda a sua relevância para melhorar a classificação em dados de elevada dimensionalidade. Este objectivo é endereçado em três passos. Primeiro, novos modelos descritivos locais – flexíveis e robustos – são propostos. Na presença de dados tabulares e redes, o foco é colocado na aprendizagem de modelos de biclustering capazes de endereçar os desafios existentes: aprender regiões com coerências não-triviais (modelos aditivos, simétricos, sobrepostos e baseados em ordenações); promover a escalabilidade das procuras; garantir a robustez a diferentes formas e graus de ruído; modelar regiões a partir de dados esparsos; e incorporar conhecimento disponível. Na presença de dados estruturados, possivelmente caracterizados por séries temporais multivariadas ou multi-conjuntos de eventos, o foco é colocado na definição de métodos determinísticos e estocásticos para a aprendizagem de modelos descritivos locais associados a cascatas de módulos ou arranjos de eventos. Segundo, novos princípios são propostos para avaliar e garantir a significância estatística destes modelos descritivos. Terceiro, as contribuições anteriores são alargadas a dados anotados, e novas funções de treino e teste são propostas para a aprendizagem de modelos de classificação. Neste contexto, este trabalho mede o impacto do uso de modelos descritivos locais (com variável estrutura, coerência e qualidade) na performance dos classificadores, e estuda e revê o seu comportamento de acordo com critérios de significância estatística. Por fim, estas contribuições são estendidas a dados com classes estruturadas para responder a problemas de previsão.

As contribuições propostas foram aplicadas num conjunto alargado de problemas reais em domínios biomédicos e sociais, incluindo a aprendizagem de modelos descritivos e classificadores em dados de expressão genética, repositórios de eventos clínicos, dados colaborativos, e redes sociais e biológicas.

Palavras Chave:

Dados com Elevada Dimensionalidade

Dados Estruturados

Biclustering

Modelos Descritivos Locais

Classificação Associativa

Análise de Dados Biomédicos

Significância Estatística

Séries Temporais Multivariadas

Multi-Conjuntos de Eventos

Dados Esparsos

Reconhecimentos

O presente documento é o resultado de um longo trabalho realizado em circunstâncias únicas, dentro do qual me revejo como humilde facilitador. Ele é primariamente o resultado das pessoas que o inspiraram, das mentes científicas que o precederam e nele cooperaram, e dos seres que comigo partilharam este percurso.

Antes demais, o meu profundo agradecimento aos meus pais, Rui e Elsa. Ao seu incessável apoio, dedicação e contínuo respeito pela forma como conduzo a minha vida. A vocês, o meu Amor.

A minha profunda gratidão a AJ Miller, Mary Luck, Inelia Benz, Almine, Manuela Melo, Luís Morgado e Leslie Temple Thurston pela forma como tocaram e transformaram a minha vida. À minha musa, Maria Flávia de Mon-saraz, por me ter revelado a Ordem da Vida. Aos guias e seres que acompanharam o meu percurso, contribuindo também para este resultado.

O meu agradecimento ao primeiro responsável por este trabalho. Sara Madeira. Obrigado. A sua integridade, confiança e compromisso para uma comunicação aberta constituíram, sem dúvida, o pilar do presente trabalho. A sua humanidade e genuína atenção foram o maior segredo para a condução deste trabalho, preenchendo as minhas madrugadas de trabalho com ânimo. Mais, a sua aguda visão científica foi essencial para o desenvolvimento dos conteúdos. Em 2012, os apontamentos gráficos e coloridos do seu moleskine azulão preencheram o meu mundo por breves dias, dando origem (quatro anos mais tarde) à actual tese.

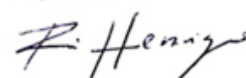
Agradeço às pessoas que contribuíram para a aprimoração desta tese. Aos membros do júri por toda a sua pronta atenção e tempo dedicados, e inúmeras partilhas em prol da qualidade do presente trabalho e da minha formação pessoal. O meu sincero agradecimento a Cláudia Antunes por ter despoletado em mim a genuína paixão pela aprendizagem automática e pelo seu papel nas contribuições dos capítulos IV-3 e VI-7. A sua presença durante os meus primeiros anos de investigação marcou positivamente o meu percurso. Ainda, o meu agradecimento a todos os revisores científicos dos artigos decorrentes deste trabalho. Ao Francisco Ferreira pelo seu papel na produção das interfaces gráficas do software decorrente desta tese. Aos meus colegas de equipa, Telma Pereira, Sofia Teixeira e Rita Levy, pela sua presença radiosa nos meus dois últimos anos de trabalho.

Agradeço à Fundação para a Ciência e Tecnologia e aos cidadãos portugueses o financiamento do meu doutoramento, através da bolsa SFRH/BD/75924/2011, que possibilitou a realização harmoniosa desta tese. Acredito que as contribuições decorrentes deste trabalho demonstram o meu empenho para avançar o estado da ciência computacional em Portugal e não só. Quero ainda agradecer às instituições de acolhimento, Inesc-ID, e de atribuição do grau, IST (Universidade de Lisboa), pela plataforma de suporte conduçiva à realização dos trabalhos.

A minha gratidão a Marta Oliveira, Pedro Gonçalves, Pedro Cosme, Gonçalo Ferreira, Sílvia Pina, Elsa Torres, Mikel Damon Miller, Alexandre Camões, João Belchior, Samantha Nogueira, Sylvia Alves, Teresa Castanheira, Pedro Policarpo, Maria Júdice, César Moniz, Oet Grebke e António Barreto. Amigos de alma, cujos percursos se cruzaram com o meu de uma forma irreversível. E, claro, ao meu irmão Miguel.

A todos vocês, e ao leitor atento, dedico esta tese com verdade e humildade.

8 de Outubro de 2015,



Notation

Some structures apart from the usual text, figures and tables are used in this work. Their inclusion aims to better organize the conveyed ideas so its content can be easily assimilated by the reader.

Definitions. The introduction of critical concepts are framed by a light blue box. Exemplifying:

Def. 1 A **definition** is a passage describing, and possibly formalizing, the meaning of a concept.

Requirements and Contributions. Key premises for the thesis validation are framed by a green box:

R1 This is an illustrative *requirement*.

Conceptual Maps. Taxonomies are introduced in the beginning of each section to guide the reading. Their coloring aims to better separate concepts, not having other meaning than that.

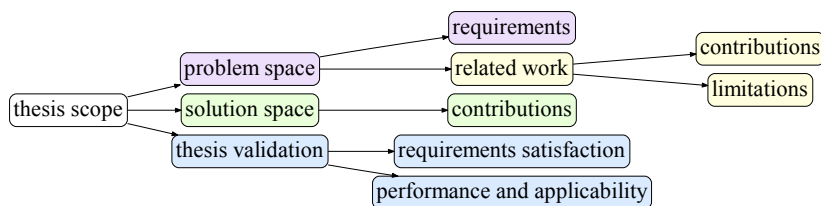


Figure 1: Illustrative conceptual map to structure the covered contents of one section, thus promoting clarity.

Framed text. Colored frames are either used to illustrate basic concepts (Basics label) or to expose contextual contents and complementary readings (Pointers label) in order to guarantee that the main text remains concise.

Basics: Suggestion

Experts may choose to skip these frames.

Pointers: Further information

Examples of pointers include alternative lines of research and applications not directly related with the task under analysis.

Source Code. Algorithms are presented using pseudo-code. Illustrating:

Algorithm 1: The Min-Error algorithm with an earliest first heuristic

Input: Set of tasks and processors

Output: Mapping of tasks to processors

while *Unscheduled tasks remaining* **do**

foreach *Processor j* **do**

if $FT(a,j) \leq FT(a,c)$ **then** $c = j$;

 Schedule(Task *a* on Processor *c*) ;

Cross-Referencing. The document is organized in seven books (I-VII), each grouping a set of chapters. References to sections, figures, tables and the previous structures inside the referee chapter are presented standardly (e.g. *Section 1.1*), while references to contents outside of the chapter are preceded by the book index (e.g. *Section I-1.1*).

Contents

I	Foundations	1
1	Introduction	2
1.1	Universe of Discourse	3
1.1.1	Input Data	3
1.1.2	Output Models	6
1.1.3	Learning Function	10
1.2	Problem Motivation	12
1.2.1	Problems of Dimensionality Reduction	13
1.2.2	Relevance of Local Models: Applications	15
1.3	Thesis Requirements	15
1.4	Solution Space	17
1.4.1	Scientific Dissemination	18
1.5	Contents	18
II	Performance Guarantees of Models Learned from High-Dimensional Data	21
1	Performance Guarantees of Classification Models	24
1.1	Background	25
1.2	Related Work: Limitations and Contributions	26
1.3	Solution: Principles to Bound and Compare Performance	29
1.3.1	Principles for Robust Assessments	29
1.3.2	Assessment Methodology: Integrating the Proposed Principles	35
1.4	Results and Discussion	36
1.5	Summary of Contributions and Implications	40
2	Performance Guarantees of Local Descriptive Models	41
2.1	Bounding and Comparing Descriptive Models	42
2.2	Performance Views for Biclustering Models (Local Descriptors in Tabular Data)	43
2.3	Performance Views for Local Descriptors learned from Structured Data	47
2.3.1	Triclustering Models	47
2.3.2	Cascade Models	48
2.4	Assessment Methodology	50
2.5	Concluding Note	51
3	Synthetic Data Generation for Robust Assessments	52
3.1	Generation of Synthetic Biclustering Data	53
3.1.1	Background	53
3.1.2	Related Work	55
3.1.3	Generating Biclustering Data	55
3.1.4	Data Benchmarks with BiGen	58
3.1.5	Generating Sparse and Network Data	59
3.1.6	Software Description	60
3.2	Generation of Structured Data	61
3.2.1	Generation of Three-Way Time Series with Planted Cascades	61
3.2.2	Generation of Collections of Events with Informative Arrangements	63
3.3	Generation of Labeled Data with Planted Local Regularities	64

3.4	Summary of Contributions and Implications	66
-----	---	----

III Learning Local Descriptive Models from Tabular Data **67**

1	A Structured View on Pattern-based Biclustering	73
1.1	On the Relevance of Pattern-based Biclustering	74
1.1.1	Potentialities of Pattern-based Biclustering	75
1.1.2	Challenges of Pattern-based Biclustering	75
1.1.3	Applicability	75
1.2	Background	76
1.2.1	Pattern Mining	77
1.2.2	Pattern-based Biclustering	78
1.3	Related Work	79
1.4	Structured View on Pattern-based Biclustering	80
1.4.1	Mining Options	81
1.4.2	Mapping Options: Preprocessing Input Data	88
1.4.3	Closing Options: Postprocessing Biclustering Solutions	90
1.5	Comparison of Pattern-based Biclustering Approaches	92
1.6	Summary of Contributions and Implications	93
2	Biclustering Robust to Noise, Missings and Discretization Problems	94
2.1	Affecting the Quality	95
2.2	Robustness to Discretization Problems	96
2.3	Robustness to Noise	97
2.4	Robustness to Missing Values	98
2.5	Customizing Biclustering Structures	99
2.6	BicPAM Algorithm	100
2.6.1	Integrating Contributions	100
2.6.2	Stopping Criteria	101
2.7	Results and Discussion	101
2.7.1	Comparison of Biclustering Approaches in Synthetic Data	102
2.7.2	BicPAM Performance in Synthetic Data	104
2.7.3	Results in Real Data	106
2.8	Summary of Contributions and Implications	109
3	Additive, Multiplicative and Symmetric Models for Tabular Data	110
3.1	Limitations of Related Work	111
3.1.1	Additive and Multiplicative Models	111
3.1.2	Symmetric Models	111
3.1.3	Biclustering Tabular Data	112
3.2	Solution	112
3.2.1	Additive and Multiplicative Models	112
3.2.2	Symmetric Models	113
3.2.3	Biclustering Tabular Data	114
3.3	Results and Discussion	115
3.3.1	Results in Synthetic Data	115
3.3.2	Results in Real Data	118
3.3.3	Comparison of Pattern-based Biclustering Approaches	121
3.4	Summary of Contributions and Implications	121

4	Flexible Plaid Models: Meaningful Interactions between Biclusters	122
4.1	Related Work	124
4.2	Solution	126
4.2.1	Plaid Models from Exhaustive Biclustering Solutions	126
4.2.2	Flexible Composition of Biclusters	128
4.2.3	Complex Interactions	131
4.2.4	Modeling Dependencies between Biological Processes and Social Communities	131
4.3	Results and Discussion	132
4.3.1	Results in Synthetic Data	133
4.3.2	Results in Real Data	136
4.4	Summary of Contributions and Implications	139
4.4.1	Future work	139
5	Scalable Pattern-based Biclustering	140
5.1	Background	141
5.1.1	Motivation	141
5.1.2	Limitations of Related Work	142
5.2	Solution: Handling Efficiency Bottlenecks	142
5.2.1	Full Frequent-Pattern Growth	142
5.2.2	Scalable Full-Pattern Mining	144
5.2.3	Condensed and Approximative Patterns	145
5.2.4	Distributed Settings and Data Partitioning Principles	146
5.3	Results and Discussion	146
5.3.1	Results on Synthetic Data	146
5.3.2	Results on Real Data	147
5.3.3	Pattern-based Biclustering with F2G	148
5.4	Summary of Contributions and Implications	149
6	Flexible and Robust Order-Preserving Biclustering	150
6.1	Background	152
6.1.1	Order-preserving Biclustering	152
6.1.2	Item-Indexable Sequential Pattern Mining	152
6.1.3	Applications	153
6.2	Related Work	153
6.2.1	Sequential Pattern Mining	153
6.2.2	Order-Preserving Biclustering	155
6.3	Solution: BicSPAM	155
6.3.1	Efficiency: Item-Indexable Sequential Pattern Mining	157
6.3.2	Efficiency: Additional Principles	159
6.3.3	Flexibility: Affecting the Properties of Order-preserving Models	160
6.3.4	Robustness: Affecting Quality of Order-Preserving Models	161
6.4	Results	163
6.4.1	IndexSpan Assessment	163
6.4.2	Assessment of BicSPAM	165
6.4.3	Biological Relevance of BicSPAM	167
6.5	Summary of Contributions and Implications	168
7	Biclustering with Efficient Closing Options	169
7.1	Related Work: Limitations and Contributions	170
7.2	Solution: Efficient Merging Procedures	171
7.2.1	Anti-monotonic Heuristics	172
7.2.2	Merging: Maximal Circuits in Acyclic Graphs	173
7.2.3	Merging: Multi-support Frequent Itemset Mining	173
7.2.4	Integrated Mining and Merging of Biclusters	174
7.3	Solution: Removing Bottlenecks of Extension and Reduction Options	174

7.3.1	Parameterized Pattern Mining Searches	174
7.3.2	Constraint-based Guided Discovery	175
7.4	Results and Discussion	175
7.5	Summary of Contributions and Implications	177
8	Effective Biclustering of Large-Scale Network Data	178
8.1	Background	179
8.2	Related Work: Limitations and Contributions	180
8.3	Solution	182
8.3.1	Network Modules with Flexible Coherencies	183
8.3.2	Modules with Noisy and Missing Interactions	185
8.3.3	BicNET: Efficiently Biclustering Network Data	185
8.4	Results and Discussion	186
8.4.1	Results on Synthetic Data	187
8.4.2	Results on Real Data	188
8.5	Summary of Contributions and Implications	192
9	Flexible Pattern-based Biclustering: Putting All Together	194
9.1	Integrative View of Contributions	194
9.1.1	Synergies	196
9.1.2	Algorithmic Solution	197
9.1.3	Computational Complexity Analysis	197
9.2	Learning from Sparse Data: Discarding Uninformative Elements	198
9.3	Default and Dynamic Parameterizations of Pattern-based Biclustering Algorithms	199
9.4	Declarative, Graphical and Programmatic Interfaces	200
9.5	Summary of Contributions and Implications	202
10	Constraint-based Biclustering with Domain Knowledge	203
10.1	Background	204
10.2	Related Work	206
10.3	Pattern-based Biclustering with Background Knowledge	207
10.3.1	Integrating Knowledge from Semantic Sources and Literature	208
10.3.2	Constraints with Nice Properties for Biological and Social Data	208
10.3.3	Biclustering with Full-Constraints	210
10.3.4	Incorporation of Full-Constraints	210
10.4	Exploring Optimum Efficiency Gains from Constraints with Nice Properties	211
10.4.1	F2G-Bonsai: F2G with Constraints	211
10.4.2	IndexSpanPG: IndexSpan with Constraints	212
10.5	Results and Discussion	212
10.5.1	Results on Synthetic Data	213
10.5.2	Results on Biological Data	215
10.6	Summary of Contributions and Implications	217
IV	Learning Local Descriptive Models from Structured Data	219
1	Learning Cascade Models from Multivariate Time Series	223
1.1	Problem Formulation	225
1.1.1	Flexible Modeling of Modules [R1]	227
1.1.2	Discovering the Causal Dependencies [R2]	228
1.1.3	Handling the Diversity and Complexity of Responses [R3]	229
1.1.4	Dealing with Temporal Misalignments between Observations [R4]	229
1.2	Related Work: Limitations and Contributions	230
1.3	Solution	232
1.3.1	Mapping 3TS as a Sequential Database	232

1.3.2	Learning Cascade Models from Sequential Databases	233
1.3.3	Robustness of Cascade Models	235
1.3.4	Guaranteeing the Efficiency of the FReCa	235
1.3.5	Algorithm and Complexity	236
1.4	Results and Discussion	237
1.4.1	Results in Synthetic Data	238
1.4.2	Results in Real Data	241
1.5	Summary of Contributions and Implications	242
1.5.1	Future Work	242
2	Learning Local Descriptive Models from Multi-sets of Events	243
2.1	Background	244
2.2	Related Work	245
2.2.1	Learning Integrative Models from Multi-Sets of Events	246
2.2.2	Learning Descriptive Models from Repositories of Health-Records	246
2.3	Solution	247
2.3.1	Data Mapping	247
2.3.2	Learning Arrangements of Events from Time-Enriched Itemset Sequences	250
2.4	Initial Results and Discussion	251
2.5	Summary of Contributions	252
3	Stochastic Modeling of Itemset Sequences	253
3.1	Background	254
3.1.1	Probabilistic Learning from Sequences	255
3.1.2	Related Work on Markov Modeling of Patterns from Sequential Data	256
3.2	Solution	257
3.2.1	Existing HMM Architectures	257
3.2.2	Applying Existing Architectures over Itemset Sequences	258
3.2.3	Initialization of Transition-Emission Probabilities	259
3.2.4	Robust Learning Settings	260
3.2.5	Principles for Decoding of Patterns	260
3.2.6	Revised Architectures to Model Co-occurrences and Precedences	261
3.2.7	Adequate Convergence of Emissions with Multi-path Architectures	263
3.2.8	Integrative Architectures	264
3.3	Results and Discussion	265
3.3.1	Results on Synthetic Data	266
3.3.2	Results in Real Data	269
3.3.3	Discussion	269
3.4	Summary of Contributions and Implications	270
4	Advanced Stochastic Modeling of Structured Data	271
4.1	Stochastic Learning of Cascade Models from Three-Way Time Series	272
4.1.1	Intrinsic Benefits	272
4.1.2	Handling Multi-Item Assignments	273
4.1.3	Fixed Multivariate Order and Continuous Markov Assumption	273
4.1.4	Postprocessing Procedures	275
4.1.5	Remaining Considerations	275
4.2	Stochastic Learning of Local Models from Multi-sets of Events	275
4.2.1	Benefits of Stochastic Modeling of Arrangements of Events	276
4.2.2	Time-enriched Markov Models	277
4.2.3	Stochastic versus Deterministic Learning	277
4.3	Initial Results and Discussion	278
4.4	Summary of Contributions and Implications	279

V	Significance Guarantees of Local Descriptive Models	283
1	Assessing the Statistical Significance of Biclustering Solutions	287
1.1	Background	288
1.1.1	Limitations	289
1.2	Related Work	290
1.2.1	Significance of Constant Biclusters	290
1.2.2	Deviation from the Expectations	291
1.3	Solution	292
1.3.1	Assessing Perfect Constant Biclusters: Integrating Statistical Principles	292
1.3.2	Robust and Efficient Corrections	294
1.3.3	Global Constraints	295
1.4	Results and Discussion	296
1.5	Summary of Contributions and Implications	299
2	Significance of Biclusters with Flexible Coherencies	300
2.1	Solution: Significance of Biclusters with Flexible Coherency Assumptions	301
2.1.1	Additive Model	301
2.1.2	Multiplicative Model	302
2.1.3	Plaid Model	303
2.1.4	Order-Preserving Model	303
2.1.5	Symmetric Models	303
2.1.6	Global Constraints	304
2.2	Solution: Significance Assessment of Biclusters from Noisy and Sparse Matrices	304
2.2.1	Biclustering Models with Arbitrary-High Levels of Noise	304
2.2.2	Biclustering Models from Sparse Data	305
2.2.3	Combining Significance and Homogeneity Views	305
2.3	Results and Discussion	306
2.4	Summary of Contributions and Implications	309
3	Assessing the Significance of Real-valued Biclusters	310
3.1	Solution: Significance of Real-valued Biclusters	311
3.1.1	Assessing Real-Valued Biclusters	311
3.1.2	Dealing with Continuous Ranges of Shifting Factors	312
3.1.3	Dealing with Continuous Ranges of Scaling Factors	313
3.2	Solution: Tabular Data with Non-Identically Distributed Features	314
3.3	Results and Discussion	315
3.4	Summary of Contributions and Implications	317
4	Significance of Local Descriptive Models from Structured Data	318
4.1	Background	319
4.2	Related Work	319
4.3	Solution	320
4.3.1	Assessing the Significance of Sequential Patterns	321
4.3.2	Controlling False Positive and Negative Rate	321
4.3.3	Assessing the Significance of Highly-Probable Paths in Graphs	322
4.3.4	Assessing the Significance of Cascade Models and Arrangements of Events	323
4.3.5	Using Significance Criteria to Guide the Learning	324
4.4	Summary of Contributions and Implications	324
VI	Learning Effective Classifiers from Local Descriptive Models	325
1	Effective Associative Classification from Discriminative Biclusters	331
1.1	Background	332
1.1.1	Problems of Classification from High-Dimensional Data	333

1.1.2	Motivating Associative Classification	334
1.1.3	Limitations of Existing Associative Classifiers	334
1.2	Related Work	335
1.2.1	Discovery of Discriminative Regions	335
1.2.2	Associative Training Functions	336
1.2.3	Associative Testing Functions	337
1.3	Solution	337
1.3.1	Discovery of Discriminative Biclusters	338
1.3.2	Training	340
1.3.3	Testing	341
1.4	Results and Discussion	342
1.4.1	Results on Synthetic Data	342
1.4.2	Results on Real Data	344
1.5	Conclusion and Implications	345
2	Classification from Regions with Non-Constant Coherency	346
2.1	Background	347
2.2	Solution	349
2.2.1	Discriminative Regions with Flexible Homogeneity	349
2.2.2	Training: Scoring Noisy and Non-Constant Regions	350
2.2.3	Testing: Matching Observations against Non-Constant Regions	350
2.2.4	Algorithm	351
2.3	Results and Discussion	352
2.4	Summary of Contributions and Implications	355
3	Advanced Aspects of Associative Classification	356
3.1	Learning from Sparse Data	357
3.2	Integrating Associative with Global Functions	358
3.3	Stochastic Learning from Generative Biclustering Models	359
3.4	Learning from Complex Tabular Data	360
3.5	Effective Incorporation of Constraints	360
3.6	Summary of Contributions	361
4	Learning Associative Classifiers from Structured Data	362
4.1	Background	364
4.2	Related Work	365
4.2.1	Deterministic and Stochastic Learning of Classifiers from Temporal Data	365
4.2.2	Learning Integrative Models from Structured Data	366
4.3	Solution	367
4.3.1	Data Mappings to Learn from Structured Data	367
4.3.2	Pattern-based Classifiers for Labeled Structured Data	368
4.3.3	Stochastic Classifiers for Structured Data	371
4.4	Results and Discussion	372
4.4.1	Discussion	374
4.5	Summary of Contributions and Implications	375
5	Classification from Significant Regions	376
5.1	Background	377
5.2	Related Work	378
5.3	Solution	379
5.3.1	Significance of Discriminative Biclusters: Local Assessments	379
5.3.2	Using Significance to Shape Associative Models	380
5.3.3	Enhancing Local Classifiers for Tabular Data	381
5.3.4	Enhancing Local Classifiers for Structured Data	383
5.4	Results and Discussion	384

5.5	Summary of Contributions and Implications	387
6	Learning Significant and Accurate Decisions	388
6.1	Solution	389
6.1.1	Significance of Training Functions	389
6.1.2	Significance of Testing Functions	389
6.1.3	Indicative Significance of Classification Decisions	390
6.1.4	Combining Significance and Accuracy Views	390
6.2	Results and Discussion	391
6.3	Conclusion	392
7	Multi-period Classification for Predictive Tasks	393
7.1	Background	394
7.1.1	Formalization	394
7.1.2	Applications	395
7.2	Contributions and Limitations from Related Work	395
7.3	Solution: Methods for Multi-period Classification	397
7.3.1	Hybrid Single-Output Classifiers	398
7.3.2	Cluster-based Multi-Period Classification	398
7.3.3	Extension of CPMC to Capture Local Temporal Dependencies	401
7.3.4	CMPC with Sliding Windows	402
7.4	Evaluation Metrics	402
7.4.1	Baseline Evaluation	403
7.4.2	Multi-period Classification with Ordinal Labels	404
7.4.3	Compact Performance Views	404
7.4.4	Complementary Evaluation Metrics	405
7.5	Results and Discussion	405
7.5.1	Comparing Multi-period Classifiers	407
7.5.2	Discussion	411
7.6	Summary of Contributions and Implications	412
VII	Closing	413
1	Final Discussion	414
1.1	Hypothesis Validation	414
1.2	Thesis Contributions and Implications	414
1.3	Future Work	421

I Foundations

Introduction

Learning from high-dimensional data, where the high number of features can exceed the number of observations, is challenged by an inherent complexity and generalization difficulty. In these data contexts, these challenges can be minimized by focus the learning on specific regions of interest (such as subsets of features) [316, 302]. However, the lack of flexibility on how the existing learning methods select such regions is associated with three major problems: 1) the inclusion of non-relevant regions (promoting overfitting), 2) the exclusion of relevant regions (promoting underfitting), and 3) the modeling of apparently relevant regions, yet not statistically significant [105, 316, 22]. As illustrated in Figure 1.1, learning from high-dimensional data is a challenging task since not all regions are equally informative and, even when informative, regions may not be statistically significant. Furthermore, they can be significantly informative yet non-significantly discriminative.

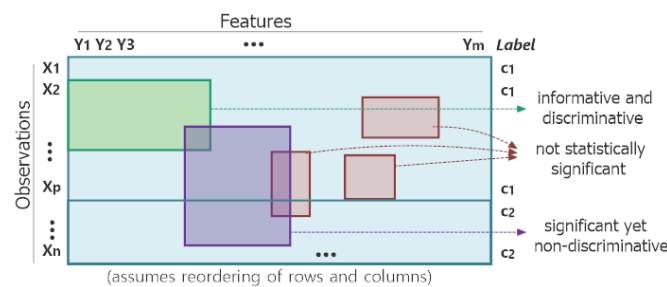


Figure 1.1: Learning from high-dimensional data: relevance of selecting coherent, discriminative, significant regions.

This work aims to tackle these challenges by learning flexible, robust and statistically significant descriptive models and associative classification models from relevant regions of a high-dimensional data space. This learning task is addressed for tabular and structured data. In this context, our goal is to systematically study how does the performance of descriptive and classification models vary with the homogeneity, discriminative power and statistical significance of the selected regions. The underlying *hypothesis* is that the adequate selection and composition of regions improves the performance guarantees of local descriptive models and (associative) classifiers learned from high-dimensional data. As a result, this understanding opens a window of opportunity: new principles can be inferred to revise the behavior of existing learning methods.

The importance of this thesis is driven by two major observations. First, the need to validate the increasing number of scientific statements from the analysis of high-dimensional data without proper statistical assessments [316]. This is particularly critical across biomedical domains, due to the severity of implications that some statements might have on human health and upcoming research. Second, the need to face the increasing dimensionality of the available data, without being susceptible to the problems of feature selection and peer procedures for dimensionality reduction. The focus on a small subset of features is commonly associated with the exclusion of relevant regions and inclusion of non-relevant elements, contributing to the over/underfitting risk.

The in-depth study of how to optimize the performance of the target learning methods, while guaranteeing their statistical significance, is thus critical to answer a wide-set of real-world learning problems. In this work, we address the tasks of learning from: 1) tabular data from biomedical domains with a high number of molecular units or clinical features per sample or patient, and social domains with a high number of rated items, traits or behavioral

features per subject; 2) weighted graphs given by large-scale biological and social networks; 3) sequential data associated with (multivariate) time series with a high number of time points and/or (sliding) features; and 4) structured data mapped from high-dimensional multi-sets of events, such as repositories of health records, trading decisions, (e-)commerce operations and browsing events.

This chapter is organized as follows. *Section 1.1* formalizes the universe of discourse of this thesis. *Section 1.2* explores the current limitations and opportunities of learning from high-dimensional data. *Section 1.3* structures the problem space according to its requirements. *Section 1.4* provides a high-level view on the contributions of this thesis. Finally, a roadmap for an easy exploration of the contents in this dissertation is provided in *Section 1.5*.

1.1 Universe of Discourse

This section formalizes the target learning task. According to Figure 1.2, follows a characterization of the possible *inputs* (high-dimensional data) in *Section 1.1.1*, the desirable *outputs* (learned models) in *Section 1.1.2*, and the learning functions in *Section 1.1.3*.

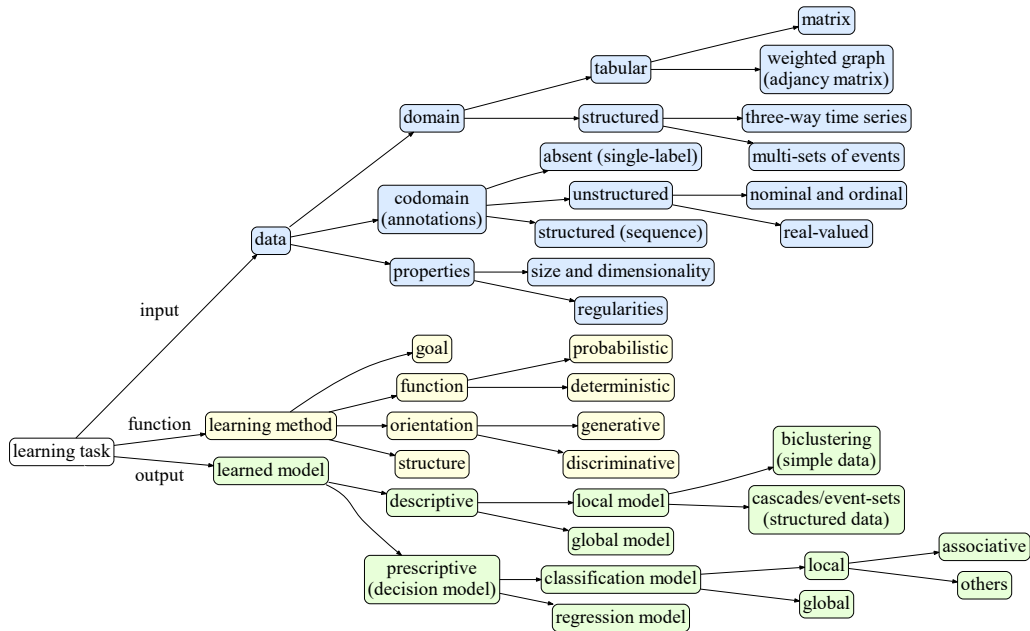


Figure 1.2: Learning task according to its input, function and output.

1.1.1 Input Data

Learning from high-dimensional data has both challenges dependent and independent of the input data format. In this work, we first tackle the task of learning from (high-dimensional) data given by real-valued matrices and weighted graphs. Second, we move towards learning from structured data given by multivariate time series, itemset sequences, multi-sets of events and multi-dimensional data.

Basics 1.1 Notation

A random variable (or random vector/matrix in bold) is a function denoted by a capital letter (\mathbf{X} , \mathbf{Y} , \mathbf{A} , \mathbf{C}) from an event space into a sample space (denoted by caligraphic style: \mathcal{X} , \mathcal{Y} , \mathcal{A} , \mathcal{C}), with outcomes denoted by the corresponding lower case (\mathbf{x} , \mathbf{y} , \mathbf{a} , c).

A dataset is defined by a set of observations, a sample of a (possibly) random vector \mathbf{X} taking values on a sample space \mathcal{X} with probability function $P_{\mathbf{X}}(\mathbf{x})$, or simply $P_{\mathbf{X}}$. In labeled datasets, the \mathbf{X} observations are (possibly) conditionally dependent on the assigned labels $c \in C$, also referred as classes, and thus described by a class-conditional probability function $P_{\mathbf{X}}(\mathbf{x}|c)$, or simply $P_{\mathbf{X}|C}$. As we move from tabular to structured datasets, the observations take values on a structured sample space \mathcal{X} .

Data Structures

Below we introduce tabular data structures. Real-valued matrices and weighted graphs can be seen as specializations of tabular data where features are numeric and thus $\mathcal{X} = \mathbb{R}^m$ (where m is the number of features). Missing interactions from graphs are seen as interactions with a zero weight. Tables 1.3-1.6 illustrate these data structures.

Def. 1.1 A **real-valued matrix** \mathbf{A} with n observations (rows) $\mathbf{x}_i \in \mathbb{R}^m$, m features (columns) $\mathbf{y}_j \in \mathbb{R}^n$, and $n \times m$ elements $a_{ij} \in \mathbb{R}$ is a (n, m) -space. Let Σ be a set of categoric values (classes), a labeled real-value matrix given by a (n, m) -space is described by n labeled observations (\mathbf{x}_i, c_i) , where $\mathbf{x}_i \in \mathbb{R}^m$ and $c_i \in \Sigma$.

Def. 1.2 A **tabular dataset** \mathbf{A} has n observations $\mathbf{x}_i \in \mathcal{X}$ (possibly labeled), m features $\mathbf{y}_j \in \mathcal{Y}_j$, and $n \times m$ elements $a_{ij} \in \mathcal{Y}_j$, where \mathcal{Y}_j defines the domain of the \mathbf{y}_j feature: either nominal, ordinal or numeric.

Basics 1.2 Tabular data structures

Figures 1.3 and 1.4 provide a real-valued matrix (\mathbf{A}_1) and an alternative tabular dataset (\mathbf{A}_2). \mathbf{A}_1 is a labeled $(n=6, m=7)$ -space with 2 classes (4 c_1 -conditional observations and 2 c_2 -conditional observations). \mathbf{A}_2 has 5 observations and 6 features, each feature \mathbf{y}_i taking values on a specific sample space \mathcal{Y}_i with either numeric values (\mathcal{Y}_1 and \mathcal{Y}_5), nominal values with varying cardinality (\mathcal{Y}_3 , \mathcal{Y}_4 and \mathcal{Y}_6), or ordinal values (\mathcal{Y}_2). Illustrating the concept of data elements: $a_{2,3}=3.9$ in \mathbf{A}_1 and $a_{2,3}=b$ in \mathbf{A}_2 .

Figure 1.3: Illustrative real-valued matrix (\mathbf{A}_1)

	\mathbf{y}_1	\mathbf{y}_2	\mathbf{y}_3	\mathbf{y}_4	\mathbf{y}_5	\mathbf{y}_6	\mathbf{y}_7	class
\mathbf{x}_1	2.7	-0.9	2.2	2.7	-0.9	1.1	0.9	c_1
\mathbf{x}_2	0.8	-2.1	3.9	0.1	-2.1	-0.1	1.9	c_1
\mathbf{x}_3	1.3	-3.8	2	-2.8	-3.8	0	0.1	c_1
\mathbf{x}_4	-2.7	-1.8	3.7	1.9	-2.2	-0.9	2.1	c_1
\mathbf{x}_5	0.7	2.9	0.2	-2	-0.8	1.9	-1.1	c_2
\mathbf{x}_6	-2.6	-3.1	-0.1	3.9	0.9	-2	1.2	c_2

Figure 1.4: Illustrative tabular dataset (\mathbf{A}_2)

	\mathbf{y}_1 ($\mathcal{Y}_1=\mathbb{R}$)	\mathbf{y}_2 ($\mathcal{Y}_2=\mathbb{N}$)	\mathbf{y}_3 ($ \mathcal{Y}_3 =3$)	\mathbf{y}_4 ($ \mathcal{Y}_4 =7$)	\mathbf{y}_5 ($\mathcal{Y}_5=\mathbb{N}$)	\mathbf{y}_6 ($\mathcal{Y}_6=\{Y,N\}$)	class
\mathbf{x}_1	0.3	3	A3	A4	63	Y	c_1
\mathbf{x}_2	1	5	B3	A4	22	Y	c_1
\mathbf{x}_3	-2.1	3	A3	F4	31	N	c_1
\mathbf{x}_4	-0.1	1	C3	E4	28	Y	c_2
\mathbf{x}_5	3.2	4	B3	A4	42	N	c_2

Def. 1.3 A **weighted bipartite graph** is defined by two disjoint sets of nodes $\mathbf{X}=\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ (observations) and $\mathbf{Y}=\{\mathbf{y}_1, \dots, \mathbf{y}_m\}$ (features) where $\mathbf{x}_i \in \mathbb{R}^m$ and $\mathbf{y}_j \in \mathbb{R}^n$, and weighted interactions $a_{ij} \in \mathbb{R}$ between nodes \mathbf{x}_i and \mathbf{y}_j . A **weighted graph** is defined by a set nodes $X=\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ (observations and features) where $\mathbf{x}_i \in \mathbb{R}^n$ and $a_{ij} \in \mathbb{R}$ interactions between nodes \mathbf{x}_i and \mathbf{x}_j . Given a set of labels Σ , nodes can be labeled ($\mathbf{x}_i, c_i \in \Sigma$).

Basics 1.3 Network data: (weighted) graphs

Figure 1.5 depicts a weighted graph with labeled nodes. Figure 1.6 provides the result of mapping this graph into a real-valued matrix (\mathbf{A}_3), with 5 observations and 5 features. The weight of interaction between nodes with identifiers i and j corresponds to the a_{ij} element in the mapped matrix.

Figure 1.5: Illustrative weighted graph (\mathbf{A}_3).

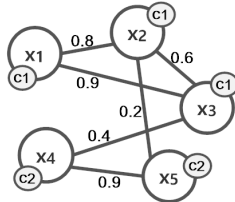


Figure 1.6: Real-valued matrix (\mathbf{A}_3) mapped from Figure I-1.5.

	\mathbf{y}_1 (\mathbf{x}_1)	\mathbf{y}_2 (\mathbf{x}_2)	\mathbf{y}_3 (\mathbf{x}_3)	\mathbf{y}_4 (\mathbf{x}_4)	\mathbf{y}_5 (\mathbf{x}_5)	class
\mathbf{x}_1	0	0.8	0.9	0	0	c_1
\mathbf{x}_2	0.8	0	0.6	0	0	c_1
\mathbf{x}_3	0.9	0.6	0	0.4	0	c_1
\mathbf{x}_4	0	0	0.4	0	0.9	c_2
\mathbf{x}_5	0	0.2	0	0.9	0	c_2

The introduced data structures are consider to be the default input along *Books II* and *III* of this document. Since new data structures are becoming increasingly prominent, bringing new challenges towards traditional learning tasks, we also consider multivariate time series, sequential databases and multi-sets of events. Figures 1.7-1.9 illustrate these data structures. We also provide mappings of multi-dimensional and relational databases into multi-sets of events in order to guarantee a wide-coverage of real-world (high-dimensional) data structures. An analysis of their application domains is provided in the next chapter.

Def. 1.4 A **three-way time series** (or cube) is a set of observations $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, where each observation \mathbf{x}_i defines a matrix with m features $\mathbf{y}_j \in \mathbb{R}^p$, p time points $\mathbf{t}_k \in \mathbb{R}^m$, and elements $a_{ijk} \in \mathbb{R}$ relating observation \mathbf{x}_i , feature \mathbf{y}_j and time point \mathbf{t}_k . A cube is also referred as a *multivariate time series* database, where each time series has a m order and p time points. Given a set of labels Σ , observations (time series) can be labeled $(\mathbf{x}_i, c_i \in \Sigma)$.

Basics 1.4 Multivariate time series

Figure 1.7 instantiates a labeled time series database \mathbf{A}_4 , with 4 multivariate time series labeled with a class (*time series* \Rightarrow *class*), each time series (matrix) has a 5 multivariate order with measurements along 4 time points. Alternatively, Figure 1.7 can be seen as an integer cube with 4 matrices (or observations), each with 5 rows and 4 columns.

Figure 1.7: Illustrative set of multivariate time series (\mathbf{A}_4)

	$\mathbf{x}_1 \Rightarrow c_1$				$\mathbf{x}_2 \Rightarrow c_1$				$\mathbf{x}_3 \Rightarrow c_2$				$\mathbf{x}_4 \Rightarrow c_2$						
	\mathbf{t}_1	\mathbf{t}_2	\mathbf{t}_3	\mathbf{t}_4	\mathbf{t}_1	\mathbf{t}_2	\mathbf{t}_3	\mathbf{t}_4	\mathbf{t}_1	\mathbf{t}_2	\mathbf{t}_3	\mathbf{t}_4	\mathbf{t}_1	\mathbf{t}_2	\mathbf{t}_3	\mathbf{t}_4			
\mathbf{y}_1	1	0	-2	-2	\mathbf{y}_1	1	-1	-2	-2	\mathbf{y}_1	1	-1	0	1	\mathbf{y}_1	0	-2	-1	0
\mathbf{y}_2	1	1	-2	-2	\mathbf{y}_2	0	0	-2	-2	\mathbf{y}_2	2	2	0	-1	\mathbf{y}_2	2	2	0	0
\mathbf{y}_3	1	2	2	2	\mathbf{y}_3	-1	2	2	2	\mathbf{y}_3	2	2	-2	-2	\mathbf{y}_3	2	2	-2	-2
\mathbf{y}_4	0	2	2	2	\mathbf{y}_4	0	2	2	2	\mathbf{y}_4	0	1	2	2	\mathbf{y}_4	-1	0	2	2
\mathbf{y}_5	-2	1	0	1	\mathbf{y}_5	1	0	1	0	\mathbf{y}_5	-1	-1	2	2	\mathbf{y}_5	-2	-1	2	2

Def. 1.5 Given a set of items \mathcal{L} , let an (itemset) sequence be an ordered set of itemsets $\langle I_1, \dots, I_m \rangle$, where $I_i \subseteq \mathcal{L}$. A *sequential database* is a set of n sequences (observations). Sequences can be labeled.

Def. 1.6 Let an event μ from a source (observation) \mathbf{x}_i be a tuple (y_j, a_{ijk}, t_{ijk}) , where $y_j \in \mathcal{Y}_j$ is the type of event (feature), a_{ijk} is its value and t_{ijk} the timestamp. A repository of *multi-sets of events* is a set of n observations $\mathbf{x}_i \in \mathcal{X}$, where each observation is a set of timestamped events. Observations can be labeled.

Basics 1.5 Sequential databases and multi-sets of events

Figures 1.8 and 1.9 show respectively an instance of a sequential database (\mathbf{A}_5) and of a multi-set of events (\mathbf{A}_6).

In Figure 1.8, itemset sequences were represented with a compact format: co-occurring items are delimited by curve parentheses and itemsets are concatenated. Illustrating, $\mathbf{x}_1 = \langle \{a, h\}, \{d\}, \{a, b\}, \{b, e, g\}, \{a, c, f\} \rangle = (ah)d(ab)(beg)(acf)$. \mathbf{A}_5 is a set of 5 labeled itemset sequences (3 c_1 -conditional and 2 c_2 -conditional observations) with items in \mathcal{L} ($|\mathcal{L}|=8$). Illustrating further properties, \mathbf{x}_4 is an ordered set of 6 itemsets with a total of 12 items and an average number of 2 items per itemset.

The labeled multi-sets of events provided in Figure 1.9 (\mathbf{A}_6) has 4 observations (4 distinct sources of events) and 3 features (3 distinct types of events) with different feature domains ($\mathcal{Y}_1 = \mathbb{N}$, $\mathcal{Y}_2 = \{y\}$ and $\mathcal{Y}_3 = \{a, b, c\}$). Each observation has an arbitrary number of timestamped events per feature. For instance, \mathbf{x}_1 has a total of 4 event occurrences, 2 associated with \mathbf{y}_1 feature ($(3, t_2)$ and $(5, t_4)$), no event associated with \mathbf{y}_2 feature and two events associated with \mathbf{y}_3 that co-occur in time point t_1 .

Figure 1.8: Illustrative sequential database (\mathbf{A}_5)

	sequence ($\mathcal{L}=\{a,b,c,d,e,f,g,h\}$)	class
\mathbf{x}_1	$(ah)d(ab)(beg)(acf)$	c_1
\mathbf{x}_2	$(bd)(ab)(be)(bf)a$	c_1
\mathbf{x}_3	$(de)h(ab)g(bef)(fg)$	c_1
\mathbf{x}_4	$b(ab)(abce)(de)b(df)$	c_2
\mathbf{x}_5	$(ad)(ac)f(ad)(cdf)(ab)g$	c_2

Figure 1.9: Illustrative multi-sets of events (\mathbf{A}_6)

	\mathbf{y}_1 ($ \mathcal{Y}_1 =\mathbb{N}$)	\mathbf{y}_2 ($ \mathcal{Y}_2 =1$)	\mathbf{y}_3 ($ \mathcal{Y}_3 =3$)	class
\mathbf{x}_1	$\{(3, t_2), (5, t_4)\}$	\emptyset	$\{(a, t_2), (c, t_2)\}$	c_1
\mathbf{x}_2	$\{(2, t_2), (3, t_3), (4, t_4)\}$	$\{(y, t_3)\}$	$\{(c, t_1)\}$	c_1
\mathbf{x}_3	\emptyset	$\{(y, t_2), (y, t_4)\}$	$\{(b, t_3), (c, t_4)\}$	c_2
\mathbf{x}_4	$\{(2, t_1), (1, t_4)\}$	$\{(y, t_1)\}$	$\{(b, t_1), (a, t_3), (c, t_3)\}$	c_2

Labels: Data Codomain

As introduced, a labeled dataset is a sample from a class-conditional probability function $P_{\mathbf{X}|C}$, where \mathbf{X} is a random vector taking values on a (possibly structured) sample space \mathcal{X} (the *domain*), and C is a random variable with classes from a sample space \mathcal{C} (often referred as *codomain*). Let Σ be a set of labels, the classes can be *nominal*, $\mathcal{C}=\Sigma$ (default case), *ordinal* when Σ is an ordered set, or *numeric* when $\mathcal{C}=\mathbb{R}$. Like domains, codomains can be structured. In particular, we tackle the case where labels are associated with *categoric vectors*, $\mathcal{C}=\Sigma^h$.

Data Properties

A dataset is primarily characterized by the number of observations (*size*), dimensionality, and data regularities.

The *dimensionality* of a dataset is given by the number of columns/features in tabular data (m); number of nodes in weighted graphs; product of the number of features (multivariate order) and time points in multivariate time series ($m \times p$); average number of items per itemset sequence in sequential databases; and average number of events per observation in multi-sets of events. In this context, the criteria to decide whether a dataset is high-dimensional deserves some attention. **High dimensionality** has been seen not only as a product of dimensionality, but also dependent on the complexity of the learning task [553, 316]. There is a considerably agreement that a dimensionality superior to 100 can be considered already high for common learning tasks (based on the suggested cut-off thresholds to apply feature selection) [589, 215, 343]. Complementarily, the higher the learning complexity, the lower the dimensionality threshold to consider a dataset to be high-dimensional. As such, high-dimensionality can be seen as a result of three major aspects:

- number of observations and classes. The lower the ratio $n/|C|$ (where $|C|=1$ for non-labeled data), the higher is the learning complexity due to an increased difficulty to generalize;
- type of input data. In structured data contexts, the learning complexity increases more rapidly with an increasing number of features (or types of events) than with an increasing number of time points/partitions. Illustrating, for a multivariate time series database, the ratio $m\sqrt{p} > 100$ can be considered to be a more fair verification of high-dimensionality;
- the regularities of the input data. Illustrating, the higher the number of correlated features, the higher the learning complexity. A high number of studies aim to theoretically or empirically predict the minimum number of observations for an adequate learning based on the regularities of a given dataset with fixed dimensionality [218, 110, 178, 54]. High-dimensionality is here assumed when the number of available observations is lower than the expected minimum number of observations.

Basics 1.6 Data dimensionality

Consider the data from Figures 1.3-1.7. Their dimensionality is: $\dim(\mathbf{A}_1)=m=7$, $\dim(\mathbf{A}_2)=m=6$, $\dim(\mathbf{A}_3)=n=5$ and $\dim(\mathbf{A}_4)=m \times p=20$.

Let the number of itemsets of an itemset sequence \mathbf{x} be $|\mathbf{x}|$, the i^{th} itemset of \mathbf{x} be \mathbf{x}^i , and the number of items of an itemset I be $|I|$, then $\dim = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{|\mathbf{x}^i|} |\mathbf{x}_i^j|$.

Given a multi-set of events, let the set of events of type $j \in J$ from source \mathbf{x} be \mathbf{x}^j , then $\dim = \frac{1}{n} \sum_{i=1}^n \sum_{j \in J} |\mathbf{x}_i^j|$.

Accordingly, the dimensionality of datasets in Figures 1.8 and 1.9 is respectively $\dim(\mathbf{A}_5) = 11.2$ and $\dim(\mathbf{A}_6) = 4.75$

1.1.2 Output Models

Given a dataset \mathbf{A} characterized by a set of underlying stochastic regularities, $P_{\mathbf{A}}$ (given by $P_{\mathbf{X}}$ or $P_{\mathbf{X}|C}$), a *learning task* aims to infer a model M from this (n, m) -space such that the error over $P_{\mathbf{A}}$ is minimized.

In this thesis, we tackle different learning tasks according to three major qualities: **flexibility** (the ability to affect the properties of the output model), **robustness** (the ability to deal with noisy and missing data) and statistical **significance** (the ability to exclude spurious regularities).

Two major types of models are considered: *descriptive* and *classification* models. These models can be further categorized according to the extent of the space coverage (*global* or *local*) and properties.

Descriptive Models

Def. 1.7 A **descriptive model** abstracts either locally or globally the regularities of a (possibly labeled) dataset, $M(\mathbf{A})$. A *regularity* is a trend in data. While *unsupervised* descriptors ($|C|=1$) model observations, $P_{\mathbf{X}}$, *supervised* descriptors ($|C|>1$) model class-conditional observations, $P_{\mathbf{X}|C}$.

A descriptive model (Def.1.9), also referred as an observation model, is either unsupervised or supervised depending on whether there is knowledge regarding the assignment of probability functions (labels per observation). Supervised descriptors are also referred as class-conditional observation models. Understandably, supervised descriptive models differ from decision models, such as classification models, since their goal is description and not

the labeling of new observations.

Two distinct criteria of locality can be considered to identify whether a model is global or local. First criterion: the model is able to separate groups of observations with distinct regularities. Under this criterion, a descriptive model that approximates a distribution for each feature of a tabular dataset is not local. As such, descriptive models that define a multivariate distribution per class are global and thus not able to accurately describe datasets where groups of observations with a shared class show distinct regularities. Contrasting, clustering models are able to accommodate this locality criteria (Def.1.8). In particular, there are dedicated research streams on clustering to describe not only tabular data, but also multivariate time series, sequential databases and multi-sets of events [74, 36]. In tabular datasets, clustering can be alternatively applied to group subsets of features with correlated values across observations. This possibility can be seen as an alternative form of locality.

Def. 1.8 Given a dataset with \mathbf{X} observations, a **clustering model** is a set of subsets of observations (clusters), $\{\mathbf{X}_1, \dots, \mathbf{X}_l\}$ where $\mathbf{X}_i \subseteq \mathbf{X}$, with intra-cluster and inter-cluster guarantees of (dis)similarity between observations.

Basics 1.7 Global mixtures versus clustering models

Considering the illustrative dataset \mathbf{A}_1 (Figure 1.3) with 2 classes and 7 real-valued features. Assuming independence between features, let us consider the simplistic task of learning a global mixture given by class-conditional multivariate Gaussian distributions. Given \mathbf{A}_1 , then $\mathbf{y}_1|c_1 \sim N(\mu=\frac{1}{4}\sum_{i=1}^4 a_{i1}=0.5, \sigma^2=\frac{1}{4-1}\sum_{i=1}^4 (a_{i1}-0.5)^2=5.3)$, $\mathbf{y}_1|c_2 \sim N(-1.0, 5.4)$, $\mathbf{y}_2|c_1 \sim N(-2.2, 1.5)$, $\mathbf{y}_2|c_2 \sim N(-0.1, 18.0)$, and so forth. Understandably, this model suffers from a major drawback. The inability to distinguish between subsets of observations with a shared class yet with different regularities leads to: biases in the Gaussian's mean and an increased variance that blurs the discriminative power of a feature. On top of this observation, there is a high overfitting propensity for data with a limited number of observations, such as \mathbf{A}_1 . The high variance associated with the observed values for the illustrated features \mathbf{y}_1 and \mathbf{y}_2 shows that they are insufficient to effectively characterize and distinguish c_1 and c_2 -conditional regularities.

Contrasting with this global mixture, let us consider a mixture given by a clustering model with 2 clusters of observations per class based on their Euclidean distance, where a cluster is characterized by the average value per feature (mean centroid). Given \mathbf{A}_1 , the similarity between 2 observations is given by $d(\mathbf{x}_a, \mathbf{x}_b) = \sqrt{\sum_{j=1}^7 (a_{aj} - a_{bj})^2}$, leading to the following groups of cluster for class c_1 : cluster₁={ $\mathbf{x}_2, \mathbf{x}_4$ } and cluster₂={ $\mathbf{x}_1, \mathbf{x}_3$ }. The mean centroids of these clusters are [-0.95,-1.95,3.8,1,-2.15,-0.5,2] and [2,-2.35,2.1,-0.05,-2.35,0.55,0.5], respectively. For the given distance, this clustering model is able to achieve a reasonable intra-cluster similarity, low inter-cluster similarity and some notable differences between clusters from c_1 and c_2 classes.

Second criterion of locality: the model is able to identify regions given by subsets of observations and subsets of features in tabular data (or subsets of time points, items and events in structured data). Illustrating, in the presence of real-valued matrices, a local descriptive model is a composition of learned regions from the original space, where each region \mathbf{B}_i is a (r_i, s_i) -space where $r_i \leq n \wedge s_i \leq m$. We consider this criterion to be the default locality criterion in this work due to its higher flexibility to discard non-informative and non-discriminative regions, a critical condition when learning from high-dimensional data (Figure 1.1). Under this criterion, clustering models, are considered to be global. Contrasting, more flexible descriptive models, such as biclustering models, are local.

Local Descriptive Models

Def. 1.9 A **local descriptive model** is a composition of regions from a (possibly structured) dataset. A *region* is a subset of overall data elements that satisfies certain homogeneity and (possibly) discriminative criteria.

Local descriptors aim to learn relevant regions, subspaces from a (possibly structured) sample data space. Given a tabular dataset, an element a_{ij} relates the \mathbf{x}_i observation and \mathbf{y}_j feature or nodes \mathbf{x}_i and \mathbf{x}_j . Given a three-way time series, an element a_{ijk} relates the \mathbf{x}_i observation (time series), \mathbf{y}_j feature and \mathbf{t}_k time point. Given a sequential database and multi-set of events, an element corresponds respectively to an occurring item and event.

As introduced, the properties of a local descriptive model depend on the structure of the input data. Below, we define flexible descriptive models for real-valued matrices, multivariate time series and itemset sequences. For these data structures, regions of interest are respectively given by biclusters, triclusters and sequential patterns. The formalized concepts associated with these descriptive models are instantiated in Basics 1.8, 1.9 and 1.10. To

preserve conciseness, an in-depth analysis of these models, as well as of additional variants, is provided throughout *Books III* and *IV* of this document.

Def. 1.10 Given a real-valued matrix \mathbf{A} with n observations (rows) \mathbf{X} in \mathbb{R}^m and m features (columns) \mathbf{Y} in \mathbb{R}^n , a *bicluster* $\mathbf{B} = (\mathbf{I}, \mathbf{J})$ is a region/subspace of the original (n, m) -space, where $\mathbf{I} \subseteq \mathbf{X}$ is a subset of rows and $\mathbf{J} \subseteq \mathbf{Y}$ a subset of columns. The **biclustering** task aims to find a set of biclusters $\{\mathbf{B}_1, \dots, \mathbf{B}_l\}$ such that each bicluster \mathbf{B}_i satisfies specific criteria of *homogeneity*, *discriminative power* in the presence of labels, and statistical *significance*.

Basics 1.8 Biclustering real-valued matrices

Considering the introduced \mathbf{A}_1 matrix in Figure 1.3, the $(2,4)$ -space given by $\mathbf{B}_1 = (\mathbf{I} = \{x_2, x_4\}, \mathbf{J} = \{y_2, y_3, y_5, y_7\})$ is coherent, discriminative and significant, and can thus be seen as one bicluster from a biclustering model learned from \mathbf{A}_1 . To facilitate the analysis of the properties of \mathbf{B}_1 , Figure 1.10 provides a variant matrix of \mathbf{A}_1 where the values were rounded and both rows and columns were reordered. First, \mathbf{B}_1 has approximately constant values across observations, a common form of homogeneity. Second, the combination of values per bicluster's row, $\{-2, 4, -2, 2\}$, is supported by 2 observations with class c_1 and 0 observations with class c_2 , indicating a potentially high discriminative power. Finally, half of the total elements from c_1 -conditional data are covered by the bicluster, possibly indicating its statistical significance.

Figure 1.10: Illustrative bicluster in \mathbf{A}_1 matrix (rounded values and reordered rows/columns)

	y ₁	y ₂	y ₃	y ₅	y ₇	y ₄	y ₆	C
x ₁	3	-1	2	-1	1	3	1	c ₁
x ₂	1	-2	4	-2	2	0	0	c ₁
x ₄	-3	-2	4	-2	2	2	-1	c ₁
x ₃	1	-4	2	-4	0	-3	0	c ₁
x ₅	1	3	0	-1	-1	-2	2	c ₂
x ₆	-3	-3	0	1	1	4	-2	c ₂

Def. 1.11 Given a real-valued cube (multivariate time series database) \mathbf{A} with n observations (matrices) \mathbf{X} , m features (rows) \mathbf{Y} and p time points (columns) \mathbf{T} : a *tricluster* $\mathbf{B} = (\mathbf{I}, \mathbf{J}, \mathbf{K})$ is a subspace of the original space, where $\mathbf{I} \subseteq \mathbf{X}$ and $\mathbf{J} \subseteq \mathbf{Y}$ are subsets of observations and features, and $\mathbf{K} \subseteq \mathbf{T}$ is a subset of contiguous time points. Given \mathbf{A} , the **triclustering** task aims to find a set of triclusters $\{\mathbf{B}_1, \dots, \mathbf{B}_l\}$ such that each tricluster \mathbf{B}_i satisfies specific criteria of *homogeneity*, *discriminative power* in the presence of labels, and statistical *significance*.

Def. 1.12 Given a real-valued cube \mathbf{A} with n observations and a set of triclusters (modules) supported by the same subset of observations, there is a high chance that these modules are correlated. A *cascade* (or frequent response) R is a set of l modules $\{\mathbf{B}_1, \dots, \mathbf{B}_l\}$ related through r temporal dependencies $\mathbf{D} = \{d_1, \dots, d_r\}$, where d_i is a sequential constraint defining either a parallel occurrence $\{\mathbf{B}_i, \mathbf{B}_j\}$ or precedence $\mathbf{B}_i \Rightarrow \mathbf{B}_j$ between two modules. Given \mathbf{A} , the task of **modeling cascades** aims to learn a set of cascades $\{R_1, \dots, R_s\}$ satisfying specific criteria of *homogeneity*, *discriminative power* in the presence of labels, and statistical *significance*.

Basics 1.9 Modeling triclusters and cascades from real-valued cubes

Considering the integer cube \mathbf{A}_4 provided in Figure 1.7, some of its regularities $P_{\mathbf{A}_4}$ can be given by diverse coherent modules. Figure 1.11 illustrates some of these modules given by triclusters (subsets of observations, features and time points). We highlight 4 triclusters: 2 triclusters for c_1 -conditional observations ($\mathbf{B}_1 = (\mathbf{I}_1 = \{x_1, x_2\}, \mathbf{J}_1 = \{y_3, y_4\}, \mathbf{K}_1 = \{t_2, t_3, t_4\})$ and $\mathbf{B}_2 = (\mathbf{I}_2 = \{x_1, x_2\}, \mathbf{J}_2 = \{y_1, y_2\}, \mathbf{K}_2 = \{t_3, t_4\})$) and 2 triclusters for c_2 -conditional observations ($\mathbf{B}_3 = (\mathbf{I}_3 = \{x_3, x_4\}, \mathbf{J}_3 = \{y_2, y_3\}, \mathbf{K}_3 = \{t_1, t_2\})$ and $\mathbf{B}_4 = (\mathbf{I}_4 = \{x_3, x_4\}, \mathbf{J}_4 = \{y_3, y_4, y_5\}, \mathbf{K}_4 = \{t_3, t_4\})$). All of these triclusters: 1) show constant values per feature, a commonly accepted form of homogeneity; 2) are supported by observations of a single class, and thus discriminative; and 3) appear to be statistically significant as they are supported by all of the observations of a particular class and include at least 20% of the total elements from these observations.

Figure 1.11: Triclusters (and implicit cascades of triclusters) from the \mathbf{A}_4 cube

$x_1 \Rightarrow c_1$					$x_2 \Rightarrow c_1$					$x_3 \Rightarrow c_2$					$x_4 \Rightarrow c_2$				
	t ₁	t ₂	t ₃	t ₄		t ₁	t ₂	t ₃	t ₄		t ₁	t ₂	t ₃	t ₄		t ₁	t ₂	t ₃	t ₄
y ₁	1	0	-2	-2	y ₁	1	-1	-2	-2	y ₁	1	-1	0	1	y ₁	0	-2	-1	0
y ₂	1	1	-2	-2	y ₂	0	0	-2	-2	y ₂	2	2	0	-1	y ₂	2	2	0	0
y ₃	1	2	2	2	y ₃	-1	2	2	2	y ₃	2	2	-2	-2	y ₃	2	2	-2	-2
y ₄	0	2	2	2	y ₄	0	2	2	2	y ₄	0	1	2	2	y ₄	-1	0	2	2
y ₅	-2	1	0	1	y ₅	1	0	1	0	y ₅	-1	-1	2	2	y ₅	-2	-1	2	2

Understandably, due to the temporal nature inherent to \mathbf{A}_4 , relations between the discovered triclusters can be hypothesized, leading to meaningful cascades of coherent behavior. Considering triclusters \mathbf{B}_1 and \mathbf{B}_2 , they seem to occur in parallel, being the coherency of \mathbf{B}_2 possibly triggered by the starting of \mathbf{B}_1 . Alternatively, \mathbf{B}_3 and \mathbf{B}_4 triclusters appear to be related through a causal relation. From this observation, we can infer two cascades: R_1 with co-occurring modules $\{\mathbf{B}_1, \mathbf{B}_2\}$, and R_2 with precedent modules $\mathbf{B}_3 \Rightarrow \mathbf{B}_4$.

Def. 1.13 Let a sequence of itemsets $\langle I_{11} \dots I_{1n} \rangle$ be contained in another sequence of itemsets $\langle I_{21} \dots I_{2m} \rangle$ if $\exists_{1 \leq i_1 < \dots < i_n \leq m} : I_{11} \subseteq I_{2i_1}, \dots, I_{1n} \subseteq I_{2i_n}$. Given a sequential database \mathbf{A} with n itemset sequences (observations), a *sequential pattern* s is an itemset sequence contained in a significant subset of the observations. Given \mathbf{A} , the **sequential pattern mining** task aims to discover a set of sequential patterns satisfying specific criteria of statistical significance, discriminative power in the presence of labels, and dissimilarity between patterns.

Basics 1.10 Modeling sequential databases

Most of existing local descriptive models for sequential databases are a composition of temporal patterns. Some of these temporal patterns assume temporal contiguity (such as motifs and strings), while others exclude the possibility to model co-occurrences. Contrasting, local descriptive models given by sequential patterns flexibly capture frequent co-occurrences and precedences (Def.1.13). Given \mathbf{A}_4 database provided in Figure 1.7 and a strict criteria of significance by requiring patterns to be supported by all observations of a given class and to have a minimum number of 6 items, 2 sequential patterns can be retrieved: $s_1 = d(ab)(be)f$ and $s_2 = a(ac)d(df)$. s_1 is supported by all c_1 -conditional observations and s_2 by all c_2 -conditional observations. Both have 4 frequent precedences and an average number of 1.5 co-occurring items per itemset. Besides the significance criteria, these sequential patterns are dissimilar and appear to be discriminative as they are supported by observations of a single class only.

Classification Models

We now move from descriptive to prescriptive settings to be able to answer a wider range of learning tasks.

Def. 1.14 Given a set of labeled observations, a *decision model* is a mapping function between observations and classes, $M : X \rightarrow C$. A decision model is a **classification model** in the presence of categoric labels, $C = \Sigma$, and a *regression model* in the presence of numeric labels, $C = \mathbb{R}$.

Decision models are inferred from the conditional regularities of the input space, $P_{X|C}$.

Given a set of labeled observations (\mathbf{x}_i, c) , classifiers learn a mapping from X to C given by *decision rules* for labeling (unlabeled) observations. When observations are labeled with real values, we are in the presence of parameter estimation problems (also called point estimation problems), in which an unknown scalar quantity can be estimated using regression models.

Similarly to descriptive models, decision models can either be conceptually divided as global or local, depending on whether relations are inferred from the overall data space or from informative and discriminative regions.

Local Classification Models

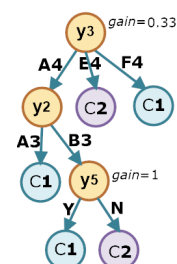
Def. 1.15 Given a set of labeled observations, a **local classification model** is a meaningful composition of decision rules inferred from regions of interest (*supervised local descriptive models*) to label new observations.

Similarly to the introduced descriptive models, the properties of the regions used within a local classifier highly depend on the input data. Illustrating, given an labeled real-valued matrix in a (n, m) -space, a decision rule $M_i(\mathbf{x})$ is inferred from a discriminative subspace \mathbf{B}_i , where \mathbf{B}_i is a (r_i, q_i) -space where $r_i < n \wedge q_i < m$. A widely known local classifier is a decision tree, where each path from the root to the leaf defines a decision rule associated with a region with specific interestingness criteria such as high information gain.

Basics 1.11 Learning decision trees

Given a set of labeled observations, a decision tree is a local classifier that gathers class decisions from conjunctions of tests on the values of discriminative features (see Figure 1.12). Decisions trees are commonly learned by iteratively selecting a feature with the highest information gain (or lowest entropy) and splitting data accordingly for subsequent branched decisions. Given a \mathbf{X} set of n observations with labels in C that are divided $\{\mathbf{X}^1, \dots, \mathbf{X}^{|\mathcal{Y}^j|}\}$ according to their values on feature \mathbf{y}_j , information gain is given by $H(\mathbf{X}) - \sum_{i=1}^{|\mathcal{Y}^j|} \frac{|\mathbf{X}^i|}{n} H(\mathbf{X}^i)$ where $H(\mathbf{X}) = -\sum_{c_i \in C} \frac{n_i}{n} \log_2(\frac{n_i}{n})$ is its entropy. Accordingly, Figure 1.12 shows the learned decision tree for the tabular dataset \mathbf{A}_2 in Figure 1.4. Understandably, paths from root to leaf form a region in the original dataset given by a subset of observations respecting the accepted values and the set of tested features.

Figure 1.12: Decision tree learned from \mathbf{A}_2 data.



Associative classification models are a specialization of local classification models. They define a set of weighted decision rules from informative and discriminative regions, thus combining simplicity and flexibility.

Def. 1.16 Given a labeled dataset \mathbf{A} , an *associative model* is a composition of p weighted association rules, where each rule $\mathbf{B} \Rightarrow^s C$ has s -weight and maps a region of interest \mathbf{B} (rule’s antecedent) with a subset of classes $C \subset C$ (rule’s consequent). Given \mathbf{A} , an **associative classification model** defines a matching criteria M to label a new observation \mathbf{x}_{new} against a (possibly pre-computed) associative model learned from \mathbf{A} .

Basics 1.12 Learning a simplistic associative classifier

Considering the \mathbf{A}_4 dataset provided in Figure 1.7, 4 regions of interest that can be retrieved and mapped as the following set of rules: $\mathbf{B}_1 \Rightarrow^{s_1} \{c_1\}$, $\mathbf{B}_2 \Rightarrow^{s_2} \{c_1\}$, $\mathbf{B}_3 \Rightarrow^{s_3} \{c_2\}$, $\mathbf{B}_4 \Rightarrow^{s_4} \{c_2\}$ (see *Basics 1-1.9*). From these rules, scores can be inferred s_i based for instance on the discriminative power and significance of each region ($s_1 > s_2 \wedge s_4 > s_3$). Given a new observation, the simplest way to compute the strength of each class is to sum the score of rules with matched regions. Assuming $s_1=2.4$, $s_2=1.6$, $s_3=1.4$, $s_4=3.2$, and that the decision model M considers a valid match between the newly observed region and a scored region if over 70% of the elements of these regions are approximately equal. Given the observation in Figure 1.13, we can see that 3 regions satisfy this criterion: \mathbf{B}_1 , \mathbf{B}_2 , and \mathbf{B}_3 , and thus the strength of each class is $c_1 = \frac{s_1+s_2}{s_1+s_2+s_3} = 74\%$ (the decision) and $c_2 = \frac{s_3}{s_1+s_2+s_3} = 26\%$.

Figure 1.13: New observation \mathbf{x}_{new} for \mathbf{A}_4 cube.

	t_1	t_2	t_3	t_4
y_1	-1	0	-2	-2
y_2	2	2	-1	-2
y_3	1	2	2	0
y_4	1	2	2	2
y_5	2	1	0	-1

Since the underlying goal of this work is to explore the impact of selecting relevant regions when learning from high-dimensional data, the focus is placed on local (descriptive and classification) models. In this context, we use the term *local model* – a composition of regions of interest – not only as the foundation for learning descriptors but also for learning decision rules.

1.1.3 Learning Function

The chosen learning method determines the properties of the learned models and thus their adequacy to answer a given problem. The quantification of the performance of descriptive and decision models is formally expressed via a loss function, L . For descriptive models, the loss function is typically given by: 1) match scores between the learned regularities against new observations or expectations (when assuming background knowledge regarding P_A) or, alternatively, by 2) merit functions that measure the coherency of the learned abstractions (in the absence of testing observations and background knowledge of P_A). For classification models, a loss function measures the incurred cost of erroneous decisions, $L(y, \hat{y})$. Although this quantification can be also seen optimistically as an utility function to measure the gain of correct decisions, its symmetric function defines a loss function and therefore we use the term loss function interchangeably. Illustrative loss functions for classification include functions based on accuracy or sensitivity metrics in the presence of nominal classes, and the normalized or root mean squared errors in the presence of ordinal classes.

Def. 1.17 Given a (labeled) dataset \mathbf{A} , a descriptive or classification *learning task* aims to learn a model M that achieves a specific optimality criterion with regards to a specific loss function L or set of loss functions.

The properties of the target learning models are mainly driven by the learning function. The learning function can either be *probabilistic* or *deterministic*, as well as *generative* or *discriminative*.

Probabilistic functions place assumptions on the stochasticity of observations to model their regularities P_A . This is done by either fitting observations against mixtures or more structured models (Basics 1.13). The learning task thus consists of estimating the parameters of these models, and it is thus often seen as an optimization problem. Contrasting, deterministic functions typically rely on greedy or exhaustive searches to extract relevant data aspects, which can be seen as an implicit model of the true data regularities P_A . Probabilistic and deterministic functions should not be confused with probabilistic and deterministic outputs of classification models. Given a set of labels C , the output or decision of a classifier is probabilistic when discloses the probability of labeling a given observation

for each class in C , and deterministic when simply returns the class with higher probability.

In labeled data contexts, these probabilistic and deterministic functions can either be used to learn generative or discriminative models. In generative contexts, each class-conditional probability function $P_{X|c}$ is learned separately. Generative learning needs to be sufficiently powerful to model regularities specific to a single class for data contexts with subtle differences between class-conditional functions. Discriminative learning aims to discover meaningful boundaries that separate observations from different classes. Discriminative functions can either be derived directly from data, such as decision trees, or from generative functions by focusing on their class-conditional differences (Basics 1.14). Illustrating, the target associative classifiers typically rely on the generative learning of class-conditional regions of interest. Yet, the desirable focus on dissimilar regions between classes also requires discriminative learning. Alternatively, some learning approaches combine generative and discriminative functions via generative embeddings [344, 386]. In the context of unlabeled data, generative and discriminative learning is respectively associated with the description and differentiation of specific regions of interest. However, to preserve simplicity, this work only applies these terms in labeled data contexts.

Basics 1.13 Probabilistic models: unstructured vs. structured, generative vs. discriminative

Given A_1 dataset, the learned multivariate Gaussian mixture: $\mu|_{c_1}=[0.5,-2.2,3,0.5,-2.3,0,1.3]$ and $\mu|_{c_2}=[-1,-0.1,0.2,1,0.1,-0.1,0]$ (see Basics 1.7) is an unstructured and generative probabilistic model. Given an observation x_{new} , the probability of x_{new} being generated by each class-conditional mixture (also referred as fit) determines the strength of each class. Nevertheless, as illustrated in Figure 1.14, these class-conditional mixtures can be use as input for a discriminative function to place decisions. Considering the same A_1 dataset, the learning function can consider more complex stochastic assumptions, possibly given by structured probabilistic models such as the model illustrated in Figure 1.15.

Figure 1.14: Discriminative decisions from multivariate Gaussian distributions learned from A_1 data.

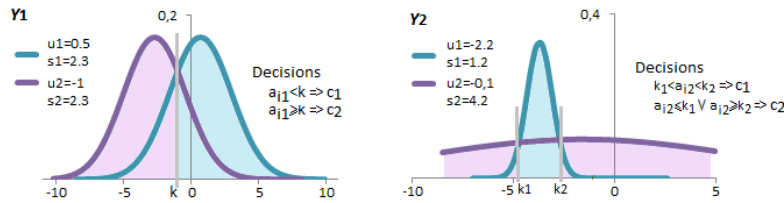


Figure 1.15: Structured generative model learned from A_1 .

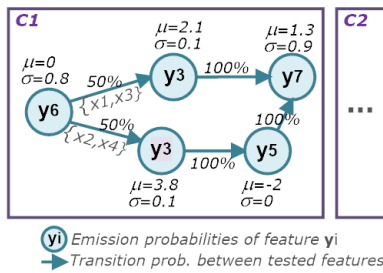
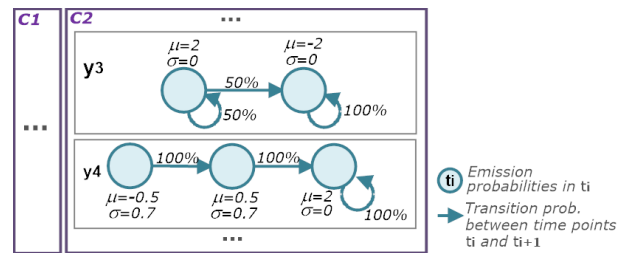


Figure 1.16: Structured generative model learned from A_4 .



Considering the A_4 three-way time series. Similarly, both unstructured and structured probabilistic functions can be learned from A_4 . In this data context, a multivariate Gaussian mixture can be learned with parameters dependent on time (e.g. $\mu(t)|_{c_1}=[\mu_{y_1}(t)|_{c_1}, \dots, \mu_{y_5}(t)|_{c_1}]=[0.38t^2-2.93t+3.63, -t+1.75, \dots, -0.25t^2+1.55t-1.75]$ when assuming a polynomial regression) or by matrices where the Normal assumption is considered for each time point (e.g. $\mu(t_1)|_{c_1}=[1.0,0.5,0,0,-0.5]$). For testing new observations, a generative fitting schema can be considered or discriminative decisions inferred from the underlying mixtures. Contrasting, an illustrative structured model learned from A_4 where temporal dependencies are explicitly modeled is depicted in Figure 1.16.

Basics 1.14 Deterministic models: unstructured vs. structured, generative vs. discriminative

Given A_2 dataset, the illustrated decision tree in Fig.1.12 is a structured and discriminative deterministic model. Given A_4 dataset, the illustrated associative classifier in Basics 1.12 given by a set of weighted rules is an unstructured and discriminative deterministic model. Given A_5 dataset, the c_1 -conditional $d(ab)(be)f$ and c_2 -conditional $a(ac)d(df)$ sequential patterns (see Basics 1.13) can be either seen as a generative model to test the fit of new observations or as a discriminative model when dissimilarity guarantees are provided. In fact, the associative models learned from A_4 and A_5 data require both generative and discriminative learning functions to, respectively, model class-conditional regularities and guarantee their discriminative power.

1.2 Problem Motivation

Figure 1.17 lists the major challenges and commonly applied principles to learn from high-dimensional data.

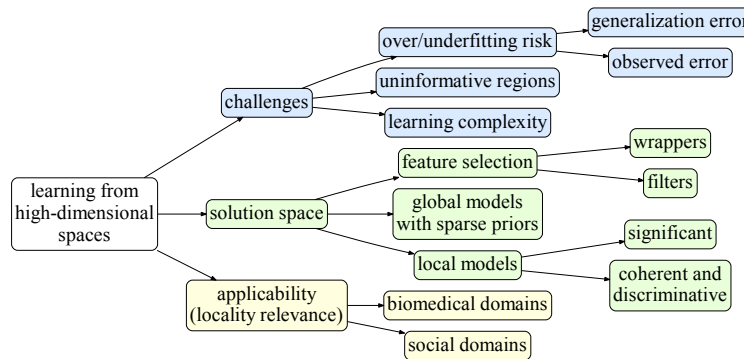


Figure 1.17: Motivating the learning in high-dimensional spaces: open challenges, contributions and applications.

A well-known challenge of learning from high-dimensional data is associated with the propensity of the resulting models to either overfit or underfit the observed data [646, 316, 173]. Minimizing this risk requires essentially an optimal trade-off between the *observed error* – error estimated from assessing the learned model on the observed data –, and the *generalization error*, often given by the mean and variability of the error estimates collected from assessing the model on new sets of observations [182]. In this context, adjusting the complexity (also referred capacity) of the learning function is necessary to achieve good generalization [26]. Complex functions guarantee a low observed error but often perform poorly on unseen data (overfitting propensity), while overly simple functions may not be able to model relevant data regularities (underfitting propensity).

However, in data contexts where the number of features exceeds the number of observations, the complexity term cannot be explored since models may not be able to generalize. This property is often referred as perfect overfitting towards the observed data [646]. To illustrate this problem, let us consider the following simplistic global model: a linear hyperplane $M(\mathbf{x})$ in \mathbb{R}^m defined by a vector $\mathbf{w} \in \mathbb{R}^m$ and point b to either separate two classes, $\text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$, predict a real-valued outcome, $\mathbf{w} \cdot \mathbf{x} + b$, or describe the input observations, $\mathbf{X} \sim \mathbf{w} \cdot \mathbf{x} + b$. As illustrated in Figure 1.18, a linear hyperplane in \mathbb{R}^m can perfectly model up to $m + 1$ observations, either as a global classifier $\mathcal{X} \rightarrow \{\pm 1\}$, as a regression model $\mathcal{X} \rightarrow \mathbb{R}$ or as a global descriptive model of \mathbf{X} . Although the assessment of these models using the same observations is associated with a zero observed error, in the presence of new observations, the generalization error can be significantly high due to the risk of perfect overfitting towards the training data.

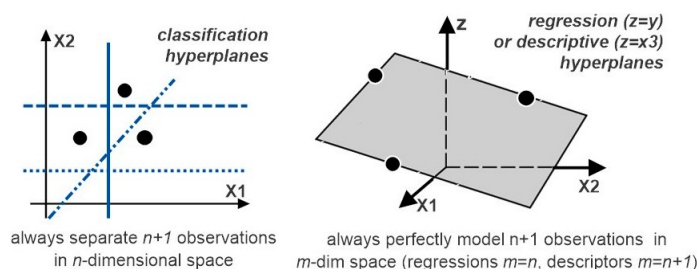


Figure 1.18: Linear hyperplanes cannot generalize when the number of features (data dimensionality) is larger than the number of observations (data size), $m \geq n + 1$.

As a result of these observations, learning from specific regions of interest from high-dimensional data has been presented as an option to avoid perfect overfitting. However, assessing the statistical impact of selecting regions is critical since small regions are highly prone to be relevant by chance [343].

In tabular data contexts, regions given by a small number of features and/or observations can be highly coherent (descriptive tasks) or highly discriminative (classification tasks), yet their probability of occurrence might not be statistical significant (see Basics 1.15). This observation is also valid in structured data contexts, where a region is additionally associated with a subset of time points, item occurrences or events.

Basics 1.15 Statistical significance of regions

In tabular data contexts, a *region* is a subset of observations and features with a specific form of homogeneity. Given a tabular dataset \mathbf{A} , the *statistical significance* of a region defines the probability of its occurrence against a null data model to deviate from expectations. In this context, we use the term region to either describe an *observed region* in the data space, as well as an *unobserved region* (whose statistical significance can be also assessed). Given these considerations, the probability of occurrence of an observed region is non-necessarily 100% since this probability is computed against data expectations. According to Figure 1.1, we identify three red regions as not statistically significant. This is a likely condition since their low number of observations and features increases their probability to occur. *Book V* is dedicated to adequately assessing the statistical significance of regions with varying properties.

Illustrating, consider a real-valued matrix defining a $(n=50, m=10000)$ -space with an Uniform distribution of values per feature $y_j \sim U(-1, 1)$ and two balanced classes. Consider a region given by a subset of the original features, $\mathbf{B}=(\mathbf{I}=\mathbf{X}, \mathbf{J} \subseteq \mathbf{Y})$, defining a $(50, 5)$ -space. The combination of values for the selected subset of 5 features (assuming an entropy ratio above 90% according to Basics 1.11) is highly likely to occur by chance and therefore this region is not statistically significant. Alternatively, let us neglect the labels and consider a $(n=20, m=10)$ -space. The probability that this region $\mathbf{B}=(\mathbf{I}, \mathbf{J})$ has constant values $\forall_{i \in \mathbf{I}} \forall_{j \in \mathbf{J}} a_{ij} \in [0.2, 1]$ across observations is 44% assuming a simplistic binomial calculus, $\binom{m}{|\mathbf{J}|} \sum_{x=|\mathbf{I}|}^n \binom{n}{x} p_{\varphi_{\mathbf{B}}}^x (1 - p_{\varphi_{\mathbf{B}}})^{n-x}$. Again, this region is not statistically significant.

Although the selection of small regions is highly prone to be either informative or discriminative by chance, many classifiers: 1) rely on feature selection to deal with high-dimensionality, or 2) infer decisions from regions given by (possibly small) subsets of features. Illustrative classifiers with propensity towards this behavior are decision trees. Decision trees (see Fig.1.12) typically select a minimum subset of features, whose combination of values is able to discriminate a specific class. As a result, decision trees and peer classifiers show a high variable performance (when assessed from a collection of error estimates) and generalization error.

Understandably, the selection of non-significant regions is associated with the risk of underfitting the observed data. This is the tackled problem in this work since this risk is not structural, meaning that it can be minimized. For this aim, the impact of mapping an original data space into a set of regions needs to be addressed.

In addition to this problem, the selection of uninformative regions increases the learning complexity and can introduce unnecessary biases on the learned models.

1.2.1 Problems of Dimensionality Reduction

Let us further explore the facets of this problem. Three major learning options have been considered for the learning of descriptive and classification models from high-dimensional data.

First, *feature selection* methods have been applied as a filter or a wrapper. Filters select subsets of features as an independent preprocessing stage according to some measure of feature relevance, which often neglects the statistical significance of the selected spaces [682]. Wrappers can be alternatively applied to minimize this problem since they can estimate the generalization error of the model by evaluating the learned model against multiple subsets of features [216, 558]. However, minimizing the generalization error does not guarantee that the selected subsets of features are statistically significant. Additionally, wrappers degrade the learning efficiency and are dependent on the chosen model, that is, there are no guarantees that a subset of features chosen for one model is adequate for other models.

Pointers 1.16 The problem of feature extraction

A rather less-studied problem is associated with the learning from tabular data spaces with features extracted from structured data spaces (e.g. physiological signals, repositories of health-records). In order to guarantee that a reasonable set of informative features is extracted, existing studies tend to generate a wide-range of statistical, temporal and geometric features^a [285, 397, 354]. Understandably, in the presence of a limited number of observations, an informative feature can easily be discriminative by chance. Furthermore, this problem is aggravated for feature extraction due to biases towards the extraction of purely discriminative features.

^aIllustrative methods for feature extraction from temporal structured data include rectangular tonic-phasic windows; moving and sliding features (as moving and sliding mean and median); transformations (Fourier, wavelet, empirical, Hilbert, singular-spectrum); principal, independent and linear component analysis; projection pursuit; nonlinear auto-associative networks; multidimensional scaling; and self-organizing maps.

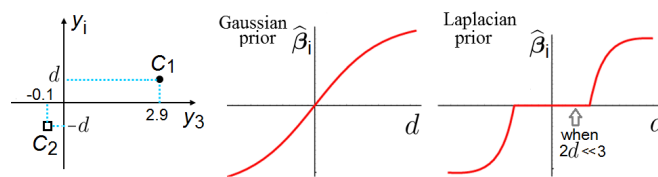
In real-valued data contexts, an alternative simplistic way of reducing the dimensionality of a given dataset is to use a mapping function, also referred as a projection or hyper-dimensional transformation, from the observed data space into a new data space with lower dimensionality $\phi : \mathbb{R}^m \rightarrow \mathbb{R}^d$ where $d < m$. An illustrative projection of \mathbf{A}_1 data space (Table 1.3) is $\phi(\mathbf{x}_i) = \phi(a_{i1}, \dots, a_{im}) = (a_{i3}, a_{i5}, 2a_{i7}, a_{i1} \times a_{i6})$. Contrasting with feature selection, projections can affect the value distributions of features, thus often facilitating the subsequent learning task. However, even in the presence of complex mapping functions, these procedures are not able to flexibly select an arbitrary number of regions from the input data space.

Second, and complementarily to feature selection, global models can be learned using *sparse kernels*. A sparse kernel is a parametric learning function that it is able to guarantee a focus on relevant regions by placing assumptions to collapse or disregard parameters associated with uninformative regions, thus minimizing the learning complexity and fostering the model's generalization [153, 647]. Sparse kernels are often associated with (but not limited to) the learning of probabilistic models [215]. In these contexts, irrelevant and redundant parameters rapidly converge to zero [378, 214]. For this end, specific a priori knowledge regarding the probability function P_A , referred as prior, have been used to promote sparsity of both unstructured and structured models for both descriptive and classification tasks (see Basics 1.17). Although the covered sparse kernels offer the possibility to discard non-informative data and to balance the over/underfitting by controlling the number of iterations associated with the learning of a parametric model, they show some inherent challenges. Since sparsity is determined by the model's parameters, it is not expressive enough to guarantee a flexible selection of regions of interest due to two major challenges. First, although recent contributions can be used to avoid the need to specify or estimate the degree of sparseness of the models [217], there is still a high complexity associated with the definition of sparse priors. Second, sparsity is primarily used to either discard non-relevant features and/or specific ranges of values per feature, thus preventing the flexible selection of subsets of both observations and features/events/time points.

Basics 1.17 Illustrative discriminative and generative sparse kernels

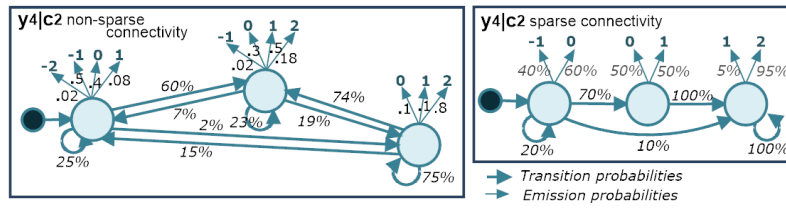
Let a support vector machine be a parametric learning function aiming to learn a hyperplane from a given feature space: $M(\mathbf{x}) = \sum_{i=1}^n w_i x_i + b = \mathbf{w}^T \mathbf{x} + b$ where m is the dimensionality and \mathbf{w} is the vector of parameters. The hyperplane can be learned to approximate observations (description and regression) or, alternatively, to separate observations with distinct labels (classification). The learning function can be applied over the original or projected feature spaces. Since not all features are equally informative, sparsity is used to guarantee that less informative features are discarded by placing a loss assumption that forces some of \mathbf{w} elements to converge to zero [624]. This assumption guarantees the learning of observation models with lower generalization error and the learning of classification models with less propensity towards overfitting (hyperplane with larger margins separating observations). Given the c_1 -conditional observations from \mathbf{A}_1 (Table 1.3), $\mathbf{w} = [0.2, -1.3, 1.7, 0.2, -1.4, -0.01, 0.7]$ defines a descriptive hyperplane where $\{w_1, w_4, w_6\}$ converge to zero with sparsity enforcement. Given the same set of c_1 -conditional observations, now consider the learning of a mixture given by a vector of parameters \mathbf{w} applied over the original space ($\mathbf{w} = [w_1, \dots, w_m]$) or projected feature space ($\mathbf{w} = [w_1, \dots, w_p]$ where $p \in \mathbb{N}^+$). The mixture illustrated in Basics 1.7 placed a Gaussian assumption, where $\mathbf{w} = [\mu|c_1, \sigma|c_1]$. Assuming a Laplacian (rather than a Gaussian) prior, \mathbf{w} can be optimized to yield a maximum posterior estimate with certain sparseness degree by removing irrelevant and redundant parameters. The Laplacian prior sets estimates as 0 when they are non-discriminative (Fig.1.19). For the given c_1 -conditional observations, y_4 is non-discriminative and y_2 is redundant with y_5 , thus $w_4 = w_2 = 0$.

Figure 1.19: Laplacian priors to discard non-informative features from \mathbf{A}_1 dataset (charts adapted from Figueiredo [213]).



Contrasting with the previous models, now consider a structured model given by an automaton with fully-interconnected transitions and certain emissions per states. In this context, a parametric learning function can be given to learn the probabilities associated with the transitions and emissions of the underlying automaton. By placing assumptions that enforce the delineate convergence of the probability of transitions and emissions, sparse models (low number of high probable paths) can be learned. Given the y_3 feature of c_2 -conditional times series from \mathbf{A}_4 (including $s_1 = \langle 0, 1, 2, 2 \rangle$ and $s_2 = \langle -1, 0, 2, 2 \rangle$), Fig.1.20 illustrates structured models in the absence and presence of sparse priors (assuming a mixture of Dirichlets [95]).

Figure 1.20: Impact of sparse kernels to enforce path convergence of structured models learned from A_4 time series.



Pointers 1.18 Complementary readings on sparse kernels

Sparse kernels can be considered when learning both descriptive and decision models [216, 78]. Sparsity can be accommodated for unstructured generative models given by mixtures with varying properties [254, 217] or more structured generative models such as hidden Markov model, dynamic Bayesian networks or neural networks by placing assumption on the underlying lattice connectivity [143, 95]. Well-known ways of obtaining sparse global models include parametric functions with a Laplacian prior [232, 504, 384] or support vector machines [254, 153]. Other illustrative sparse kernels include multinomial sparse logistic regressions [78], variants of the expectation-maximization algorithm with sparse priors [505, 215], mixtures of Dirichlets [95], among others [378, 214]. In order to avoid the need to specify or estimate the degree of sparseness of the resulting models, a hierarchical interpretation of the Laplacian prior has been applied with the Jeffreys' hyper-prior [217].

Third, some learning functions infer descriptions and decisions from sets of regions of the original data space. These functions are associated with local descriptive models, such as biclustering models, and local decision models, such as associative classification models. The problem of how to guide the learning of these models to adequately select and compose regions of interest is the central task of this work. Naturally, the implications of this study can be further used to assess and extend methods for dimensionality reduction (including but not limited to feature selection) as well as to affect the learning of global models.

1.2.2 Relevance of Local Models: Applications

To further motivate the relevance of learning local descriptive models, Table 1.1 provides a set of biomedical and social data domains characterized by the presence of meaningful local regions. These data domains are characterized by a high-dimensionality associated with a high number of genes per sample, health-records per patient, molecules per biological network, time points per physiological signal, browsing actions per user, trading decisions per business, or interactions per user in social contexts.

	Data	Illustrative subspaces with relevance for learning tasks
Biomedical	physiological [108, 201, 211]	Sets of (sliding) features and signal partitions with coherent values across case or stimuli-elicited responses.
	clinical [302, 122]	Groups of patients with correlated clinical features or health records (shared treatments, diagnoses, prescriptions).
	structural variations [207, 324]	Correlated groups of mutations and copy number variations.
	biological networks [53]	Modules of genes, proteins or metabolites with meaningful interaction (from matrices with pairwise connections).
	gene expression [429, 312]	Groups of genes involved in functional processes and pathways only active under certain conditions.
	genome-wide [662, 640]	Conserved functional subsequences (sequence alignments), factor binding sites and insertion mutagenesis.
	other	Local regularities in translational [175], chemical [415] and nutritional data [393].
Social	social networks [257]	Groups of individuals with correlated activity and intercommunication; groups of contents based on accessors.
	text mining [38, 171]	Content-related documents and web pages (from matrices weighting categories/words across text segments).
	(e-)commerce [35]	Hidden browsing patterns containing relationships between sets of (web) users, (web) pages and operations.
	financial trading [334]	Indicators producing similar profitability for specific trading points (buy, hold and sell signals) in the stock market.
	collaborative filtering [159]	Groups of users who share preferences and behavioral patterns for a subset of available actions.

Table 1.1: Disclosing the meaning of regions across (high-dimensional) biomedical and social data contexts.

1.3 Thesis Requirements

The underlying **hypothesis** is that *learning from relevant regions of high-dimensional data improves the performance guarantees of local descriptive models and (associative) classification models*. Naturally, testing this hypothesis leads us into the *how*. First, how does performance vary with the properties of the selected regions? Second, how can

this understanding be used to improve the learning of descriptive and classification models? Figure 1.21 lists the key requirements and premises to validate the target hypothesis.

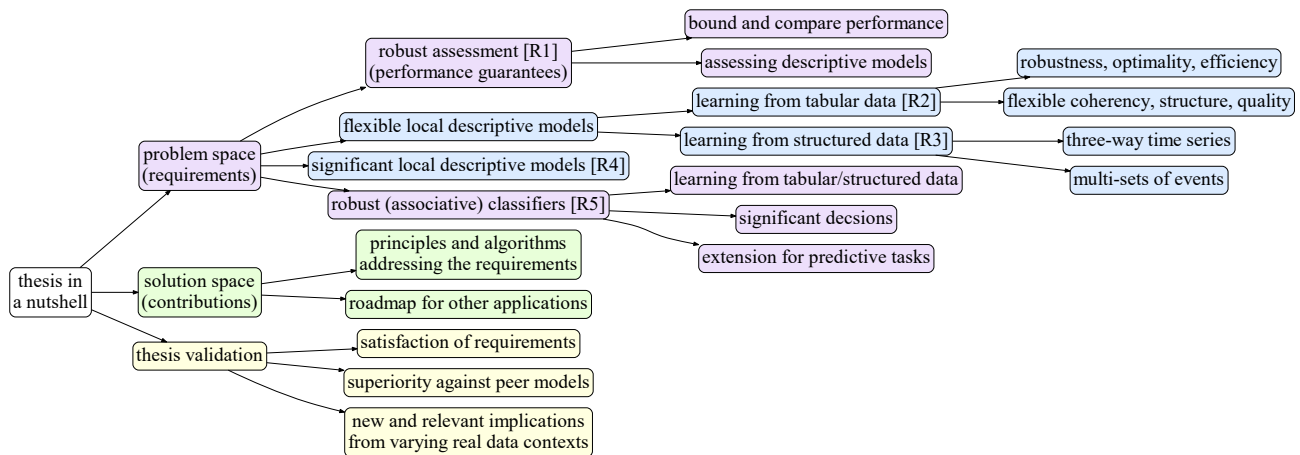


Figure 1.21: Structured view of the thesis scope: requirements and premises to validate the underlying hypothesis.

First, in order to validate the proposed hypothesis, we decompose its assertion according to an incremental set of five major requirements. These requirements define the problem space.

R1 Robust assessment of descriptive and classification models learned from high-dimensional data.

By satisfying the first requirement, we have a systematic way to validate our hypothesis, that is, to measure and compare the impact of modeling regions with varying properties of interest on the target learning tasks.

R2 Learning of flexible and robust local descriptive models from tabular data.

The satisfaction of this requirement allows the systematic exploration of the impact that distinct biclustering models have in the ability to learn from high-dimensional data. This requires the scalable discovery of flexible structures of biclusters with parameterizable homogeneity criteria, yet offering optimality guarantees to properly assess their impact on descriptive and prescriptive tasks.

R3 Learning of flexible and robust local descriptive models from structured data.

Although the satisfaction of **R2** already covers different high-dimensional data contexts, such as matrices and network data, it excludes other data structures that are becoming increasingly relevant, such as multivariate time series, sequential databases and multi-sets of events. These learning challenges are thus specifically addressed under this requirement.

R4 Guarantee the statistical significance of local descriptive models.

To answer the introduced need to assess the impact of reducing dimensionality or selecting regions of interest (Section 1.2), we require the target local descriptive models to be statistically significant. Addressing this requirement implies the presence of a robust statistical assessment to guarantee that regions with varying coherence and quality (either from tabular or structured data) are not prone to occur by chance. This allows the inference of constraints based on the properties of these regions and the original data, that can be used to guide the learning.

R5 Learning effective classifiers from flexible, robust and statistically significant local descriptive models.

This requirement combines the previous focus on local descriptive models with the need to guarantee their discriminative power in labeled data contexts. Its satisfaction allows the assessment of the impact that the coherency, quality, significance and discriminative power of the selected regions have in the performance of classification models. The significance assessment of descriptive models is also extended for (associative) classification models and

used to affect the learning. All the previous contributions are thus used at this point to guarantee both the accuracy and statistical significance of classification decisions. As a result, an integrative view of the pros and cons of learning local descriptive models to perform classification from distinct high-dimensional data domains is required.

Finally, this requirement is further extended in this work to guarantee the ability to learn from structured codomains given by sequences of classes for the adequate answering of predictive tasks.

Table 1.2 provides a non-exhaustive decomposition of these five structural requirements.

Table 1.2: Decomposition of the five requirements: list of the tackled requirements.

<i>Requirement</i>
R1: Robust assessment of models learned from high-dimensional data; R1.1: Performance guarantees of classification models; R1.2: Performance guarantees of local descriptive models; R1.3: Adequate generation of synthetic data for non-biased and complete assessments;
R2: Learning biclustering models from tabular data; R2.1: Flexible structures of biclusters with optimality guarantees; R2.2: Biclustering models with varying coherency: additive, multiplicative, plaid and order-preserving models; R2.3: Robustness of biclustering models to: 1) different forms of noise, 2) discretization and 3) missings; R2.4: Scalability of biclustering searches (with optimality guarantees); R2.5: Extension of contributions towards network data; R2.6: Effective and efficient learning in the presence of background knowledge; R2.7: Sound integration of previous contributions;
R3: Learning local descriptive models from structured data; R3.1: Learning cascade models from three-way time series; R3.2: Learning arrangements of events from multi-sets of events; R3.3: Stochastic modeling of structured data;
R4: Guarantee the significance of flexible local descriptive models; R4.1/4: Robust assessment of the statistical significance of discrete and real-valued biclusters; R4.2: Robust assessment of additive, multiplicative, symmetric, order-preserving and plaid models; R4.3/7: Robust assessment of regions with arbitrary-high levels of noisy and missings; R4.5: Robust assessment of cascades and arrangements of events from structured data; R4.6: Learning of local descriptive models from previous statistical views;
R5: Learning accurate and significant classification models from high-dimensional data; R5.1: Learning effective associative classifiers from tabular data; R5.2: Learning effective associative classifiers from structured data contexts; R5.3: Learning classifiers with guarantees of statistical significance; R5.4: Multi-period classification: extending previous contributions for the learning of sequences of classes;

Given the formulation of these requirements, our work becomes a matter of testing whether they can be simultaneously satisfied (solution space), and whether their satisfaction is associated with an improved learning in high-dimensional spaces. The thesis statement is thus asserted upon the verification of the three following sub-hypotheses: 1) the proposed learning models satisfy the introduced requirements, 2) these learning models offer distinctive behavior of interest against state-of-the-art learners, and 3) their application across real-world data domains can be used to unravel new, meaningful and significant relations from data.

1.4 Solution Space

Multiple contributions resulted from addressing the introduced requirement. Contributions take two major forms: 1) principles, and 2) algorithms and (assessment) methodologies that rely on one or more principles. Many of these contributions are not only relevant to tackle the target problem, but can also be applied to answer other problems. In order to not compromise the line of focus of this work, the applicability of our contributions to other problems are properly identified along the text using dedicated frames (see *Notation*). As we address and answer each requirement, we expect that the proposed research produces the following scientific contributions:

- C1.** Methods to bound and compare the performance of local descriptors and classifiers in high-dimensional data contexts, including adequate loss functions (able to measure the impact of selecting regions with varying properties), robust error estimators and generators of data for non-biased and complete assessments;
- C2.** New descriptors of tabular data able to efficiently discover flexible structures of biclusters with optimality guarantees and robustness to varying forms of noise. Algorithms to retrieve non-constant coherencies, such as plaid and order-preserving models; to guarantee an adequate analysis of varying forms of tabular data (including network data); and to effectively incorporate background knowledge;

- C3.** Structured view on the increasingly relevant problems of learning cascades models from three-way time series and arrangements of events from multi-sets of events. Principles to handle the inherent complexity and variability of local responses in these data contexts, combining temporal and cross-attribute views with (possible) misalignments across observations. Deterministic and stochastic algorithms integrating these principles;
- C4.** Statistical views to robustly assess the significance of regions from tabular and structured data with regards to their coherency, quality and size (with upper limits on the risk of false discoveries). Revised algorithms to combine homogeneity (C2-C3) and significance (C4) views for guiding the learning;
- C5.** Principles for an adequate discovery (C2-C4), composition, scoring and testing of (informative and discriminative) regions from tabular and structured data. New associative classifiers able to incorporate previous principles. Principles to assess and promote the statistical significance of classification decisions. Systematic analysis of the performance impact of varying the properties of the underlying regions and learning functions across data domains. Extension of the proposed classifiers, preserving the accuracy and significance of the proposed learning functions (C5), to learn sequences of classes for predictive tasks.

Transversally to these set of major contributions, we additionally: 1) survey the contributions and limitations of state-of-the-art methods, and experimentally compare them against the proposed methods; and 2) show the relevance of the learned models to unravel significant and non-trivial relations across data domains, with a particular incidence on biomedical domains.

An integrative view of the proposed contributions of this thesis is provided in *Chapter VII-1*.

1.4.1 Scientific Dissemination

According to the introduced groups of requirements, we list below the current status of the dissemination of the contributions from our thesis near the scientific community. This list contains only peer-reviewed publications, excluding other forms of dissemination, such as invited speeches, tutoring and teaching activities, collaborations in international projects, and scientific meetings and symposiums. From the listed articles, we highlight two publications in the *Data Mining and Knowledge Discovery* journal (one dedicated to a subset of **C3** and **C5** contributions, and the remaining to a subset of **C5** contributions) and additional publications in *Pattern Recognition*, *BMC Bioinformatics*, *IEEE Transactions in Computational Biology and Bioinformatics*, and *Algorithms for Molecular Biology* journals dedicated to disseminate **C2** contributions. Table 1.22 lists some of the publications proposed in the context of this thesis (see Appendix for additional published work).

1.5 Contents

The dissertation document is organized as a set of books. *Books II to VI* expose the core contributions of our thesis, each book tackling one of the introduced requirements. The contents within each book are carefully discussed at their start. A book is organized in chapters. Each chapter addresses a finer requirement and delivers a compact set of contributions that become available for the following chapters and books. In this way, contents are incrementally built upon previous contents, until we are able to test the cogency of the target hypothesis.

Figure 1.23 provides an illustrative view on the dependencies between books. This view supports a sound navigation through the contents provided in this dissertation.

Book II defines an assessment methodology to validate subsequent contributions. The new methods for learning flexible local descriptive models from tabular data contexts proposed in *Book III* are extended in *Book IV* towards structured data contexts, and combined in *Book V* with guarantees of statistical significance.

Book VI proposes classifiers based on the previous models and tackles the problem of guaranteeing the statistical significance of their decisions.

Finally, *Book VII* discusses the conditions on which the thesis statement is satisfied, provides an integrative view of the proposed contributions, and summarizes their major implications.

Accepted and under revision publications per book	State (July 2015)	Tackled requirements
Book II Performance Guarantees of Models Learned from High-Dimensional Data		
P1.1: R Henriques and SC Madeira, Towards Robust Performance Guarantees for Models Learned from High Dimensional Data, 2015, Chap.3, Big Data in Complex Systems, Vol.9, Studies in Big Data Series, 71-104, Springer;	Accepted	R1.1,R1.2
P1.2: BiGen: Synthetic Data Generation for Biclustering;	Under revision	R1.3
Book III Learning Local Descriptive Models from Tabular Data		
P2.1: R Henriques, C Antunes and SC Madeira, A Structured View on Pattern Mining-based Biclustering, 2015, Pattern Recognition, Elsevier;	Accepted	R2.1,R1.2
P2.2: R Henriques and SC Madeira, BicPAM: Pattern-based biclustering for biomedical data analysis, 2014, 9(1):27- , Algorithms for Molecular Biology, BioMed Central Ltd;	Accepted	R2.2,R2.3
P2.3: R Henriques and SC Madeira, Biclustering with Flexible Plaid Models to Unravel Interactions between Biological Processes, 2015, IEEE/ACM Transactions on Computational Biology and Bioinformatics;	Accepted	R2.2
P2.4: R Henriques and SC Madeira, BicSPAM: Flexible Biclustering using Sequential Patterns, 2014, BMC Bioinformatics, 15:130, BioMed Central Ltd;	Accepted	R2.2,R2.3
P2.5: R Henriques, SC Madeira and Cláudia Antunes, 2013, F2G: Efficient Discovery of Full-Patterns, In ECML/P-KDD IW on New Frontiers to Mine Complex Patterns, Springer-Verlag, Prague, Czech Republic;	Accepted	R2.4
P2.6: R Henriques, C Antunes and SC Madeira, Methods for the Efficient Discovery of Large Item-Indexable Sequential Patterns, 2014, Lecture Notes in Computer Science, 100-116, Springer I.P.;	Accepted	R2.4
P2.7: R Henriques and SC Madeira, BicNET: Efficient Biclustering of Biological Networks to Unravel Non-Trivial Modules, 2015, In Algorithms in Bioinformatics (WABI), LNCS Series, Springer-Verlag, Atlanta, GA, US;	Accepted	R2.5
P2.8: BicPAMS: Software for Biomedical Data Analysis using Integrative Pattern-based Biclustering;	Under revision	R2.7
P2.9: R Henriques and SC Madeira, Pattern-based Biclustering with Constraints for Gene Expression Data Analysis, 2015, In Computational Methods in Bioinformatics and Systems Biology (EPIA-CMBSB), LNAI Series, Springer;	Accepted	R2.6
Book IV Learning Local Descriptive Models from Structured Data		
P3.1: Modeling Regulatory Cascades from Gene Expression Multivariate Time Series;	Under revision	R3.1
P3.2: R Henriques, C Antunes and SC Madeira, Generative Modeling of Repositories of Health Records for Predictive Tasks, 2015, 29(4):999-1032, Data Mining and Knowledge Discovery, Springer US;	Accepted	R3.2,R3.3
Book V Significance Guarantees of Local Descriptive Models		
P4.1: Assessing the Statistical Significance of Flexible Biclustering Solutions;	Under revision	R4
Book VI Learning Effective Classifiers from Local Descriptive Models		
P5.1: Learning classifiers from high-dimensional data using discriminative biclusters with non-constant coherencies;	Under revision	R5.1
P5.2: Impact of modeling statistically significant regions in the performance of classifiers;	Under revision	R5.3
P5.3: R Henriques, C Antunes and SC Madeira, Generative Modeling of Repositories of Health Records for Predictive Tasks, 2015, 29(4):999-1032, Data Mining and Knowledge Discovery, Springer US;	Accepted	R5.2
P5.4: R Henriques, SC Madeira and C Antunes, Multi-period Classification: Learning Sequent Classes from Temporal Domains, 2015, 29(3):792-819, Data Mining and Knowledge Discovery, Springer US;	Accepted	R5.4

Figure 1.22: State of scientific publications made in the context of this dissertation on July 2015.

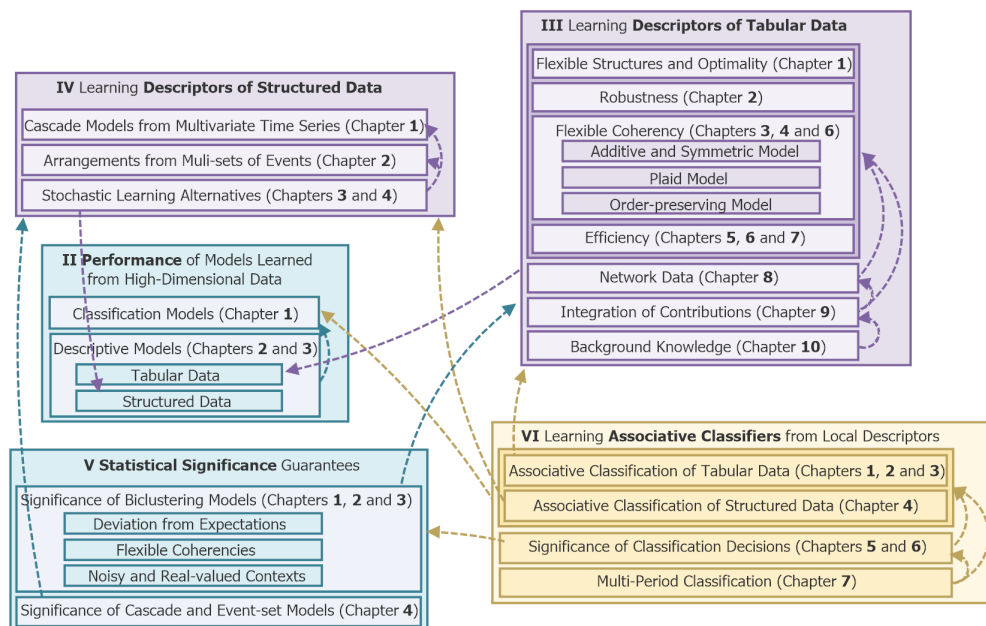


Figure 1.23: Thesis storyline: cohesive books of contents and their dependencies.