Lecture 4: Probability and Bayesian Classifier

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, $p(\Omega) = 1$, then $a \in \Omega$.
- *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

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

- 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 $b_1, ..., b_n$ with

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

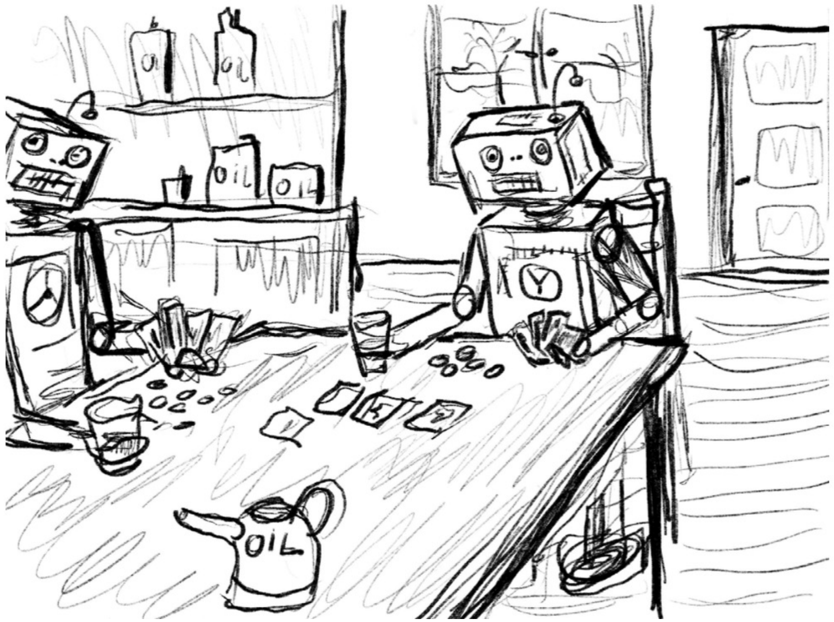
- the **law of total probability** is represented by

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

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

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

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

- The Bayes' rule follows from both equations

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

Reverent Thomas Bayes
(1702-1761)

- He set down his findings on probability in "Essay Towards Solving a Problem in the Doctrine of Chances" (1763), published posthumously in the Philosophical Transactions of the Royal Society of London.
    - The drawing after a portrait of Bayes used in a 1936 book, it is not known if the portrait is actually representing him.

Bayes' Rule

$$p(h_k|D) = \frac{p(D|h_k) \cdot p(h_k)}{p(D)} = \frac{p(D, h_k)}{p(D)}$$

- *$p(h_k)$ is called the **prior** (before)*
  - *For example, what is the probability of some illness in Portugal*
- *$p(D|h_k)$ is called **likelihood** and can can be easily estimated*
  - *For example, what is the probability that some illness generates some symptoms?*
  - *$p(D,h_k)$ is called **joint distribution***
- *$p(h_k|D)$ is called **posterior probability***

Bayes' Rule

$$p(h_k|D) = \frac{p(D|h_k) \cdot p(h_k)}{p(D)} = \frac{p(D, h_k)}{p(D)}$$

- Bayes rule can be used to determine the total posterior probability $p(h_k|D)$ of hypothesis h_k given data D
  - For example, what is the probability that some illness is present?
- The most probable hypothesis h_k out of a set of possible hypothesis $h_1, h_2, \cdots$ given some present data is according to the Bayes rule

Maximum a Posteriori (MAP) Hypothesis

- $p(h_k|D)$ and $p(D,h_k)$ are related in a linear manner

$$p(h_k|D) \propto p(D|h_k) \cdot p(h_k)$$

posterior ∝ likelihood × prior

- to determine the **maximum posteriori hypothesis** h_{MAP} we maximize

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

- we can see, the maximization is independent of *p(D)*, it follows

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

Maximum Likelihood (ML) hypothesis

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

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

Bayesian Interpretation

- In the Bayesian (or epistemological) interpretation, **probability measures a “degree of belief”** and Bayes’ rule links the degree of belief in a proposition before and **after** accounting for evidence
- with *prior probability* $p(h_k)$
- $p(D|h_k)$ represents the likelihood of the data D if we assume h_k to be true
  - if we, in fact, observe D, we can update our belief about h_k through the rule

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

Bayesian Interpretation and bias

- Objective likelihood is biased by the prior belief

$$\textit{posterior} \propto \textit{likelihood} \times \textit{prior} = \textit{likelihood} \times \textit{bias}$$

- Bias is a disproportionate weight in favor of or against an idea or thing, usually in a way that is closed-minded, prejudicial, or unfair.
- Biases can be innate or learned. People may develop biases for or against an individual, a group, or a belief.

Cancer screening

- Cancer screening aims to detect cancer before symptoms appear
- This may involve for example a blood test.
- Suppose that a patient tests positive…

- The test is secure because in **99** percent of the cases the test returns a correct positive result (= positive) in which a rare form of cancer is actually present.
- Should the doctor tell the patient, that he has cancer?

- The test has correct negative result (= negative) in 99 percent of the cases where the rare form of cancer is not present
- It is also known that 0.001 of the entire population have the rare form of cancer (h = cancer)

- $p(cancer) = 0.001$, $p(\neg cancer) = 0.999$
- $p(positive|cancer) = 0.99$, $p(positive|\neg cancer) = 0.01$,
- $p(negative|cancer) = 0.01$, $p(negative|\neg cancer) = 0.99$

- We determine h_{map} according to the linear relation

$$posterior \propto likelihood \times prior$$

$$p(cancer|positive) \propto p(positive|cancer) \cdot p(cancer) = 0.99 \cdot 0.001$$
$$p(\neg cancer|positive) \propto \cdot p(positive|\neg cancer) \cdot p(\neg cancer) = 0.01 \cdot 0.999$$

It follows

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

$$p(cancer|positive) \propto p(positive|cancer) \cdot p(cancer) = 0.99 \cdot 0.001$$

$$p(\neg cancer|positive) \propto \cdot p(positive|\neg cancer) \cdot p(\neg cancer) = 0.01 \cdot 0.999$$

It follows

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

- So, despite the positive result, **we are still more confident** that the patient is healthy than otherwise.
- The right thing to do would be to another test to try to accumulate more evidence in favor of the hypothesis that patient has the disease.

$$p(positive, cancer) = p(positive | cancer) \cdot p(cancer) = 0.99 \cdot 0.001$$

$$p(positive, \neg cancer) = p(positive | \neg cancer) \cdot p(\neg cancer) = 0.01 \cdot 0.999$$

$$p(positive | cancer) = \frac{p(positive, cancer)}{p(positive, cancer) + p(positive, \neg cancer)}$$

law of total probability: $p(positive) = p(positive, cancer) + p(positive, \neg cancer)$

$$p(positive | cancer) = \frac{p(positive | cancer) \cdot p(cancer)}{p(positive)}$$

Estimating p(h)

- Let us draw some principles to estimate

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

- Let us first start with $p(h)$
  - given no prior knowledge that *one hypothesis is more likely* than another
    - $p(h)$ can be uniformly distributed

$$\forall_{h \in H} \ p(h) = \frac{1}{|H|}$$

  - otherwise, estimate the prior base on the observed frequency

Estimating $p(D|h)$

- What choice shall we make for *P(D|h)* ?
    - Hypothesis generates data....

    - If data is **discrete**:
    - we use the frequentist approach
        - e.g. I observe 2 out of 10 individuals with blue eyes and brown in shift A and 1 out of 8 in shift B, then
        - $p(\mathbf{x} = [blue\ eyes, brown]|A) = 0.2$ and
        - $p(\mathbf{x} = [blue\ eyes, brown]|B) = 0.125$

Bayesian optimal classifier

- What is the most probable classification of the new instance given the training data?

$$h_{MAP} = \arg\max_h p(h|\mathbf{x}_{new}) = \arg\max_h \frac{p(\mathbf{x}_{new}|h)p(h)}{p(\mathbf{x}_{new})} = \arg\max_h p(\mathbf{x}_{new}|h)p(h)$$

… where the hypotheses correspond to our classes

- we ignore the denominator as it does not alter decision

- The Bayesian classifier has as many parameter as:
  - the number of priors minus 1
    - we can deduce one prior from the remaining ones
    - e.g. given h_1, h_2 and h_3, $p(h_3) = 1 - p(h_2) - p(h_1)$
  - the number of parameters associated with the class-conditional distributions, $p(\mathbf{x}|h)$

Bayesian optimal classifier: example

	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$	$class$
x_1	1	C	1	1
x_2	1	C	1	0
x_3	0	B	1	0
x_4	0	A	0	0
x_5	1	C	1	1
x_6	0	B	1	1
x_7	0	A	0	1

- Priors
  - $p(c=0)=\frac{card(c=0)}{card(\Omega)}=\frac{3}{7}$, $p(c=1)=1-p(c=0)=\frac{4}{7}$
- Joint Probability
  - $p(v_1=0, v_2=\text{A}, v_3=0, c=0)=\frac{card(v_1=0,v_2=\text{A},v_3=0,c=0)}{card(\Omega)}=\frac{1}{7}$
- Likelihood
  - $p(v_1=0, v_2=\text{A}, v_3=0|c=0)=\frac{1}{3}=\frac{card(v_1=0,v_2=\text{A},v_3=0,c=0)}{card(c=0)}=\frac{1}{3}=\frac{p(v_1=0,v_2=\text{A},v_3=0,c=0)}{p\ (c=0)}=\frac{\frac{1}{7}}{\frac{3}{7}}=\frac{1}{3}$
  - Data Joint
    - $p(v_1=0, v_2=\text{A}, v_3=0)=\frac{card(v_1=0,v_2=\text{A},v_3=0,c=0)}{card(\Omega)}=\frac{2}{7}$
  - Data: law of total probability
    - $\text{p}(v_1=0, v_2=\text{A}, v_3=0)=p(v_1=0, v_2=\text{A}, v_3=0, c=0)+p(v_1=0, v_2=\text{A}, v_3=0, c=1)=\frac{1}{7}+\frac{1}{7}$
- Posterior
  - $p(c=0|v_1=0, v_2=\text{A}, v_3=0)=\frac{p(v_1=0,v_2=\text{A},v_3=0|c=0)p\ (c=0)}{p\ (v_1=0,v_2=\text{A},v_3=0)}=\frac{\frac{3}{7}*\frac{1}{3}}{\frac{2}{7}}=\frac{\frac{1}{7}}{\frac{2}{7}}==\frac{1}{2}$

Bayesian optimal classifier: example

	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$	*class*
x_1	1	C	1	*1*
x_2	1	C	1	*0*
x_3	0	B	1	*0*
x_4	0	A	0	*0*
x_5	1	C	1	*1*
x_6	0	B	1	*1*
x_7	0	A	0	*1*

- We can "classify" new observations in the same way, e.g. $\mathbf{x}_{\text{new}} = [1, \mathrm{C}, 1]$, what is the class, $\mathrm{c} = 0$ or $\mathrm{c} = 1$?

  - $p(c = 1\,, v_1 = 1, v_2 = \mathrm{C}, v_3 = 1) = p(v_1 = 1, v_2 = \mathrm{C}, v_3 = 1 | c = 1) p\,(c = 1) = \frac{2}{7}$
  - $p(c = 0\,, v_1 = 1, v_2 = \mathrm{C}, v_3 = 1) = p(v_1 = 1, v_2 = \mathrm{C}, v_3 = 1 | c = 0) p\,(c = 0) = \frac{1}{7}$

  - $p(c = 1\,, v_1 = 1, v_2 = \mathrm{C}, v_3 = 1) > p(c = 0\,, v_1 = 1, v_2 = \mathrm{C}, v_3 = 1)$

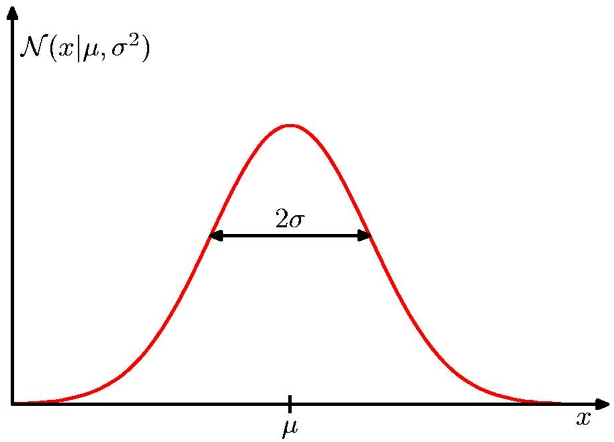
  $\mathbf{x}_{\text{new}}$ is classified with class 1

  - $p(c = 1\,| v_1 = 1, v_2 = \mathrm{C}, v_3 = 1) = \frac{p(c=1\,, v_1=1, v_2=\mathrm{C}, v_3=1)}{p(v_1=1, v_2=\mathrm{C}, v_3=1)} = \frac{\frac{2}{7}}{\frac{3}{7}} = \frac{2}{3}$
  - $p(c = 0\,| v_1 = 1, v_2 = \mathrm{C}, v_3 = 1) = \frac{p(c=0\,, v_1=1, v_2=\mathrm{C}, v_3=1)}{p(v_1=1, v_2=\mathrm{C}, v_3=1)} = \frac{\frac{1}{7}}{\frac{3}{7}} = \frac{1}{3}$

    - $p(c = 1\,| v_1 = 1, v_2 = \mathrm{C}, v_3 = 1) + p(c = 1\,| v_1 = 1, v_2 = \mathrm{C}, v_3 = 0) = 1$

Estimating $p(D|h)$

- If data is **real-valued**:
- We can use Probability Density Function of the Normal distribution

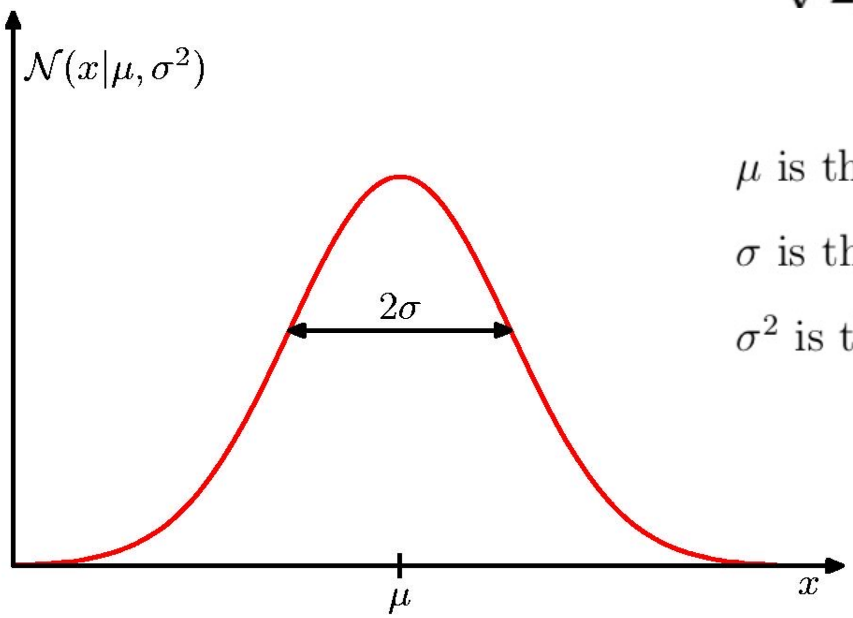


- Is this correct? **No…**
  - We assume relative probability is real probability
    - However, we do it because it is simple
    - Error for many data points small
  - How do we know that the data is described by Normal distribution?
  - This assumption can be wrong!

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

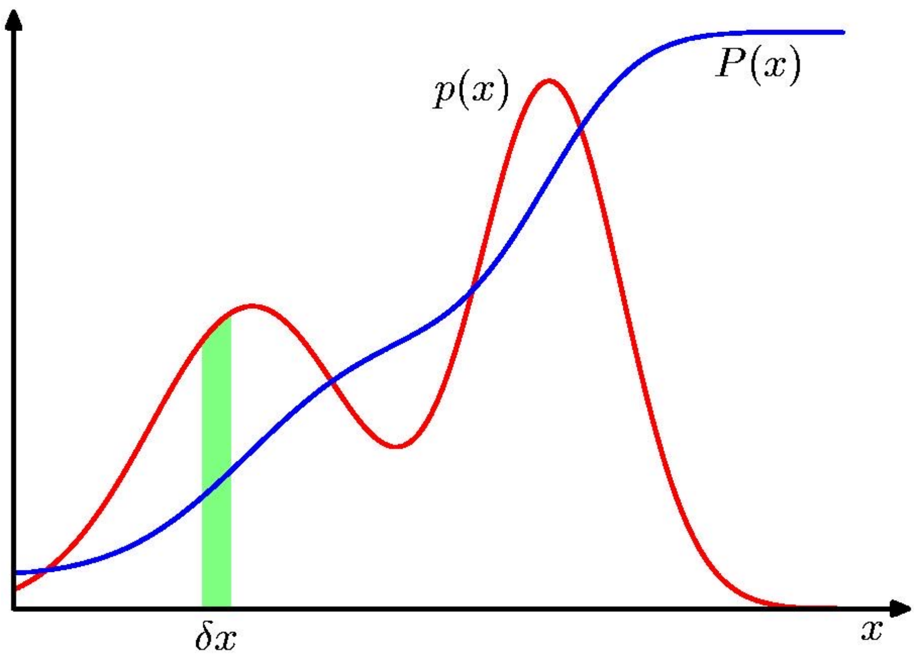
σ is the standard deviation

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)

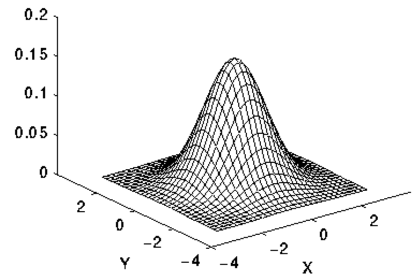
$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

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?

Normal Distribution in D dim

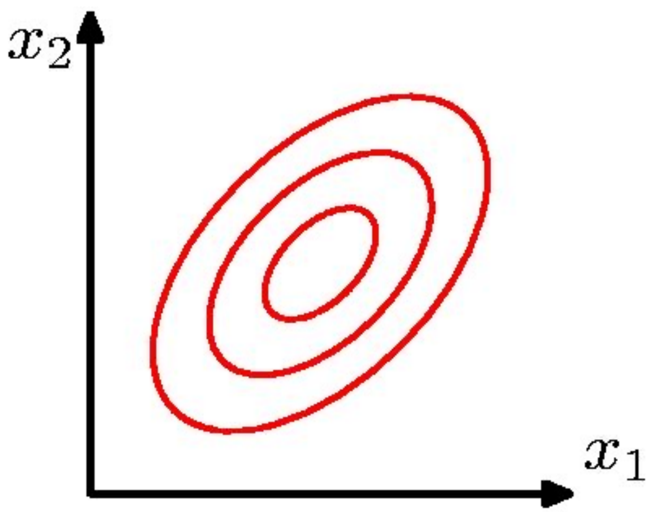


Over D dimensional space

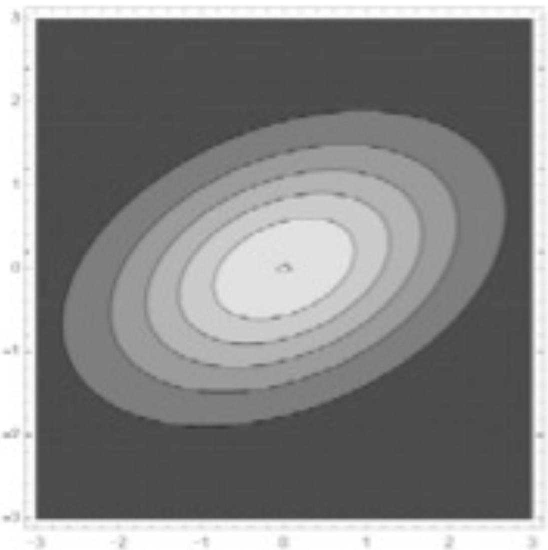
$$p(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2 \cdot \pi)^{D/2}} \cdot \frac{1}{|\Sigma|^{1/2}} \cdot \exp\left(-\frac{1}{2} \cdot (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} \cdot (\mathbf{x} - \boldsymbol{\mu})\right)$$

where

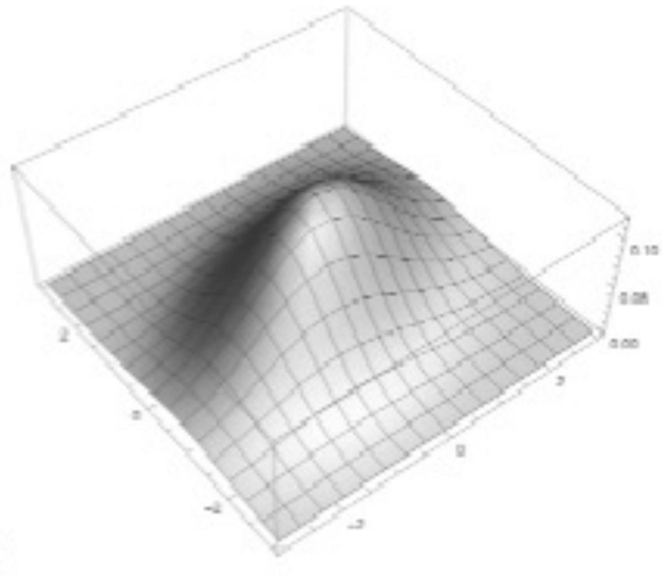
- $\boldsymbol{\mu}$ is the D dimensional mean vector
- Σ is a $D \times D$ covariance matrix
- $|\Sigma|$ is the determinant of Σ



(a)



(b)



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

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

- (b) Three dimensional plot of the Gaussian.

Covariance Matrix

- A position c_{ij}=Σ_{ij} of this matrix measures the tendency of two features, x_i and x_j, to vary in the same direction, for N features indexed by k

$$c_{ij} = \frac{\sum_{k=1}^{N}(x_{k,i} - \overline{x_i}) \cdot (x_{k,j} - \overline{x_j})}{N-1}$$

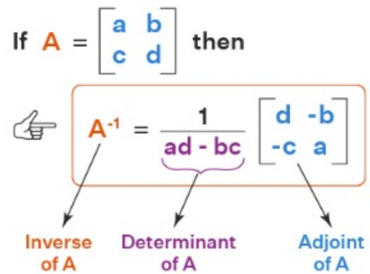
  - with $\overline{x_i}$ and $\overline{x_j}$ being the arithmetic mean of the two variables of the sample
  - Covariances are symmetric; $c_{ij} = c_{ji}$ and, so, the resulting covariance matrix Σ is symmetric and positive-definite

$$\Sigma = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mm} \end{pmatrix}$$

Multivariate Gaussian: example

Approximate a multivariate Gaussian distribution using the following points: $\{(-2,2)^T, (-1,3)^T, (0,1)^T, (-2,1)^T\}$

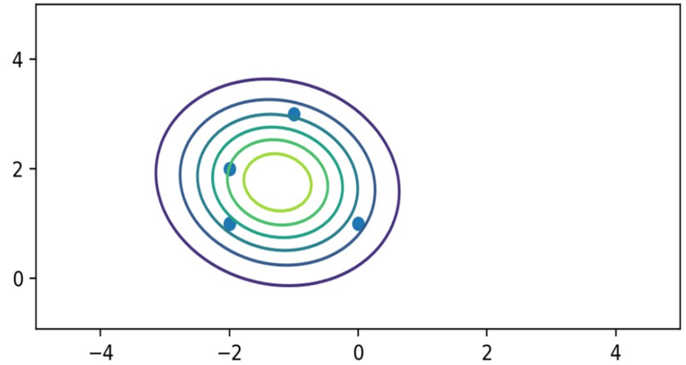
- $\mu = \frac{1}{4}\left(\begin{bmatrix}-2\\2\end{bmatrix} + \begin{bmatrix}-1\\3\end{bmatrix} + \begin{bmatrix}0\\1\end{bmatrix} + \begin{bmatrix}-2\\1\end{bmatrix}\right) = \begin{bmatrix}-1.25\\1.75\end{bmatrix}$
- $c_{12} = c_{21} = \frac{(-2+1.25)(2-1.75)+(-1+1.25)(3-1.75)+(0+1.25)(1-175)+(-2+1.25)(1-1.75)}{3}$ =-0.83
- $c_{11} = \frac{(-2+1.25)^2+(-1+1.25)^2+(0+1.25)^2+(-2+1.25)^2}{3}$ =0.92
- $c_{22} = \frac{(2-1.75)^2+(3-1.75)^2+(1-175)^2+(1-1.75)^2}{3}$ = 0.92
- $\Sigma = \begin{pmatrix}c_{11} & c_{21}\\ c_{12} & c_{22}\end{pmatrix} = \begin{pmatrix}0.92 & -0.083\\ -0.083 & 0.92\end{pmatrix}.$
- $\Sigma^{-1} = \begin{pmatrix}1.1 & 0.1\\ 0.1 & 1.1\end{pmatrix}.$ Det(Σ)=| Σ|=0.833Type equation here.



$$N(\mathbf{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{2/2}\sqrt{0.083}} exp\left(-\frac{1}{2}\left(\begin{bmatrix}x_1\\x_2\end{bmatrix} - \begin{bmatrix}-1.25\\1.75\end{bmatrix}\right)^T \begin{bmatrix}1.1 & 0.1\\0.1 & 1.1\end{bmatrix}\left(\begin{bmatrix}x_1\\x_2\end{bmatrix} - \begin{bmatrix}-1.25\\1.75\end{bmatrix}\right)\right)$$

Multivariate Gaussian: example

- What is the shape of the previous 2-dimensional Gaussian?
  - fixing $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ inspection…



- What is the probability of observing (0,0)?

- $N\left(\begin{bmatrix}0\\0\end{bmatrix} \middle| \mu, \Sigma\right) = \frac{1}{2\pi\sqrt{0.083}} exp\left(-\frac{1}{2}\left(\begin{bmatrix}0\\0\end{bmatrix} - \begin{bmatrix}-1.25\\1.75\end{bmatrix}\right)^T \begin{bmatrix}1.1 & 0.1\\0.1 & 1.1\end{bmatrix}\left(\begin{bmatrix}0\\0\end{bmatrix} - \begin{bmatrix}-1.25\\1.75\end{bmatrix}\right)\right) = 0.0145$

Bayesian optimal classifier: example

- Consider a population of 100 individuals
  - 30 individuals have phenotype A, 30 have B, and remaining ones have C
  - the expression of three genes (variables) are characterized by the following 3-dimensional Gaussians

$$N_A\left(\boldsymbol{\mu}_A = \begin{bmatrix}0.375\\0.875\\0.25\end{bmatrix}, \boldsymbol{\Sigma}_A = \begin{bmatrix}3.41 & 1.34 & 2.6\\1.34 & 2.125 & 1.18\\2.6 & 1.18 & 2.8\end{bmatrix}\right), N_B\left(\boldsymbol{\mu}_B = \begin{bmatrix}0.5\\0.125\\0.875\end{bmatrix}, \boldsymbol{\Sigma}_B = \begin{bmatrix}0.286 & 0.07 & -0.07\\0.07 & 0.125 & 0.018\\-0.07 & 0.018 & 0.125\end{bmatrix}\right), N_C\left(\boldsymbol{\mu}_C = \begin{bmatrix}0\\-0.125\\0.125\end{bmatrix}, \boldsymbol{\Sigma}_C = \begin{bmatrix}1.7 & 1.14 & 1\\1.14 & 1.55 & 0.73\\1 & 0.73 & 0.98\end{bmatrix}\right)$$

$p(A) = \frac{30}{100}$, $p(B) = \frac{30}{100}$, $p(C) = \frac{40}{100}$, prior, *called mixture parameters*

classify observations $\mathbf{x}_1 = [0, 1.1, -0.8]$

$p(\mathbf{x}_1|N_A)$=0.019. $p(\mathbf{x}_1|N_B) = 5.4E - 14.$ $p(\mathbf{x}_1|N_C)$=0.0088

$p(\mathbf{x}_1, N_A) = p(A)p(\mathbf{x}_1|N_A)$, $p(x, N_B)$= $p(B)p(\mathbf{x}_1|N_B)$, $p(\mathbf{x}_1, N_c)$= $p(C)p(\mathbf{x}_1|N_c)$

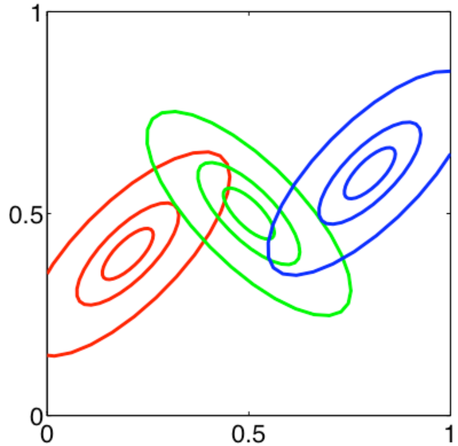
$p(\mathbf{x}_1)$= $p(\mathbf{x}_1, N_A)$+$p(x, N_B)$+ $p(\mathbf{x}_1, N_c)$

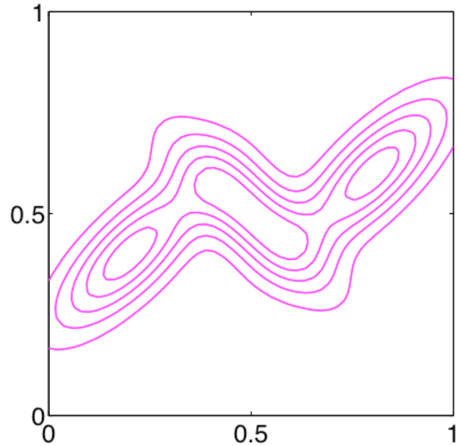
- $p(A|\mathbf{x}_1) = \frac{p(A)p(\mathbf{x}_1|N_A)}{p(\mathbf{x}_1)} = \frac{0.0057}{0.0057+0+0.0035}$ =*0.619565*, $p(B|\mathbf{x}_1) = \frac{p(B)p(\mathbf{x}_1|N_B)}{p(\mathbf{x}_1)} = \frac{0}{0.0057+0+0.0035}$ =0,
- $p(C|\mathbf{x}_1) = \frac{p(C)p(\mathbf{x}_1|N_C)}{p(\mathbf{x}_1)} = \frac{0.0035}{0.0057+0+0.0035} = 0.380435$

$\mathbf{x}_1$ is classified with phenotype A

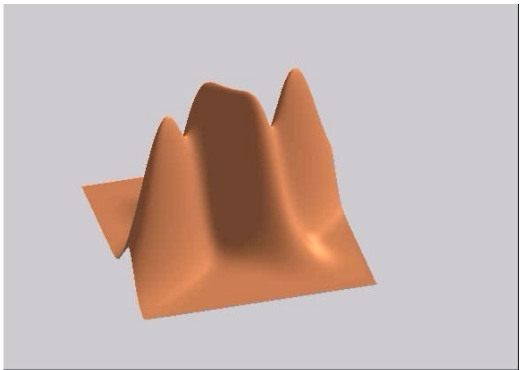
Example: Mixture of 3 Gaussians *k*

$p(\mathbf{x}|k)$





$p(\mathbf{x})$



$p(\mathbf{x})$

Bayes optimal classifier

- **Advantages**
  - when data distributions are well-approximated, provides highly **accurate** results
  - priors can be easily neglected to not bias posteriors

- **Disadvantages**
  - requires a good amount of data to estimate joint distributions
    - impracticable in the presence of **high-dimensional data**
  - can be computationally **expensive**
    - discrete data: need to compute the posterior probability for every hypothesis
    - numeric data: need to approximate distributions
      - e.g. fitting multivariate Gaussians can be expensive due covariance matrix inversion

Joint distribution

- A joint distribution for toothache, cavity, catch, *dentist‘s probe catches in my tooth* ☹
  - we need to know the conditional probabilities of the conjunction of toothache and cavity
  - what can a dentist conclude if the probe catches in the aching tooth?

$$P(cavity \mid toothache \wedge catch) = \frac{P(toothache \wedge catch \mid cavity)P(cavity)}{P(toothache \wedge cavity)}$$

  - **Problem?**
    - For *n* possible variables there are 2^n possible combinations

	toothache		no toothache	
	catch	no catch	catch	no catch
cavity	0.108	0.012	0.072	0.008
no cavity	0.016	0.064	0.144	0.576

Conditional independence

- Once we know that the patient has cavity we do not expect the probability of the probe catching to depend on the presence of toothache
  - **independence** $P(catch \mid cavity \wedge toothache) = P(catch \mid cavity)$

    $P(toothache \mid cavity \wedge catch) = P(toothache \mid cavity)$

- The decomposition of large probabilistic domains into weakly connected subsets via conditional independence is one of the most important developments in the recent history of AI

$$P(a \wedge b) = P(a)P(b)$$

$$P(toothache, catch, cavity, Weather = cloudy) =$$

$$= P(Weather = cloudy)P(toothache, catch, cavity)$$

$$P(a \mid b) = P(a)$$

$$P(b \mid a) = P(b)$$

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:

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

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

$$= \arg\max_{v_j \in V} P(a_1, a_2 \ldots a_n | v_j) P(v_j)$$

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

Estimating probabilities in small samples

- We have estimated probabilities by the times the event is observed, n_c, over total opportunities, n
  - poor estimates when n_c is very small
  - **problem**: what if none of the training instances with target value v_j have attribute value a_i? → n_c is 0!
- when n_c is very small: $\hat{P}(a_i|v_j) = \frac{n_c + mp}{n + m}$ $v_{NB} =_{v_j \in V} P(v_j) \prod_i \hat{P}(a_i|v_j)$
    - n is number of training examples for which $v = v_j$
    - n_c number of examples for which $v = v_j$ and $a = ai$
    - p is the prior estimate
    - m is the weight given to prior (i.e. number of "virtual" examples)

Naïve Bayes: comments

- **Advantages**
  - easy to implement, good results obtained in most of the cases
  - The decomposition of large probabilistic domains into weakly connected subsets via conditional independence is one of the most important developments in the recent history of AI

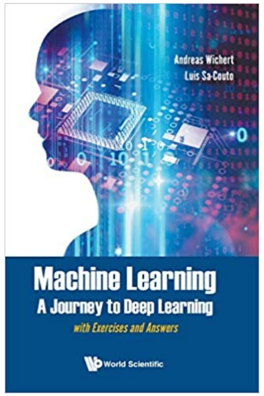
- Conditional independence assumption is often violated
- ...but it works surprisingly well anyway

Naïve Bayes: comments

- **Disadvantages**
  - Assumption: class conditional independence , therefore loss of accuracy
  - Practically, dependencies exist among variables
  - E.g., hospitals: patients: Profile: age, family history etc
    Symptoms: fever, cough etc., Disease: lung cancer, diabetes etc
  - Dependencies among these cannot be modeled by Naïve Bayesian Classifier

  - How to deal with these dependencies?
    - Bayesian - Belief Networks

Literature



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