Lecture 3: Probability and Information

Andreas Wichert
Department of Computer Science and Engineering
Técnico Lisboa

- A key concept in the field in machine learning is that of uncertainty
  - Through noise on measurements
  - Through the finite size of data sets

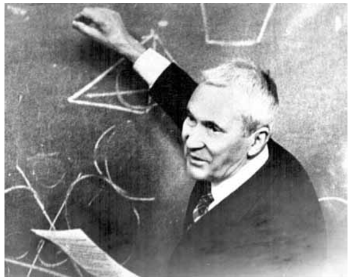
- Probability theory provides a consistent framework for the quantification and manipulation of uncertainty
- Forms one of the central foundations for pattern recognition.

Kolmogorov's Axioms of Probability (1933)

- To each sentence *a*, a numerical degree of belief between *0* and *1* is assigned

$$0 \leq p(a) \leq 1$$

$$p(true)=1, \; p(false)=0$$



- The probability of disjunction is given by

$$p(a \vee b) = p(a) + p(b) - p(a \wedge b)$$

Where do these numerical degrees of belief come from?

- Humans can *believe* in a subjective viewpoint from *experience*. This approach is called **Bayesian**
- For a finite sample we can estimate the true fraction. We count the *frequency* of an event in a *sample*. We do not know the true value because we cannot access the whole population of events. This approach is called f**requentist**
- From the true nature of the universe, for example, for a fair coin, the probability of heads is 0.5. This approach is related to the **Platonic world** of ideas. However, we can never verify whether a fair coin exists

- From the frequentist approach, one can determine the probability of an event a by counting
- If *Ω* is the set of all possible events, *p(Ω) = 1*, then *a* ∈ *Ω*.
- *card(Ω)* is the number of elements of the set *Ω*, *card(a)* is the number of elements of the set *a* and

$$p(a) = \frac{card(a)}{card(\Omega)}$$

$$p(a \wedge b) = \frac{card(a \wedge b)}{card(\Omega)}$$

- Now we can define the posterior probability, the probability of a after the evidence *b* is obtained

$$p(a|b) = \frac{card(a \wedge b)}{card(b)}$$

- using

$$p(a \wedge b) = \frac{card(a \wedge b)}{card(\Omega)}$$

- we get

$$p(a|b) = \frac{p(a \wedge b)}{p(b)} \qquad p(b|a) = \frac{p(a \wedge b)}{p(a)}$$

Bayes' Rule

$$p(a|b) = \frac{p(a \wedge b)}{p(b)} \qquad p(b|a) = \frac{p(a \wedge b)}{p(a)}$$

- The Bayes' rule follows from both equations

$$p(b|a) = \frac{p(a|b) \cdot p(b)}{p(a)}$$

Law of Total Probability

- For mutually exclusive events $b_1, \ldots, b_n$ with

$$\sum_{i=1}^{n} p(b_i) = 1$$

- the law of total probability is represented by

$$p(a) = \sum_{i=1}^{n} p(a) \wedge p(b_i) = \sum_{i=1}^{n} p(a, b_i)$$

$$p(a) = \sum_{i=1}^{n} p(a|b_i) \cdot p(b_i)$$

The Rules of Probability

Sum Rule	$p(X) = \sum_Y p(X,Y)$
Product Rule	$p(X,Y) = p(Y\|X)p(X)$

Bayes' rule

- Bayes rule can be used to determine the prior total probability *$p(h_\eta)$* of hypothesis *h_η* to given data *D*.
- For example, what is the probability that some illness is present?

$$p(h_\eta|D) = \frac{p(D|h_\eta) \cdot p(h_\eta)}{p(D)}$$

  - *$p(D|h_\eta)$* is the probability that a hypothesis *h_η* generates the data *D*
    - can be easily estimated
    - For example, what is the probability that some illness generates some symptoms?
  - The probability that an illness is present given certain symptoms, can be then determined by the Bayes rule

Maximum a Posteriori (MAP) Hypothesis

- The most probable hypothesis h_η out of a set of possible hypothesis $h_1, h_2, \cdots$ given some present data is according to the Bayes rule
- To determine the maximum a posteriori hypothesis h_{MAP} we maximize

$$h_{MAP} = \arg\max_{h_\eta} \frac{p(D|h_\eta) \cdot p(h_\eta)}{p(D)}$$

- The maximisation is independent of $p(D)$, it follows

$$h_{MAP} = \arg\max_{h_\eta} p(D|h_\eta) \cdot p(h_\eta)$$

posterior $\propto$ *likelihood* $\times$ *prior*

Maximum Likelihood (ML) Hypothesis

- If assume $p(h_\eta) = p(h_\gamma)$ for all h_η and h_γ, then can further simplify, and choose the maximum likelihood (ML) hypothesis

$$h_{ML} = \arg\max_{h_\eta} p(D|h_\eta)$$

Bayesian Learning

$$p(\mathbf{w}|D) = \frac{p(D|\mathbf{w}) \cdot p(\mathbf{w})}{p(D)}$$

- *p(D|**w**)* is evaluated on the observed data set *D* and is called likelihood function.
  It indicates how probable the observed data set is for different settings of ***w***.
- Given likelihood we can state: *posterior ∝ likelihood × prior*
  - *According to linear relation*

Example

- Does patient have cancer or not?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result (+) in only 98% of the cases in which the disease is actually present, and a correct negative result (-) in only 97% of the cases in which the disease is not present

Furthermore, 0.008 of the entire population have this cancer

Suppose a positive result (+) is returned...

$$P(cancer) = 0.008 \qquad P(\neg cancer) = 0.992$$

$$P(+|cancer) = 0.98 \qquad P(-|cancer) = 0.02$$

$$P(+|\neg cancer) = 0.03 \qquad P(-|\neg cancer) = 0.97$$

$$P(+|cancer) \cdot P(cancer) = 0.98 \cdot 0.008 = 0.0078$$

$$P(+|\neg cancer) \cdot P(\neg cancer) = 0.03 \cdot 0.992 = 0.0298$$

$$h_{MAP} = \neg cancer$$

Normalization

$$P(cancer \mid +) = \frac{0.0078}{0.0078 + 0.0298} = 0.20745$$

$$P(\neg cancer \mid +) = \frac{0.0298}{0.0078 + 0.0298} = 0.79255$$

- The result of Bayesian inference depends strongly on the prior probabilities, which must be available in order to apply the method

Naive Bayes Classifier

- Along with decision trees, neural networks, nearest neighbor, one of the most practical learning methods
- When to use:
  - Moderate or large training set available
  - Attributes that describe instances are conditionally independent given classification
- Successful applications:
  - Diagnosis
  - Classifying text documents

Naive Bayes Classifier

- Assume target function *f: X* ➔ *V*, where each instance *x* described by attributes $a_1, a_2 .. a_n$
- Most probable value of *f(x)* is:

$$v_{MAP} = \operatorname*{arg\,max}_{v_j \in V} P(v_j | a_1, a_2 \ldots a_n)$$

$$v_{MAP} = \operatorname*{arg\,max}_{v_j \in V} \frac{P(a_1, a_2 \ldots a_n | v_j) P(v_j)}{P(a_1, a_2 \ldots a_n)}$$

$$= \operatorname*{arg\,max}_{v_j \in V} P(a_1, a_2 \ldots a_n | v_j) P(v_j)$$

v_{NB}

- Naive Bayes assumption:

$$P(a_1, a_2 \ldots a_n | v_j) = \prod_i P(a_i | v_j)$$

- which gives

$$\text{Naive Bayes classifier: } v_{NB} = \operatorname*{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

Naive Bayes Algorithm

- For each target value v_j
- $\hat{P}(v_j)$ ← estimate $P(v_j)$
- For each attribute value a_i of each attribute a
- $\hat{P}(a_i|v_j)$ ← estimate $P(a_i|v_j)$

$$v_{NB} = \arg\max_{v_j \in V} \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i|v_j)$$

Training dataset

Class:
C1:buys_computer='yes'
C2:buys_computer='no'

Data sample:

X =
(age<=30,
Income=medium,
Student=yes
Credit_rating=Fair)

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
30…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Naïve Bayesian Classifier: Example

- Compute $P(X|C_i)$ for each class

  P(age="<30" | buys_computer="yes") = 2/9=0.222
  P(age="<30" | buys_computer="no") = 3/5 =0.6
  P(income="medium" | buys_computer="yes")= 4/9 =0.444
  P(income="medium" | buys_computer="no") = 2/5 = 0.4
  P(student="yes" | buys_computer="yes)= 6/9 =0.667
  P(student="yes" | buys_computer="no")= 1/5=0.2
  P(credit_rating="fair" | buys_computer="yes")=6/9=0.667
  P(credit_rating="fair" | buys_computer="no")=2/5=0.4

P(buys_computer=„yes") =9/14

P(buys_computer=„no")=5/14

- X=(age<=30 ,income =medium, student=yes,credit_rating=fair)

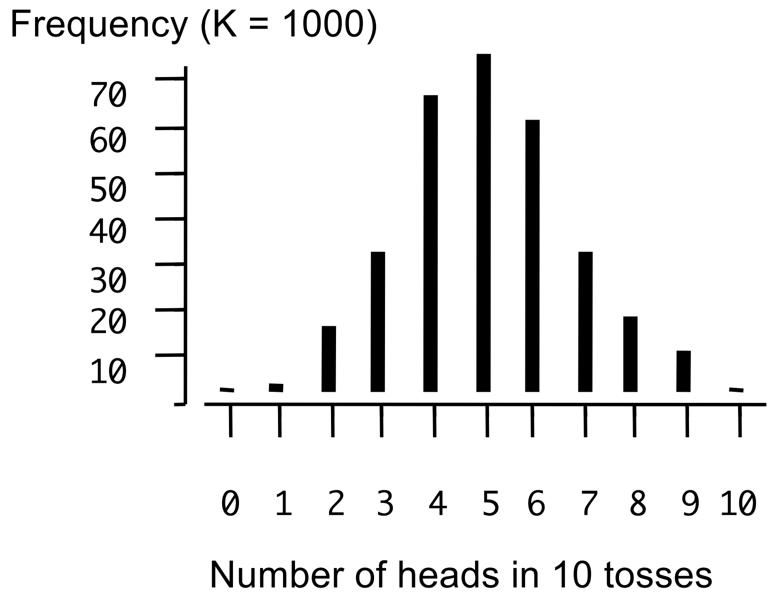
$P(X|C_1)$: P(X|buys_computer="yes")= 0.222 x 0.444 x 0.667 x 0.0.667 =0.044
$P(X|C_2)$: P(X|buys_computer="no")= 0.6 x 0.4 x 0.2 x 0.4 =0.019

$P(X|C_1)*P(C_1)$: P(X|buys_computer="yes") * P(buys_computer="yes")=0.028
$P(X|C_2)*P(C_2)$: P(X|buys_computer="no") * P(buys_computer="no")=0.007

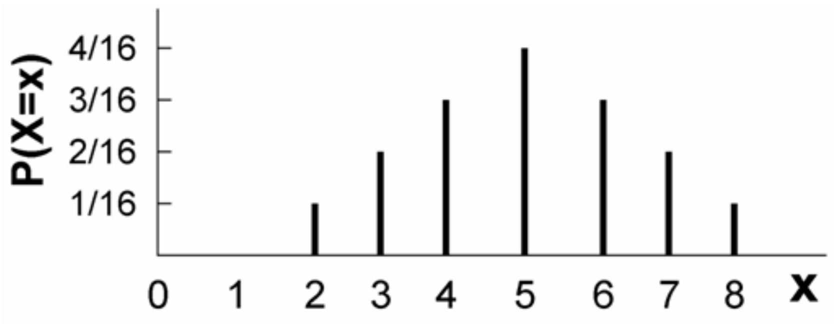
X belongs to class "buys_computer=yes" $P(C_1 |X)$ =0.028/(0.028+0.007)

Sampling of a Distribution

```
Loop K times
    r := 0                                   // r is num.heads in N
    tosses
    Loop N times                             // simulate the tosses
        Generate a random 0 ≤ x ≤ 1.0
        If x >= p increment r        // p is the probability of a head
    Push r onto sampling_distribution
Print sampling_distribution
```



x	P(x)
2	1/16
3	2/16
4	3/16
5	4/16
6	3/16
7	2/16
8	1/16



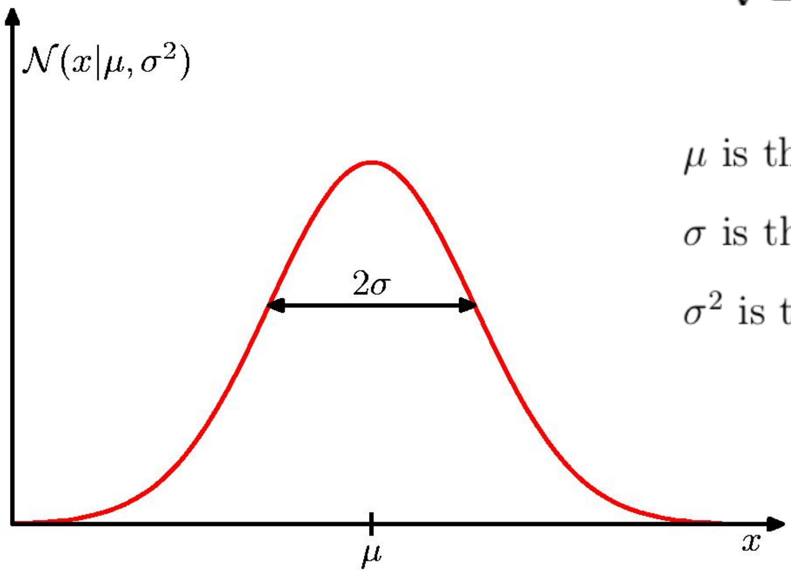
- In probability and statistics, a **probability mass function** (PMF) is a function that gives the probability that a discrete random variable is exactly equal to some value.

- Sometimes it is also known as the discrete density function. The probability mass function is often the primary means of defining a discrete probability distribution

Gaussian Distribution

- Gaussian distribution or normal is defined by the probability

$$p(x|\mu,\sigma^2) = \mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\cdot\pi\cdot\sigma}} \cdot \exp\left(-\frac{1}{2\cdot\sigma^2}\cdot(x-\mu)^2\right)$$



μ is the mean

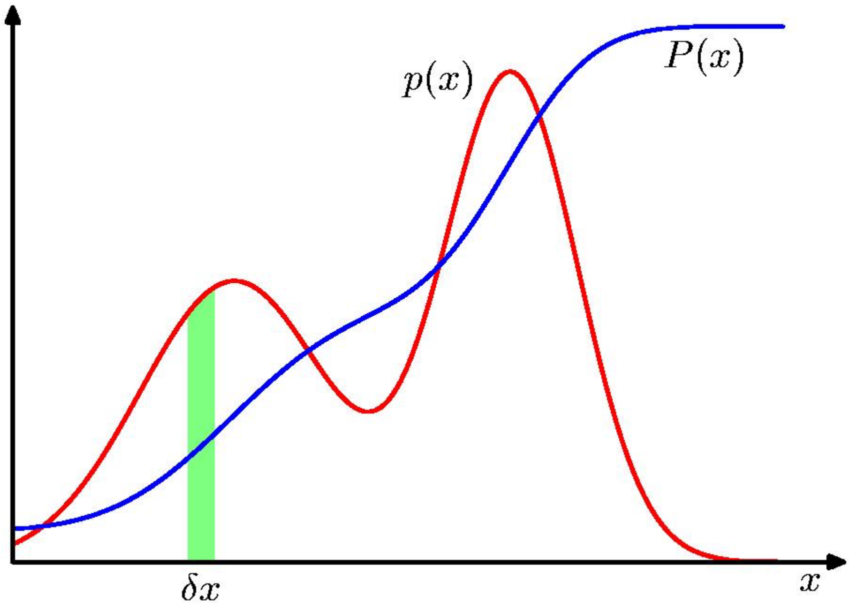
σ is the standard deviation

σ^2 is the variance

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

Probability Density Function (PDF)



$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

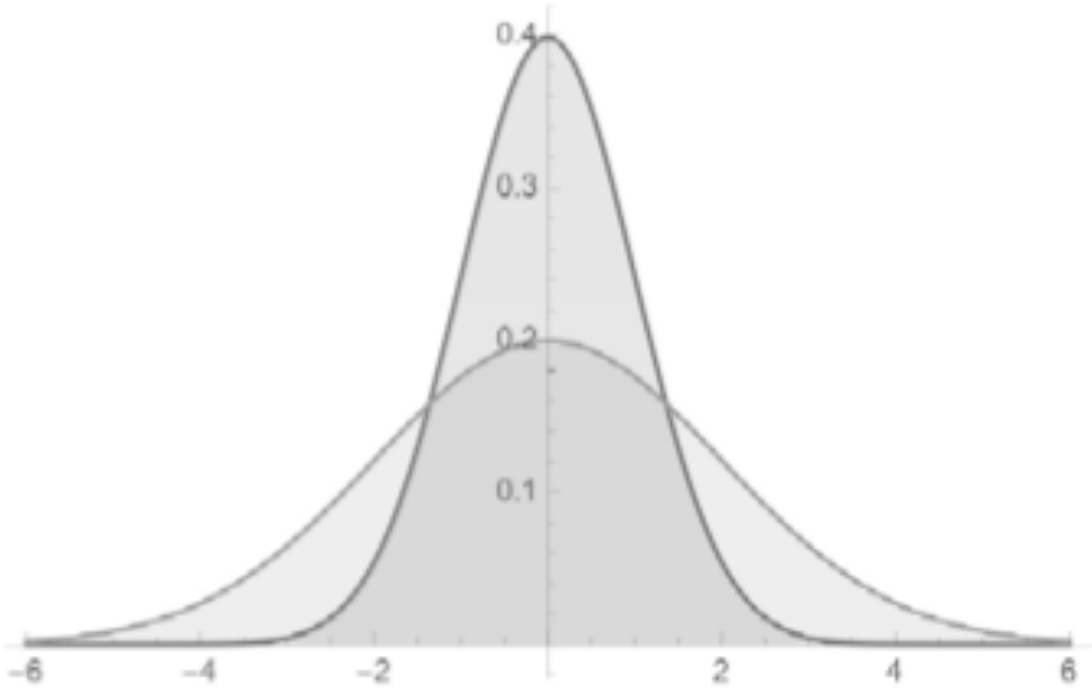
Cumulative distribution function (CDF)

$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Relative Probability

- Gaussian distribution is a type of continuous probability distribution for a real-valued random variable.
- The Gaussian distribution or normal distribution is defined as PDF (Probability Density Function) that reflects the **relative** probability.
- The **PDF may give a value greater than one** (small standard deviation).
- It is the area under the curve that represents the probability. However, the PDF reflects the relative probability.

    - Does a continuous probability distribution exist in the real world?



- Two Gaussian (normal) distribution with $\mu = 0$ $\sigma = 1$ and $\mu = 0$ $\sigma = 2$. μ describes the centre of the distribution and σ the width, the bigger σ the more flat the distribution.

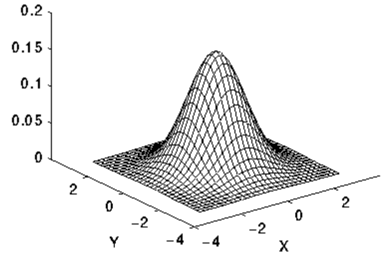
Precision

- Instead of inverting σ one uses precision which is often used in Bayesian software

$$\beta = \frac{1}{\sigma^2}, \qquad \beta^{-1} = \sigma^2$$

$$p(x|\mu, \beta) = \mathcal{N}(x|\mu, \beta^{-1}) = \frac{\beta}{\sqrt{2 \cdot \pi}} \cdot \exp\left(-\frac{1}{2} \cdot \beta \cdot (x - \mu)^2\right)$$

Normal Distribution in D dim

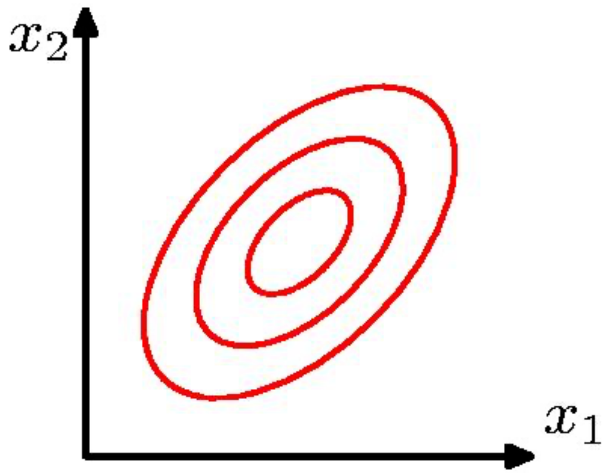


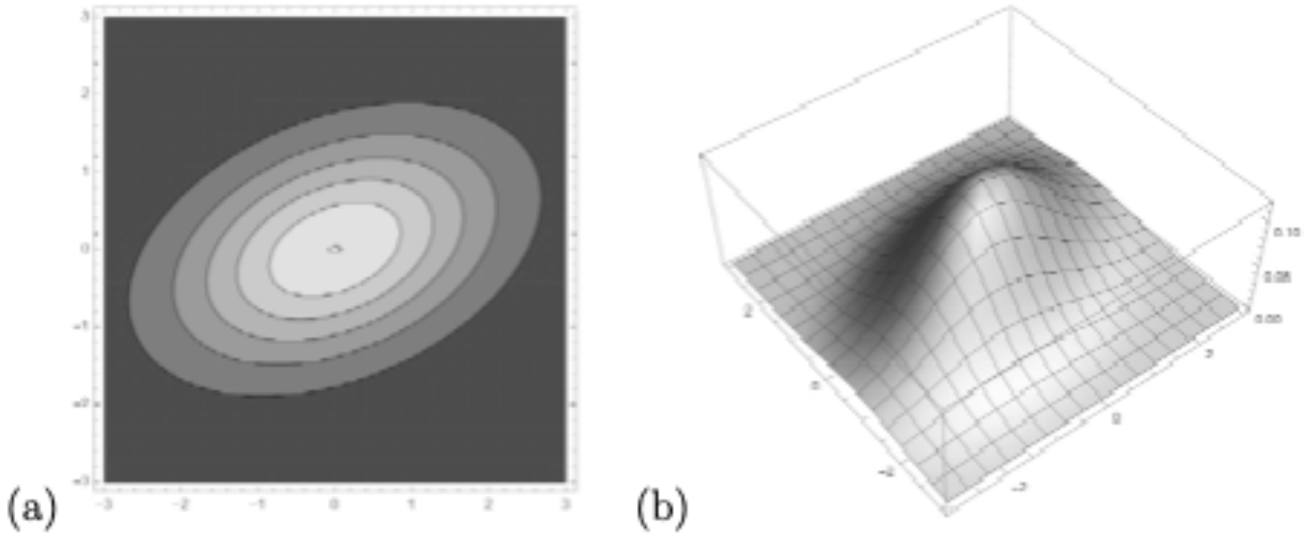
Over D dimensional space

$$p(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2 \cdot \pi)^{D/2}} \cdot \frac{1}{|\Sigma|^{1/2}} \cdot \exp\left(-\frac{1}{2} \cdot (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} \cdot (\mathbf{x} - \boldsymbol{\mu}) \right)$$

where

- $\boldsymbol{\mu}$ is the D dimensional mean vector
- Σ is a $D \times D$ covariance matrix
- $|\Sigma|$ is the determinant of Σ





- (a) The Gaussian distribution over 2 dimensional space with $\mu = (0, 0)^T$ and the covariance matrix Σ

$$\Sigma = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

- (b) Three dimensional plot of the Gaussian.

Precision

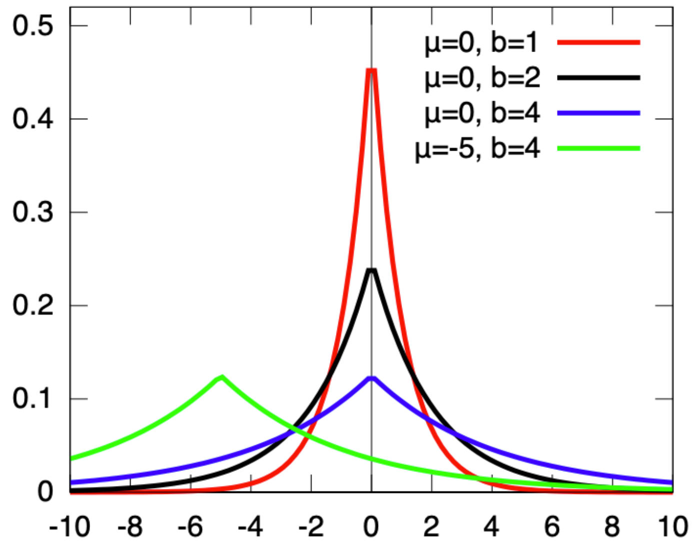
Instead of inverting Σ one uses precision matrix $\boldsymbol{\beta}$

$$\boldsymbol{\beta} = \Sigma^{-1}$$

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\beta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\beta}^{-1}) = \sqrt{\frac{|\boldsymbol{\beta}|}{(2 \cdot \pi)^D}} \cdot \exp\left(-\frac{1}{2} \cdot (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\beta} \cdot (\mathbf{x} - \boldsymbol{\mu})\right)$$

Laplace Distribution

- The probability distribution is



$$p(x|\mu, b) = Laplace(x|\mu, b) = \left(\frac{1}{2 \cdot b}\right) \exp\left(\frac{-|x-\mu|}{b}\right)$$

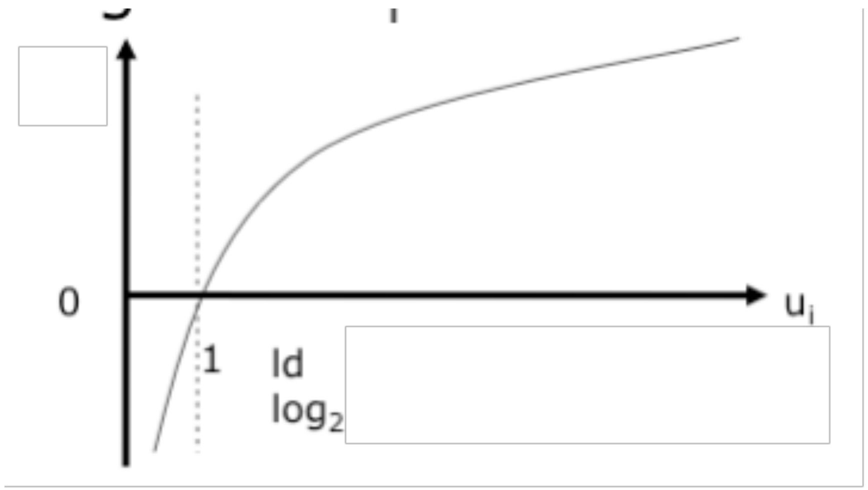
b > 0 is referred to as the diversity, is a scale parameter

Surprise

- **“Dog bites man”**
  - *No surprise*
  - *Quite common*
  - *not very informative*
- **“Man bites dog”**
  - *Most unusual*
  - *Seldom happens*
  - *Worth a headline!*

- Information is inversely related to probability

Information



$$I_i = \log_2(u_i) = \log_2(1/p_i) = -\log_2(p_i)$$

Information and probability:

- *Probabilities are* **multiplied**
- *Information is* **summed**
  - ***Use a logarithmic measure:***
  - ***I = log 1/p***
- *One unit of information (bit):*
  - **Yes/No**
  - **On/Off**
- *1 Binary symbol – use Base 2:*
  - **I = $\log_2$ 1/p bits**

Bit

J.W. Tukey

"After some more informal contacts during the first war years, on the initiative of mathematician Norbert Wiener, a number of scientists gathered in the winter of 1943-44 at a seminar, where Wiener himself tried out his ideas for describing intentional systems as based on feedback mechanisms. On the same occasion J.W. Tukey introduced the term a "bit" (binary digit) for the smallest informational unit, corresponding to the idea of a quantity of information as a quantity of yes-or-no answers."

Information Theory

- Involves the quantification of data with the goal of enabling as much data as possible to be reliably stored on a medium or communicated over a channel
- The measure of information, known as information entropy, is usually expressed by the average number of bits needed for storage or communication

- Let ***F*** be an experiment *(e.g. : two dice)*
  - Before we perform the experiment, we do not know what will be the results....
  - We are uncertain about the outcome
- How can we measure the uncertainty
- Instead of uncertainty we use the word **Entropy** of the experiment

$$0 \leq H(F) \leq \infty$$

Entropy - Information

- Experiments starts at t_0 and ends at t_1
- At t_0 we have no information about the results of the experiment
- At t_1 we have all information, so the **Entropy** of the experiment is 0
- From t_1 to t_0 we have wone **information**

Time	Entropy	Information
t_0 (before)	H(F)	0
t_1 (after)	0	H(F)

- We can describe an experiment by probabilities
- Experiment, outcome of the flip of a honest coin
- Head or Tail, both probability 0.5, the outcome can be either heat or tail, $p=(0.5,0.5)$
  - $H(F)=H(p_1,p_2)=(0.5,0.5)$

Interpretation of H(F)

- The experiment *F* was done
- Person *A* knows the outcome, person *B* not
- How to define *H*?
- *H* = number of questions to *A, B* has to pose to know the result of the experiment
  - Questions of the form yes/no

Interpretation of H(F)

- Example coin, p=(0.5,0.5)
- We can pose the question, is it tail?
- H=1

- Not interesting

- Example cards, *p=(1/2,1/4,1/4)*
  - „red“, „clubs“, „spade“

- We can ask, is the card red, if the answer is no, we have only to ask is it spade...
- If the card is red, we need only one question, else we need two questions
- We have to speak about the mean number of questions

- $H(F)=1/2*1+1/4*2+1/4*2=1.5$

  - If the card is red, we need only one question, for clubs and spade we need 2 questions...

Interpretation of H(F)

- The experiment F was done
- Person A knows the outcome, person B not
- How to define H?
- H = mean number of optimal questions to A, B has to pose to know the result of the experiment
  - Questions of the form yes/no

- For four cards of which one is the joker the probability of a joker is *0.25* and of other cards *1-0.25=0.75*, *p=(0.25,0.75)*
- In the mean we have to ask
- $1*0.25 + 1*0.75=1$
- questions to determine to determine if the card is a joker or not.

- Given n cards of which one is the joker the probability of a joker is $1/n$ and of other cards is $1-1/n$
- In the mean we have to ask

$$1 * 1/n + 1 * (1 - 1/n)$$

  questions to determine if the card is a joker or not.
- Its results in one question independent of the size of n.

- It seems some thing is missing in our definition
- Our result is correct for one independent experiment
- For several experiments the mean number of questions is lower

Real Entropy

- We define the real entropy:

  - for one experiment as $H_0(F^1)$
  - for two experiments as $H_0(F^2)$
  - ..
  - For k experiments as $H_0(F^k)$

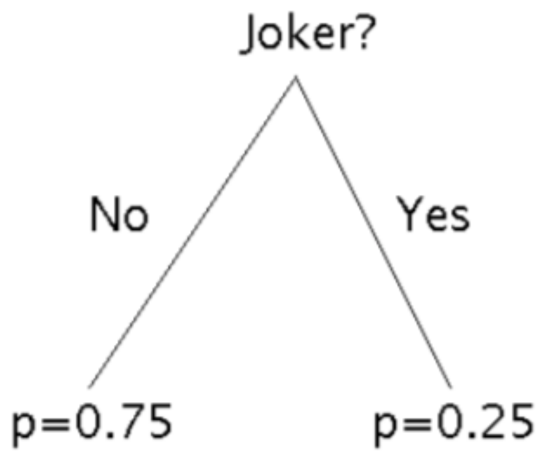
- The mean number of question for one experiment in the sequence of k experiments is

  - $1/k * H_0(F^k)$

$H_0(F^1)$

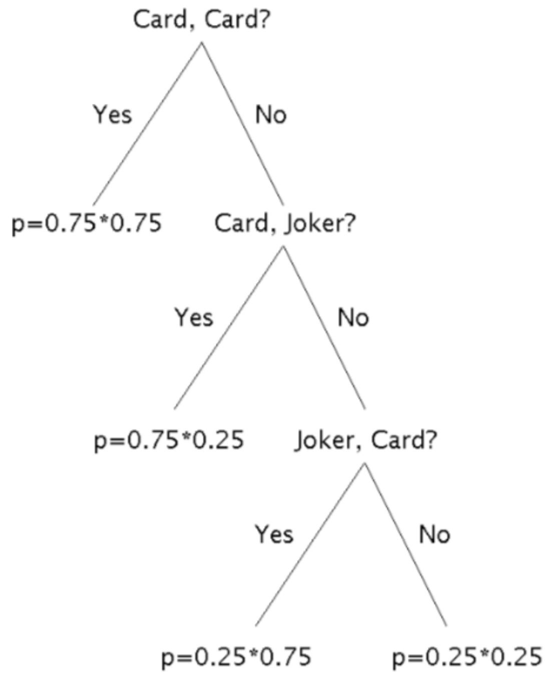
- For four cards of which one is the joker the probability of a joker is *0.25* and of other cards *1-0.25=0.75*

- $H_0(F^1)=1$
- $H_0(F^1)=1= 1*0.75 + 1*0.25=1$
- $k=1,\ 1/k\ *H_0(F^k)=1/1*H_0(F^1)=1$



$H_0(F^2)$

results	probability
card, card	$0.75 \cdot 0.75$
joker, card	$0.25 \cdot 0.75$
card, joker	$0.75 \cdot 0.25$
joker, joker	$0.25 \cdot 0.25$



$$H_0(F^2) = 1 \cdot 0.75 \cdot 0.75 + 2 \cdot 0.75 \cdot 0.25 + 3 \cdot 0.25 \cdot 0.75 + 3 \cdot 0.25 \cdot 0.25$$

$$H_0(F^2) = 1.6875 \qquad \frac{H_0(F^2)}{2} = 0.84375$$

$H_0(F^3)$

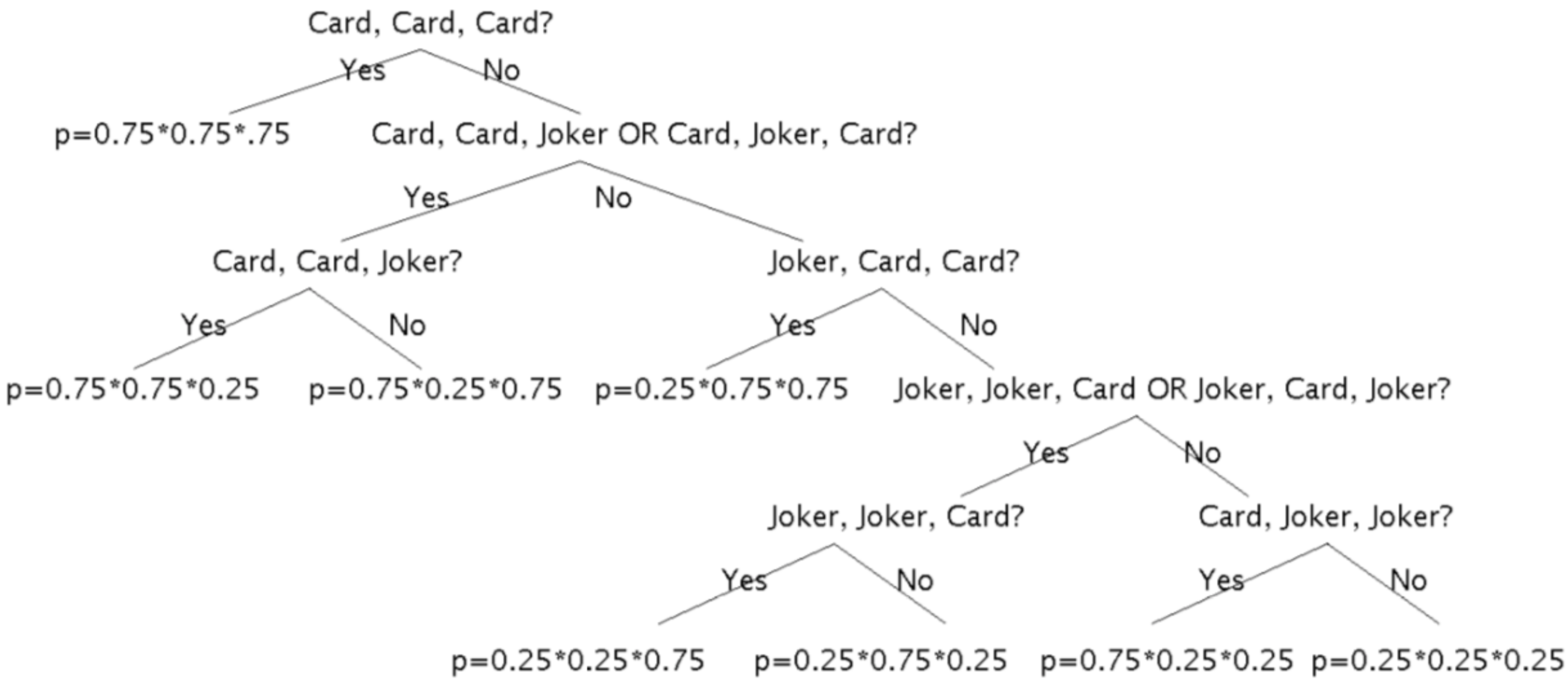
results	probability
card, card, card	$0.75 \cdot 0.75 \cdot 0.75$
card, card, joker	$0.75 \cdot 0.75 \cdot 0.25$
card, joker, card	$0.75 \cdot 0.25 \cdot 0.75$
joker. card card	$0.25 \cdot 0.75 \cdot 0.75$
joker, joker, card	$0.25 \cdot 0.25 \cdot 0.75$
joker, card, joker	$0.25 \cdot 0.75 \cdot 0.25$
card, joker, joker	$0.75 \cdot 0.25 \cdot 0.25$
joker, joker, joker	$0.25 \cdot 0.25 \cdot 0.25$

$$H_0(F^3) = 1 \cdot 0.42188 + 3 \cdot 0.14062 + 3 \cdot 0.14062 + 3 \cdot 0.14062 +$$

$$+5 \cdot 0.046875 + 5 \cdot 0.046875 + 5 \cdot 0.046875 + 5 \cdot 0.015625$$

$$H_0(F^3) = 2.4688$$

$$\frac{H_0(F^3)}{3} = 0.82292$$


Card, Card, Card?
Yes
No
p=0.75*0.75*.75
Card, Card, Joker OR Card, Joker, Card?
Yes
No
Card, Card, Joker?
Joker, Card, Card?
Yes
No
Yes
No
p=0.75*0.75*0.25
p=0.75*0.25*0.75
p=0.25*0.75*0.75
Joker, Joker, Card OR Joker, Card, Joker?
Yes
No
Joker, Joker, Card?
Card, Joker, Joker?
Yes
No
Yes
No
p=0.25*0.25*0.75
p=0.25*0.75*0.25
p=0.75*0.25*0.25
p=0.25*0.25*0.25

H(F)

Does the sequence $h_k := \frac{H_0(F^k)}{k}$, with the values $\{1, 0.84375, 0.82292, ...\}$ for $k = 1, 2, 3, ..$ have a limit for $\lim_{k\to\infty} h_k$?
It has. The limit is defined as

$$H(F) := \lim_{k\to\infty} \frac{H_0(F^k)}{k} \leq H_0(F)$$

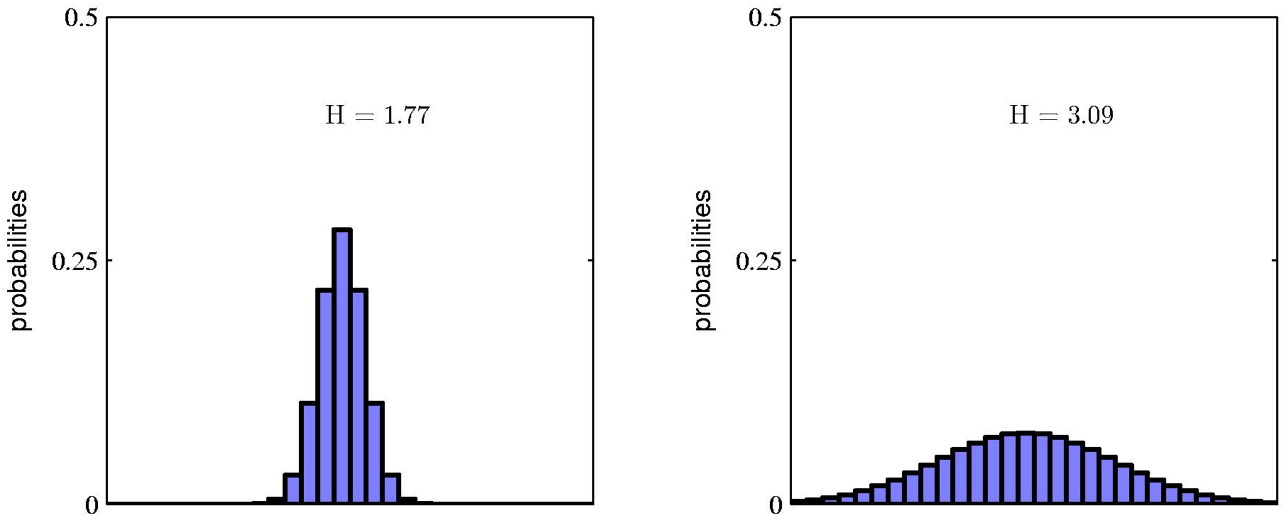
- it is called the ideal entropy, it converges to

$$H(F) = -\sum_i p_i \log_2 p_i.$$

(Ideal) Entropy

- The ideal entropy indicates the minimal number of optimal questions that *B* must pose to know the result of the experiment on
- Suppose that *A* repeated the experiment an infinite number of times
- The ideal entropy is the essential information obtained by taking out the redundant information that corresponds to the ideal distribution to which the results converge

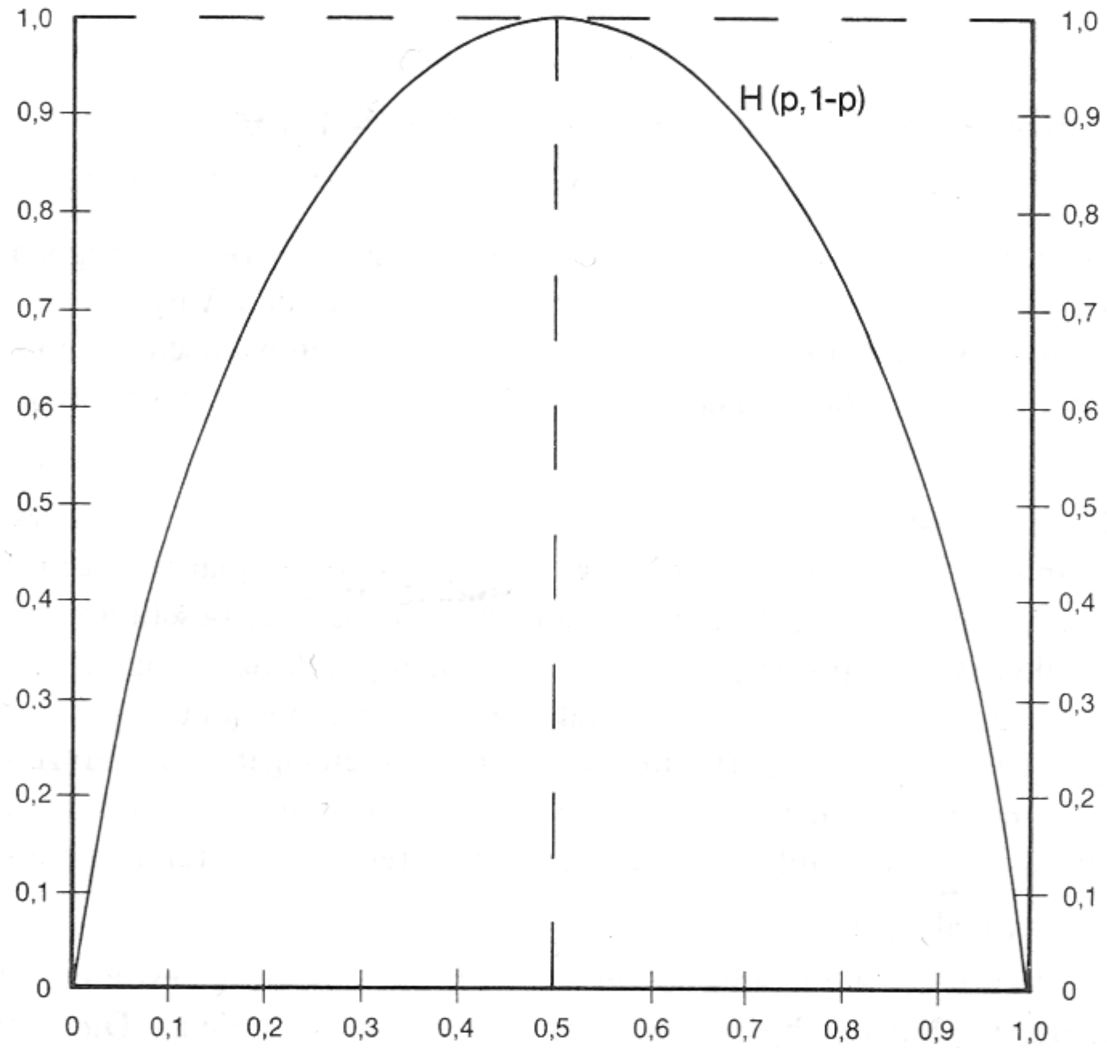
Entropy



- An experiment is described by probabilities $p=(p_1,p_2,...,p_n)$
- Does the distribution of these probabilities have an effect on the ideal entropy?
- It turns out that the ideal entropy is maximal in the case all probabilities are equal, means $p=(1/n,1/n...,1/n)$
- In this case the maximal ideal Entropy is

$$H(F) = -\sum_i p_i \log_2 p_i = -\log_2 1/n = \log_2 n$$

n=2



Entropy

- Coding theory: *x* discrete with 8 possible states; how many bits to transmit the state of *x*?

- All states equally likely

$$\mathrm{H}[x] = -8 \times \frac{1}{8}\log_2 \frac{1}{8} = 3 \text{ bits.}$$

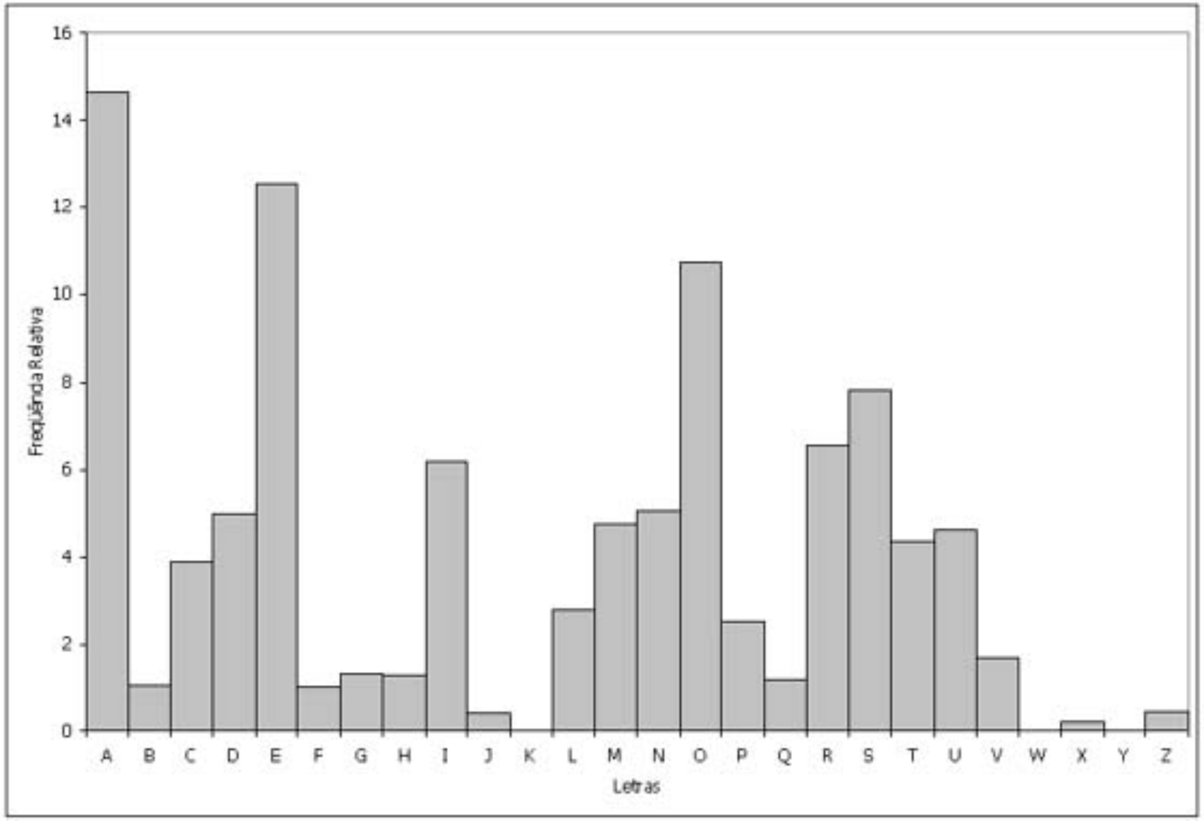
Entropy

x	a	b	c	d	e	f	g	h
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$
code	0	10	110	1110	111100	111101	111110	111111

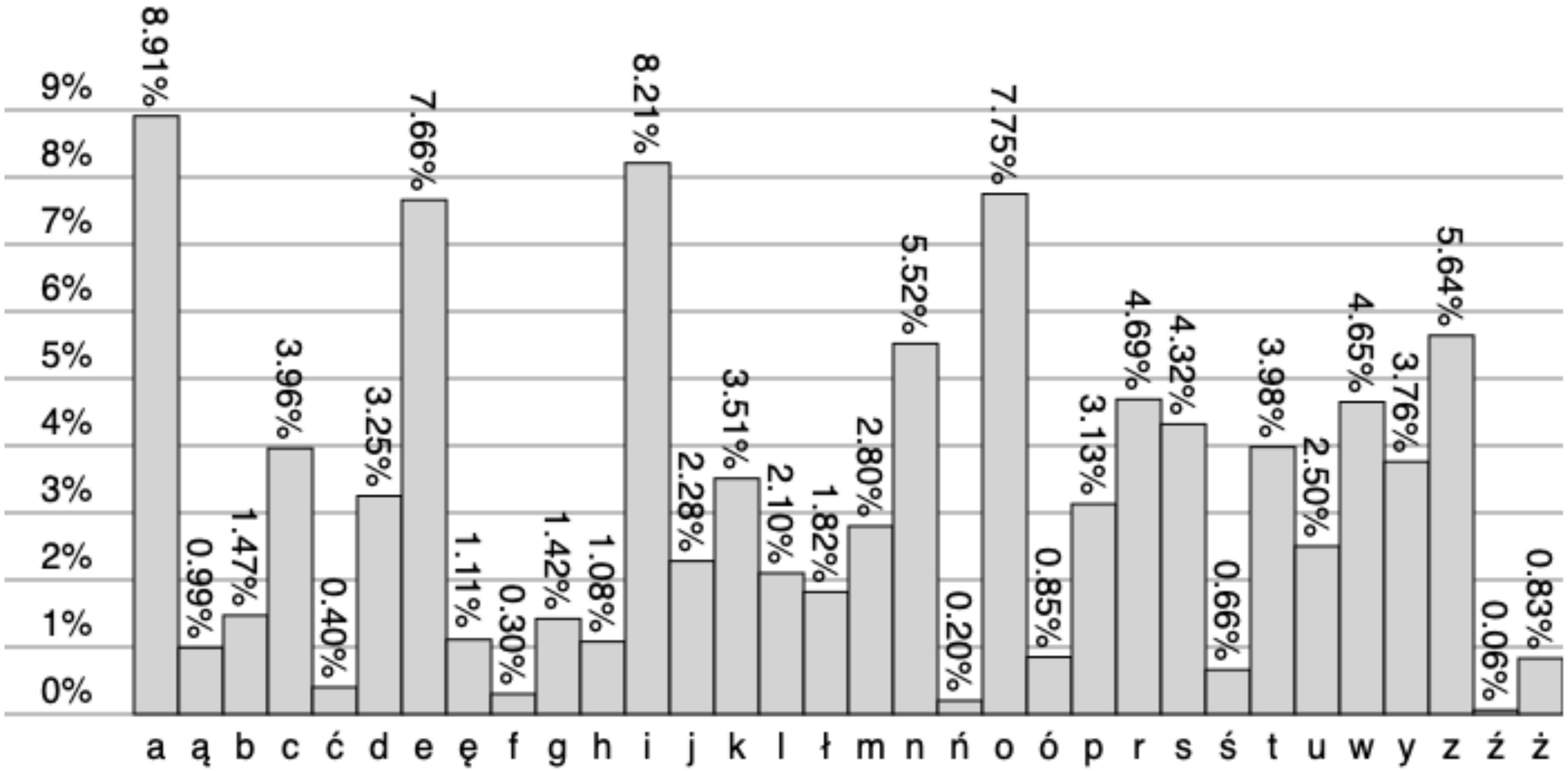
$$\begin{aligned} \mathrm{H}[x] &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} \\ &= 2 \text{ bits} \end{aligned}$$

$$\begin{aligned} \text{average code length} &= \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + 4\times\frac{1}{64}\times 6 \\ &= 2 \text{ bits} \end{aligned}$$

Frequência de uso das letras na língua portuguesa



Polish letters frequencies



- The relationship between *log$_2$* and any other base *b* involves multiplication by a constant,

$$\log_2 x = \frac{log_b x}{log_b 2} = \frac{log_{10} x}{log_{10} 2}.$$

$$H = -\frac{1}{\log_{10} 2} \cdot \sum_i^n p(m_i) \cdot \log_{10} p(m_i). = -\sum_i^n p(m_i) \cdot \log_2 p(m_i)$$

nat

$$H = -\sum_i p(x_i)\ln p(x_i) = -\sum_i p(x_i)\log p(x_i)$$

- Instead of measuring the information in bits, yes no questions, it measure the information in nepit (nat), it is the power of the Euler's number *e=2.7182818…* (sometimes also called Napier's constant).

Conditional Entropy

- Quantifies the amount of information needed to describe the outcome of a random variable *Y* given that the value of another random variable *X* is known

$$H(Y|X) = -\sum_{x \in X, y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)}\right)$$

Mutual Information

- Mutual information measures the information that *X* and *Y* share
- How much knowing one of these variables reduces uncertainty about the other

$$I(X,Y) = -\sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x) \cdot p(y)}{p(x,y)}\right)$$

- For example, if *X* and *Y* are independent, then knowing *X* does not give any information about *Y* and their mutual information is zero.

Relative Entropy

- Kullback-Leibler divergence (also called relative entropy) is a measure of how one probability distribution is different from a second
- For discrete probability distributions p and q defined on the same probability space, the Kullback-Leibler divergence between *p* and *q* is defined as

$$KL(p||q) = -\sum_{x \in X} p(x) \log q(x) - \left(-\sum_{x \in X} p(x) \log p(x) \right)$$

$$KL(p||q) = -\sum_{x \in X} p(x) \log \left(\frac{q(x)}{p(x)} \right)$$

- Example Consider some unknown distribution *p(x)*
- Suppose that we have modelled this using an approximating distribution *q(x)*
- If we use *q(x)* to construct a coding scheme for the purpose of transmitting values of *x* to a receiver, then the average additional amount of information required to specify the value of *x* as a result of using *q(x)* is *KL(p||q)*

$$KL(p||q) = -\sum_{x \in X} p(x) \log q(x) - \left(-\sum_{x \in X} p(x) \log p(x) \right)$$

Cross Entropy

- For discrete probability distributions p and q defined on the same probability space, the cross entropy between p and q is defined as

$$H(p, q) = -\sum_{x \in X} p(x) \log q(x).$$

$$H(p, q) = H(p) - KL(p||q)$$

In machine learning with the true distribution *Y* :

- is either a binary value y_k for each data element y_k of the dataset

$$H(Y,O) = -\sum_{k=1}^{N}(y_k \cdot \log o + (1-y_k)\cdot \log(1-o))$$

$$H(Y,O) = -\sum_{k=1}^{N}(y_k \cdot \log o + \neg y_k \cdot \log \neg o)$$

  and the estimated distribution is $O = (o, \neg o)$ does not need to be binary with $1 = o + \neg o$.

- or a 1-of-K representation for $\mathbf{y}_k$ vector of the dataset

$$H(Y,O) = -\sum_{k=1}^{N}\sum_{t=1}^{K} y_{kt} \cdot \log o_{kt}$$

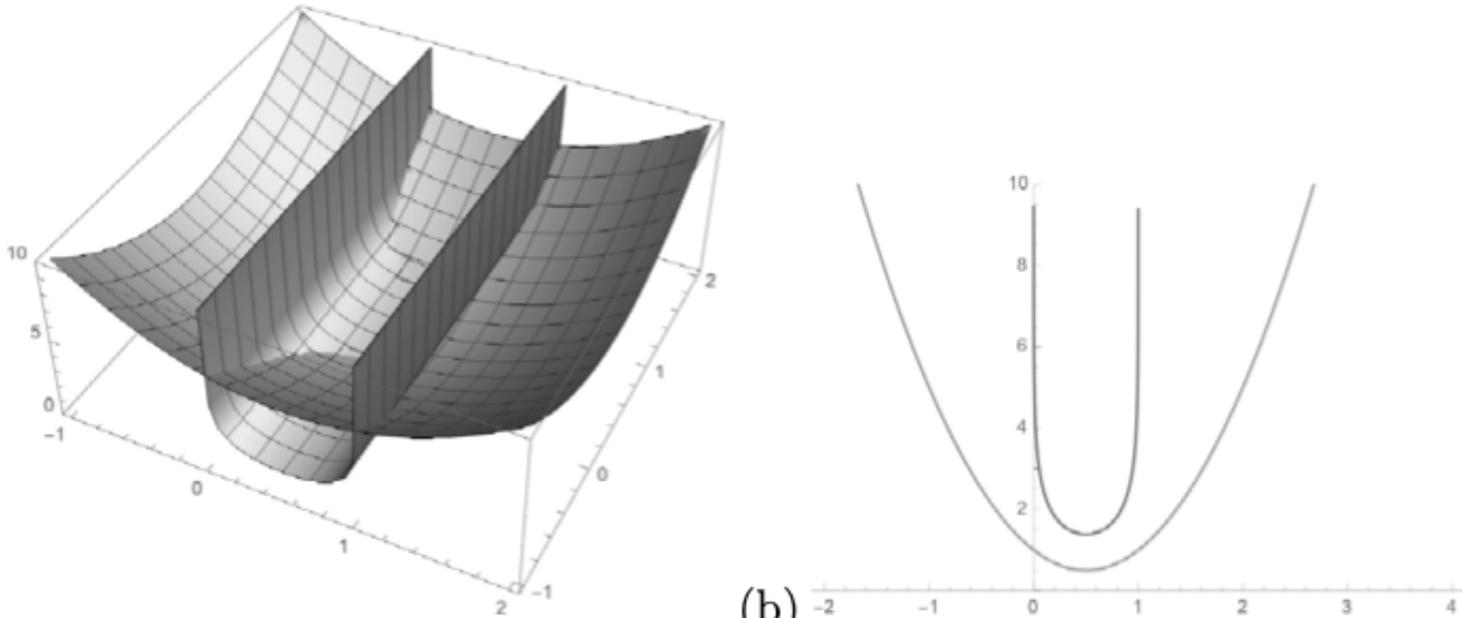
  and the estimated distribution does not need to be binary with the requirement $1 = \sum_{t=1}^{K} o_{kt}$.

- The distribution H(Y,O) defines a loss function measured

$$L(y, o) = H(Y, O)$$

- which is not a distance function since it is not symmetric and is only defined over probability distributions.

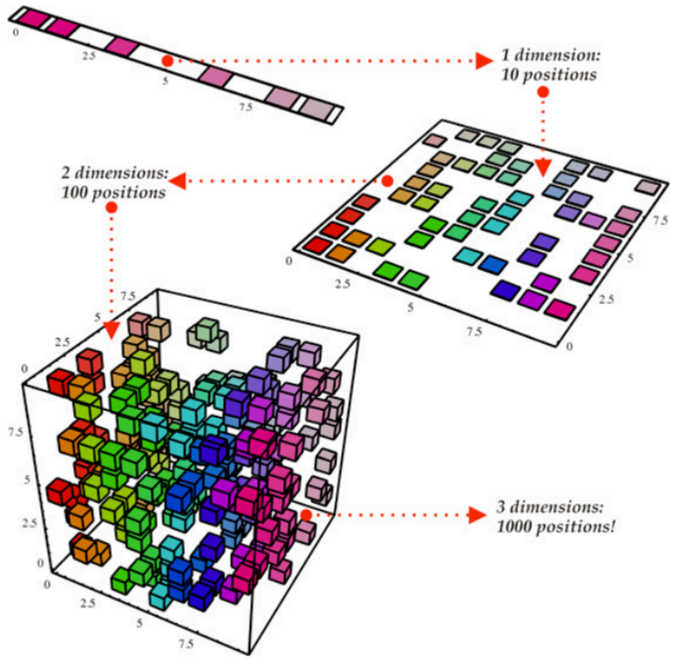
- Loss function indicates a cost function, it is equivalent to the name error function of energy function in other domains



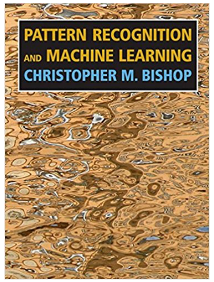
(a) (b)

The loss function that is based on cross entropy is much more steep the a possible loss or error function that is based on quadratic loss that is based on squared Euclidean distance

- What about the *vector space*?
- What the *Curse of Dimensionality?*
- How to find a minimum of a function?

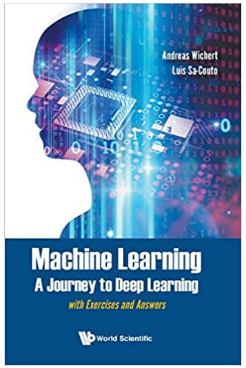


Literature



- Christopher M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer 2006
  - Section *1.2, 1.6, 2.3*

Literature



- Machine Learning - A Journey to Deep Learning, A. Wichert, Luis Sa-Couto, World Scientific, 2021
  - Chapter 2