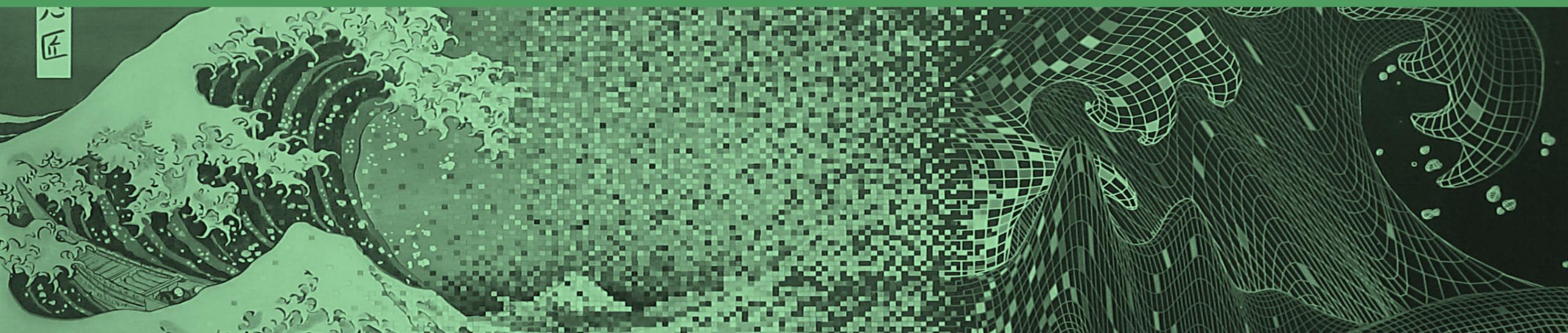


# Bayesian learning

Probability theory and Bayesian models

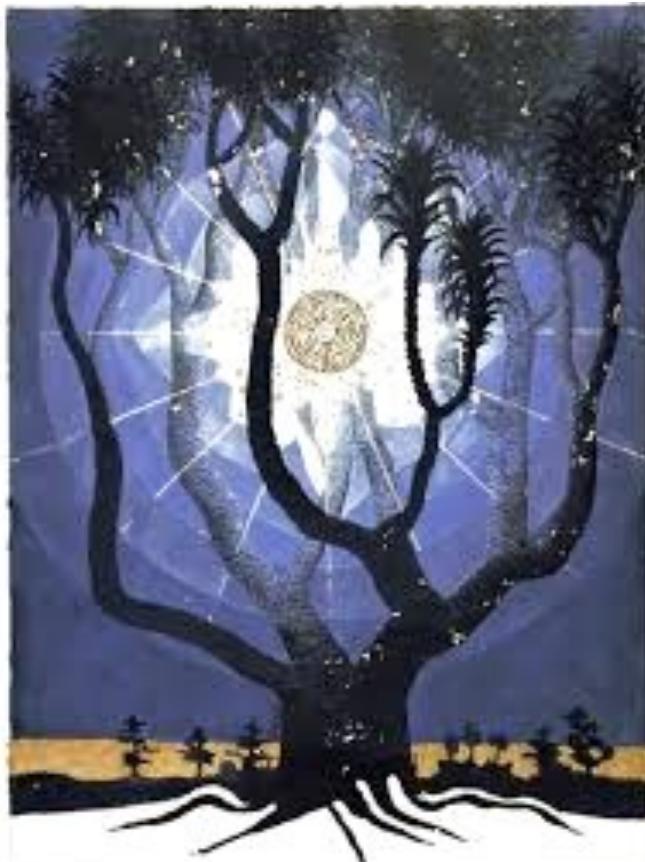


# Outline



- **Probability theory**
  - prior and posterior probability
  - maximum a posterior (MAP) and maximum likelihood (ML)
- **Bayes optimal classifier**
  - Bayesian learning in discrete spaces
  - Bayesian learning in numeric data spaces
- **Naïve Bayes classifier**
  - conditional independence
  - classification with naïve Bayes
  - estimating probabilities in small samples

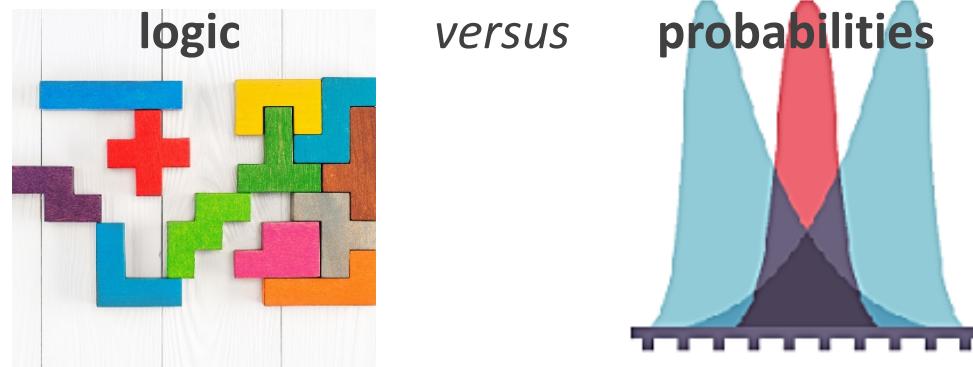
# Outline



- **Probability theory**
  - prior and posterior probability
  - maximum a posterior (MAP) and maximum likelihood (ML)
- Bayes optimal classifier
  - Bayesian learning in discrete spaces
  - Bayesian learning in numeric data spaces
- Naïve Bayes classifier
  - conditional independence
  - classification with naïve Bayes
  - estimating probabilities in small samples

# Uncertainty

- A key concept in the field in machine learning is ***uncertainty***
  - noise on measurements
  - finite size of data sets
- **Probability theory** provides a consistent framework to quantify and handle uncertainty
  - a central foundation in pattern recognition



# Kolmogorov's axioms of probability (1933)

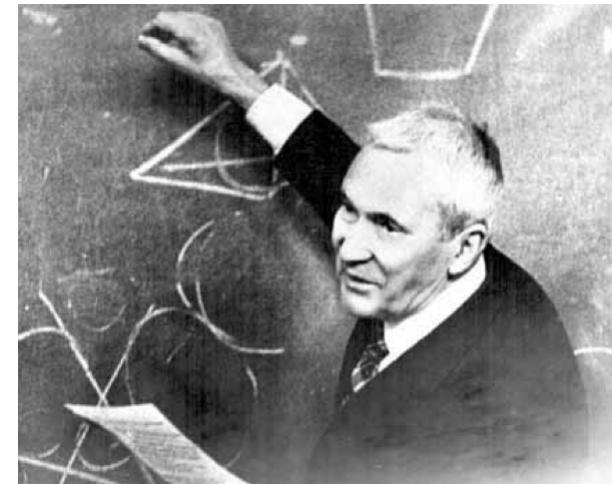
- To each statement  $a$ , a numerical degree of “belief” between 0 and 1 is assigned

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

$$p(\text{true}) = 1, \quad p(\text{false}) = 0$$

- The probability of disjunction is given by

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



# Where do numerical degrees of belief come from?

- Humans *believe* in a subjective viewpoint from *experience*
  - this approach is called **Bayesian**
- [subjectivist] For a finite sample we can *estimate* the probability of a given phenomenon
  - count the *frequency* of the event in a *sample*
  - **frequentist** approach
  - do not know the true value because we cannot access the whole population of events
- [objectivist] From the true nature of the universe, e.g. the probability of heads in a fair coin is 0.5
  - **Platonic world** of ideas!
  - we can never verify whether a fair coin exists

# Posterior probability

- If  $\Omega$  is the set of all possible events,  $p(\Omega) = 1$ 
  - $card(\Omega)$  is the number of elements of the set  $\Omega$
  - consider occurrences  $a, b \subseteq \Omega$ , then

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

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

- ... we get  $p(a|b) = \frac{p(a \wedge b)}{p(b)}$

# Bayes' rule

- From...

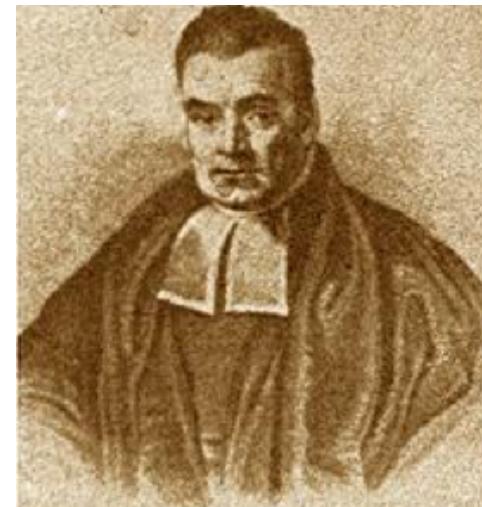
$$p(a|b) = \frac{p(a \wedge b)}{p(b)} \quad p(b|a) = \frac{p(a \wedge b)}{p(a)}$$

- ... we can infer the Bayes' rule

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

Reverent Thomas Bayes (1702-1761)

He set down his findings on probability in "Essay Towards Solving a Problem in the Doctrine of Chances" (1763)



# Law of total probability

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

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

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

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

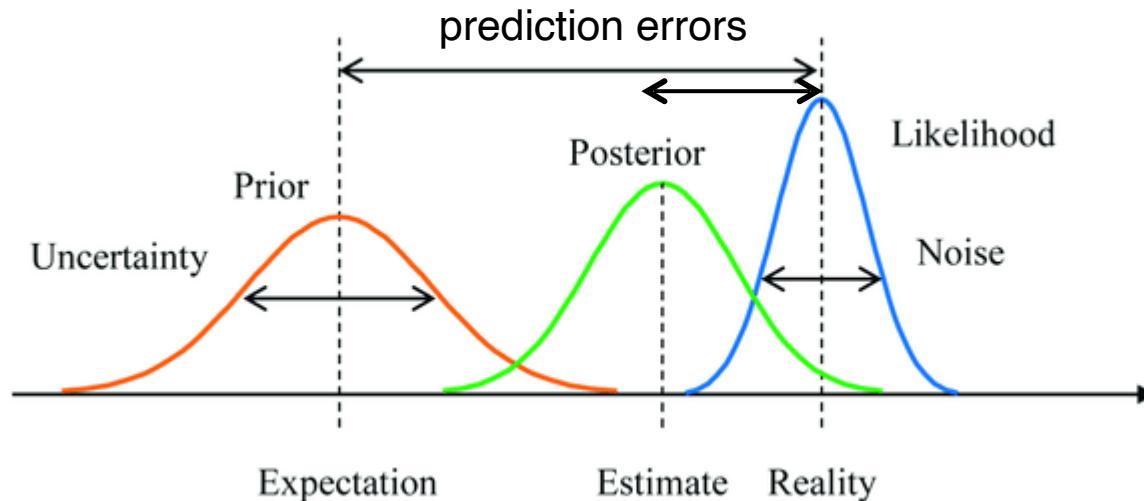
## Rules of probability

*sum rule*       $p(X) = \sum_Y p(X, Y)$

*product rule*    $p(X, Y) = p(Y|X)p(X)$

# Conditional probability: prior and posterior

- **prior** probability: before the evidence is obtained
  - $p(a)$  the prior probability that the proposition is true, e.g.  $p(cavity) = 0.1$
- **posterior** probability: after the evidence is obtained
  - $p(a|b)$ , the probability given that we know b, e.g.,  $p(cavity|toothache) = 0.8$



# Bayes theorem

- Bayes rule can be used to determine the prior total probability  $p(h)$  of hypothesis  $h$  given data  $D$ 
  - *example:* what is the probability of infection given certain symptoms?

$$p(h | D) = \frac{p(D | h)p(h)}{p(D)}$$

- - $p(h)$  = prior probability of hypothesis  $h$  – 20% of patients are infected
  - $p(D)$  = prior probability of training data  $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  – 40% of patients have identical symptoms
  - $p(D | h)$  = probability that the hypothesis  $h$  generates the data  $D$ 
    - probability that infection generates the symptoms? 70% infected patients show identical symptoms
  - $p(h | D)$  = probability of  $h$  given  $D$  –  $\frac{0.7 \times 0.2}{0.4} = 35\%$  probability of being infected given my symptoms

# Maximum a Posteriori (MAP)

- Given data  $D$  and Bayes rule assumption...
  - what is the most probable hypothesis  $h$  out of a set of possible hypothesis  $h_1, h_2, \dots$ ?
  - to determine the maximum a posteriori hypothesis  $h_{MAP}$  we maximize

$$h_{MAP} = \operatorname{argmax}_h p(h | D)$$

$$h_{MAP} = \operatorname{argmax}_h \frac{p(D | h)p(h)}{p(D)}$$

- $p(D)$  is the same for every hypothesis, hence...

$$h_{MAP} = \operatorname{argmax}_h p(D | h)p(h)$$

# Maximum Likelihood (ML)

- Assuming every hypothesis has the same probability
  - same priors  $p(h_1) = p(h_2) = \dots$
  - useful to prevent biases towards specific hypotheses (e.g., having or not having a disease)
  - we can further simplify and choose the maximum likelihood (ML) hypothesis

$$h_{ML} = \operatorname{argmax}_h p(D | h)$$

- $\text{posterior} \propto \text{likelihood} \times \text{prior}$
- $p(D | h)$  is evaluated using the observed data  $D$  and is called ***likelihood function***

# Outline



- Probability theory
  - prior and posterior probability
  - maximum a posterior (MAP) and maximum likelihood (ML)
- Bayes optimal classifier
  - Bayesian learning in discrete spaces
  - Bayesian learning in numeric data spaces
- Naïve Bayes classifier
  - conditional independence
  - classification with naïve Bayes
  - estimating probabilities in small samples

# Bayesian Learning

- For each hypothesis  $h$  in  $H$ , calculate the posterior probability

$$p(h \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid h)p(h)}{p(\mathbf{x})}$$

- Return the hypothesis  $h_{MAP}$  with the highest posterior probability

$$h_{MAP} = \operatorname{argmax}_h p(h \mid \mathbf{x})$$

- **Exercise:** does a 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 present ... and a correct negative result ( – ) in only 97% of the cases in which the disease is not present*
  - *0.008 of the entire population have this cancer*

## Suppose a *positive* result is returned...

$$P(\text{cancer}) = 0.008 \quad P(\neg\text{cancer}) = 0.992$$

$$P(+|\text{cancer}) = 0.98 \quad P(-|\text{cancer}) = 0.02$$

$$P(+|\neg\text{cancer}) = 0.03 \quad P(-|\neg\text{cancer}) = 0.97$$

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

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

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

The result of Bayesian inference strongly depends on prior probabilities, which must be representative in order to apply the MAP

# Estimating $p(h)$

- Let us draw some principles to estimate

$$p(h | \mathbf{x}) = \frac{p(\mathbf{x} | h)p(h)}{p(\mathbf{x})}$$

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

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

- otherwise, estimate the prior base on the observed frequency

# Estimating $p(D | h)$

- If data is **discrete**:
  - probability of each possible occurrence using class-conditional probability mass function
    - we can use the frequentist approach introduced in the first lectures, e.g., if we observe that:
      1. in shift A: 2 out of 10 individuals with blue eyes and brown hair, and
      2. in shift B: only 1 out of 8 in shift B,
- then  $p(\mathbf{x} = [\text{blue eyes, brown}] | A) = 0.2$  and  $p(\mathbf{x} = [\text{blue eyes, brown}] | B) = 0.125$
- If data is **real-valued**:
  - probability based on class-conditional probability density function
    - we can use empirical or theoretical distributions, e.g., if we assume
      3. age and height to be independent and uniformly distributed in class A, where
      4. age~ $U(20,30)$  and height~ $U(150,180)$ ,
- then  $p(\mathbf{x} = [22.3\text{years, } 154\text{cm}] | A) = \frac{1}{10} \times \frac{1}{30}$
- If data is **mixed**:
  - similarly to the above cases, the probability is drawn from the class-conditional joint distribution

# Bayesian optimal classifier

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

$$h_{MAP} = \operatorname{argmax}_h p(h | \mathbf{x}_{new}) = \operatorname{argmax}_h \frac{p(\mathbf{x}_{new} | h)p(h)}{p(\mathbf{x}_{new})} = \operatorname{argmax}_h p(\mathbf{x}_{new} | h)p(h)$$

... where the hypotheses correspond to our classes

- we ignore the denominator as it does not alter decision

- The Bayesian classifier has as many parameters as:

- the number of priors minus 1
  - we can deduce one prior from the remaining ones (**really??**)
- the number of parameters associated with the class-conditional distributions,  $p(\mathbf{x} | h)$

# Bayesian optimal classifier: example

- Learning the Bayesian model given by priors and posteriors

- priors  $p(c = 0) = \frac{4}{7}$ ,  $p(c = 1) = 1 - p(c = 0) = \frac{3}{7}$

- for each combination of possible values, learn the posteriors

$$p(c = 0 \mid v_1 = 0, v_2 = A, v_3 = 0) = \frac{p(c = 0)p(v_1 = 0, v_2 = A, v_3 = 0 \mid c = 0)}{p(v_1 = 0, v_2 = A, v_3 = 0)}$$

$$p(c = 0 \mid v_1 = 0, v_2 = A, v_3 = 1) = \frac{p(c = 0)p(v_1 = 0, v_2 = A, v_3 = 1 \mid c = 0)}{p(v_1 = 0, v_2 = A, v_3 = 1)}$$

...

$$p(c = 1 \mid v_1 = 1, v_2 = C, v_3 = 1) = \frac{p(c = 1)p(v_1 = 1, v_2 = C, v_3 = 1 \mid c = 1)}{p(v_1 = 1, v_2 = C, v_3 = 1)}$$

- from data  $p(v_1 = 0, v_2 = A, v_3 = 0 \mid c = 0) = 1/3$ , ...,  $p(v_1 = 1, v_2 = C, v_3 = 1 \mid c = 1) = 2/4$

- Now we can classify new observations, e.g.  $\mathbf{x}_{\text{new}} = [1, C, 1]$

- $p(c = 1 \mid v_1 = 1, v_2 = C, v_3 = 1) = \frac{1}{p(v_1 = 1, v_2 = C, v_3 = 1)} \times p(c = 1)p(v_1 = 1, v_2 = C, v_3 = 1 \mid c = 1) = \frac{1}{p(v_1 = 1, v_2 = C, v_3 = 1)} \times \frac{4}{7} \times \frac{2}{4}$

- $p(c = 0 \mid v_1 = 1, v_2 = C, v_3 = 1) = \frac{1}{p(v_1 = 1, v_2 = C, v_3 = 1)} \times p(c = 0)p(v_1 = 1, v_2 = C, v_3 = 1 \mid c = 0) = \frac{1}{p(v_1 = 1, v_2 = C, v_3 = 1)} \times \frac{3}{7} \times \frac{1}{4}$

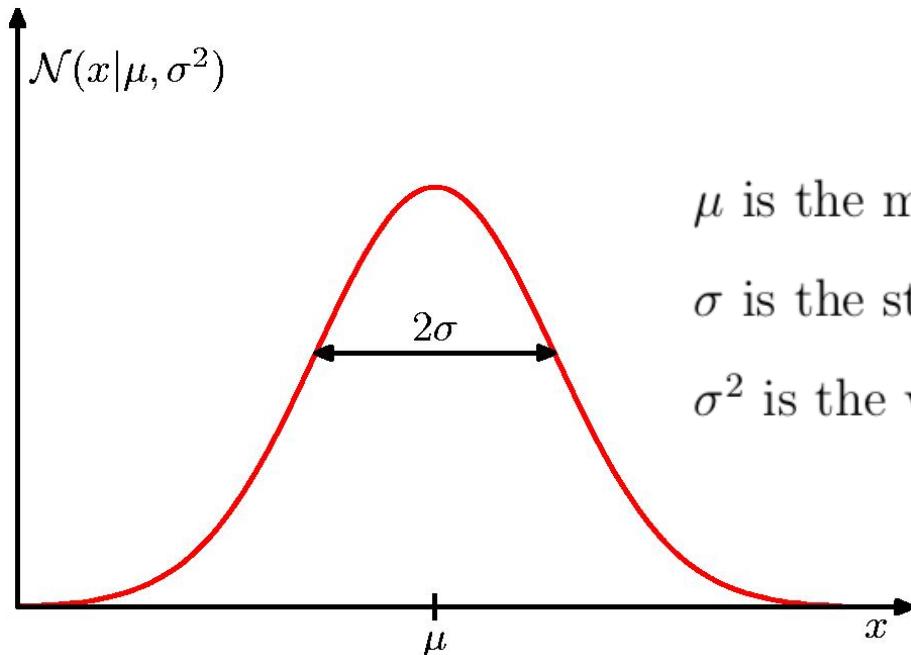
- $\mathbf{x}_{\text{new}}$  is classified with class 1

|   | v_1 | v_2 | v_3 | Class |
|---|-----|-----|-----|-------|
| 1 | 1   | C   | 1   | 1     |
| 2 | 1   | C   | 1   | 0     |
| 3 | 0   | B   | 1   | 0     |
| 4 | 0   | A   | 0   | 0     |
| 5 | 1   | C   | 1   | 1     |
| 6 | 0   | B   | 1   | 1     |
| 7 | 0   | A   | 0   | 1     |

# Recall: Gaussian distribution

- Gaussian or normal distribution defined by the probability

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



$\mu$  is the mean

$\sigma$  is the standard deviation

$\sigma^2$  is the variance

$$\int_{-\infty}^{\infty} N(x | \mu, \sigma^2) dx = 1$$

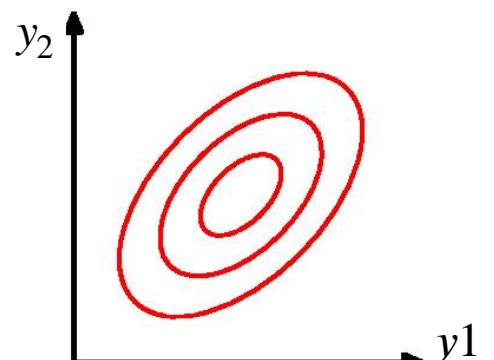
$$N(x | \mu, \sigma^2) > 0$$

# Multivariate Gaussian distribution in $m$ -dimensional spaces

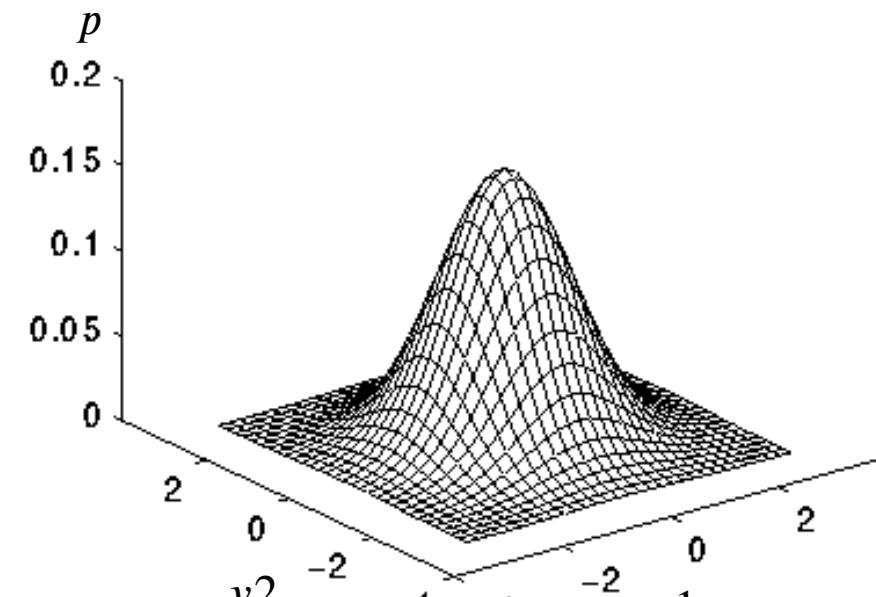
$$p(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma) = N(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{m/2} \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{u})^T \bullet \Sigma^{-1} \bullet (\mathbf{x} - \mathbf{u})\right)$$

where...

- $\mathbf{u}$  is the  $m$ -dimensional mean vector
- $\Sigma$  is a  $m \times m$  covariance matrix
- $|\Sigma|$  is the determinant of  $\Sigma$



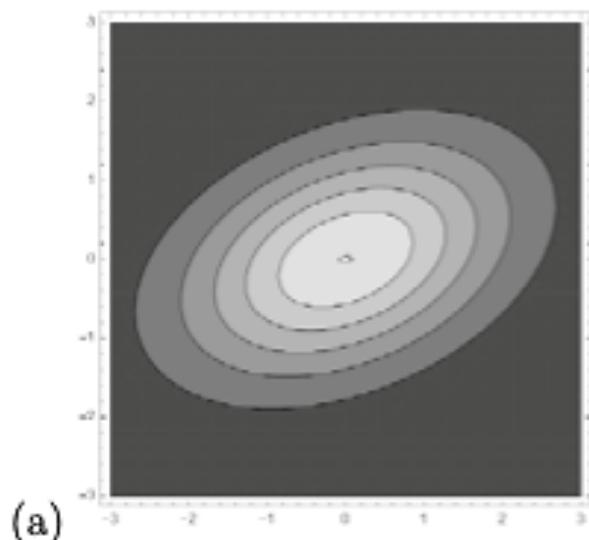
$m = 2$  dimensions



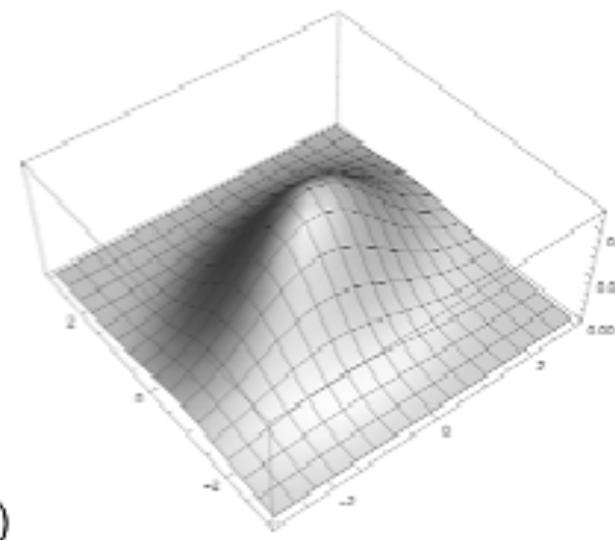
$m = 2$  dimensions

# Multivariate Gaussian distribution in $m$ -dimensional spaces

- Gaussian distribution over a  $m = 2$  dimensional space with
  - average  $\mu = (0,0)^T$
  - covariance matrix  $\Sigma = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 1 \end{pmatrix}$
  - (a) two-dimensional and (b) three-dimensional plots of the Gaussian

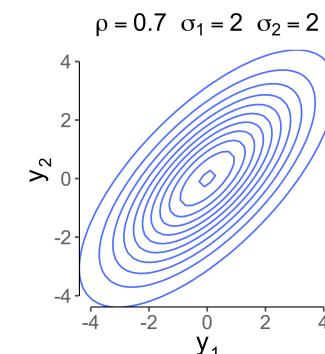
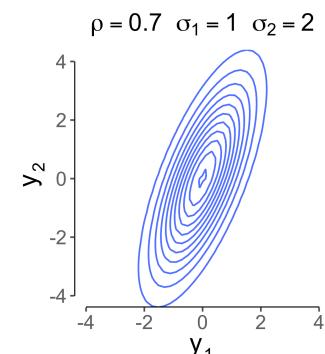
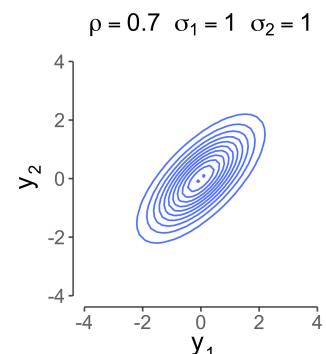
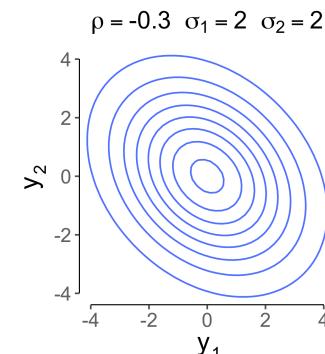
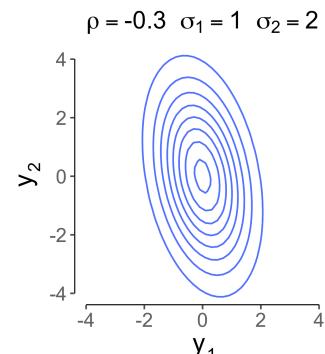
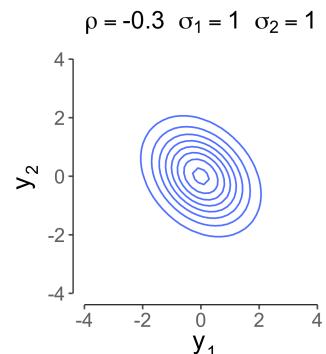
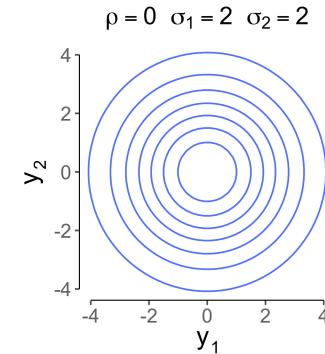
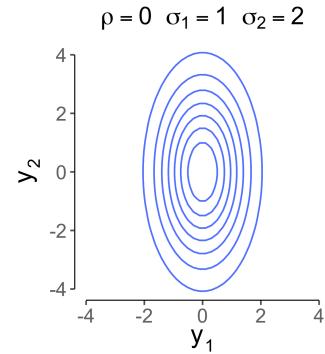
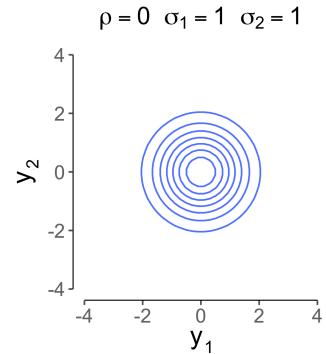


(a)



(b)

# Multivariate Gaussian: covariances



$$\Sigma = \begin{pmatrix} cov(y_1, y_1) & cov(y_1, y_2) \\ cov(y_1, y_2) & cov(y_2, y_2) \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

$$cov(y_1, y_2) = \frac{\sum_{i=1}^n (x_{1i} - \bar{y}_1) \bullet (x_{2i} - \bar{y}_2)}{n}$$

*n* population *n* versus  
sample *n* – 1

# Multivariate Gaussian: example

- Approximate a multivariate Gaussian distribution using the following points:

$$X = \{(-2,2), (-1,3), (0,1), (-2,1)\}$$

- maximum likelihood parameters

$$\mu = \frac{1}{4} \left( \begin{bmatrix} -2 \\ 2 \end{bmatrix} + \begin{bmatrix} -1 \\ 3 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} -2 \\ 1 \end{bmatrix} \right) = \begin{bmatrix} -1.25 \\ 1.75 \end{bmatrix}$$

$$\Sigma_{00} = \frac{1}{4-1} [(-2+1.25)(-2+1.25) + (-1+1.25)(-1+1.25) + (0+1.25)(0+1.25) + (-2+1.25)(-2+1.25)] \approx 0.92$$

$$\Sigma = \begin{pmatrix} \Sigma_{00} & \Sigma_{10} \\ \Sigma_{01} & \Sigma_{11} \end{pmatrix} = \begin{pmatrix} 0.92 & -0.083 \\ -0.083 & 0.92 \end{pmatrix}$$

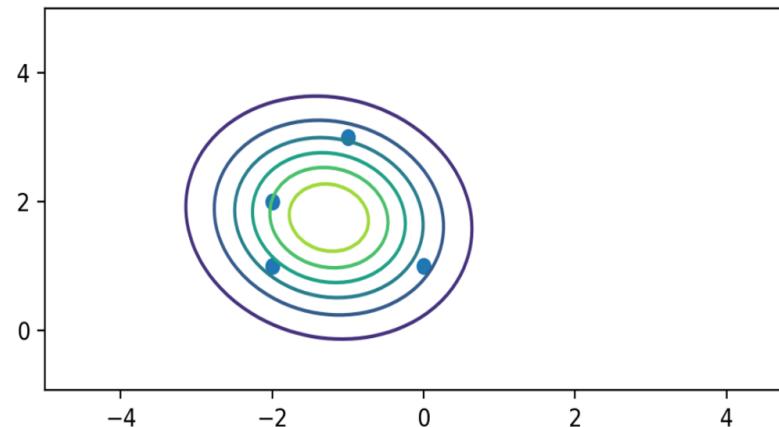
$$\Sigma^{-1} = \begin{pmatrix} 1.1 & 0.1 \\ 0.1 & 1.1 \end{pmatrix}$$

- we can write the Gaussian expression for a two-dimensional input  $\mathbf{x} = [x_1 \ x_2]^T$  as follows

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

# Multivariate Gaussian: example

- What is the shape of the previous 2-dimensional Gaussian?
  - fixing  $\mu$  and  $\Sigma$  inspection...



- What is the probability of observing  $\mathbf{x} = (0,0)^T$ ?

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

# 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

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

- classify observations  $\mathbf{x}_1 = [0, 1.1, -0.8]$  and  $\mathbf{x}_2 = [-0.3, 0.1, -0.3]$ 
  - to facilitate the calculus of class-conditional probabilities,  $p(\mathbf{x} | N)$ , we can use *scipy* or other package

$$\begin{aligned} - p(A|\mathbf{x}_1) &= \frac{p(A)p(\mathbf{x}_1|A)}{p(\mathbf{x}_1)} = \frac{1}{p(\mathbf{x}_1)} \times \frac{30}{100} \times 0.019 = 0.0057, p(B|\mathbf{x}_1) = \frac{1}{p(\mathbf{x}_1)} \times \frac{30}{100} \times 5.4E-14, p(C|\mathbf{x}_1) = \frac{1}{p(\mathbf{x}_1)} \times \frac{40}{100} \times 0.0088 = 0.0035 \\ - p(A|\mathbf{x}_2) &= \frac{p(A)p(\mathbf{x}_2|A)}{p(\mathbf{x}_2)} = \frac{1}{p(\mathbf{x}_2)} \times \frac{30}{100} \times 0.0266 = 0.008, p(B|\mathbf{x}_2) = \frac{1}{p(\mathbf{x}_2)} \times \frac{30}{100} \times 6.8E-6, p(C|\mathbf{x}_2) = \frac{1}{p(\mathbf{x}_2)} \times \frac{40}{100} \times 0.068 = 0.027 \end{aligned}$$

- $\mathbf{x}_1$  is classified with phenotype A and  $\mathbf{x}_2$  is classified with phenotype C

# Bayes optimal classifier

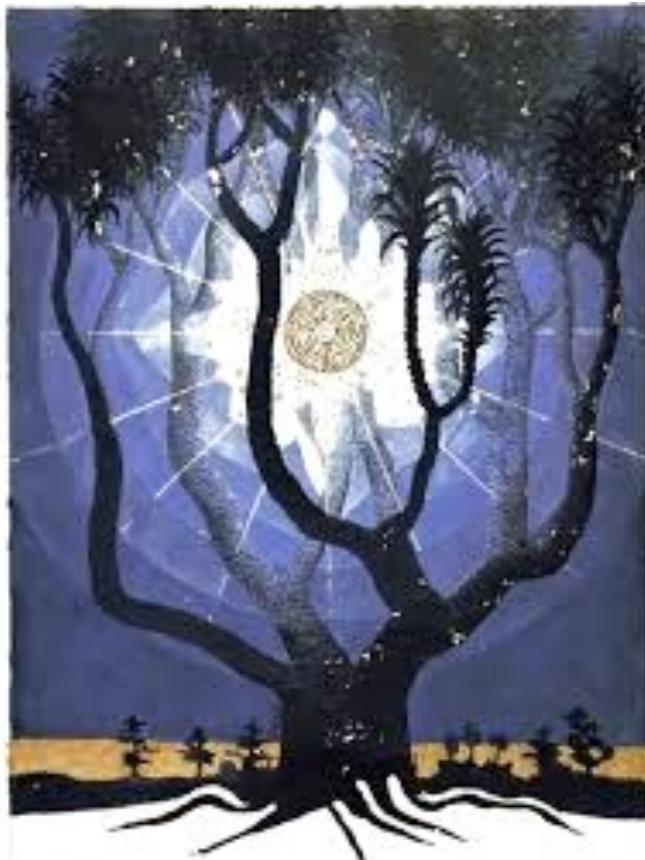
- **Advantages**

- 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

# Outline



- Probability theory
  - prior and posterior probability
  - maximum a posterior (MAP) and maximum likelihood (ML)
- Bayes optimal classifier
  - Bayesian learning in discrete spaces
  - Bayesian learning in numeric data spaces
- Naïve Bayes classifier
  - **conditional independence**
  - **classification with naïve Bayes**
  - **estimating probabilities in small samples**

# 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?

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

- **Problem?**
  - For  $m$  possible variables there are  $2m$  possible combinations

|        |       | toothache |          | no toothache |          |
|--------|-------|-----------|----------|--------------|----------|
|        |       | catch     | no catch | catch        | no catch |
| cavity | 0.108 | 0.012     | 0.072    | 0.008        |          |
|        | 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**

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

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

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

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

- 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

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

$$P(\text{toothache}, \text{catch}, \text{cavity}, \text{Weather} = \text{cloudy}) = P(\text{Weather} = \text{cloudy})P(\text{toothache}, \text{catch}, \text{cavity})$$

# Naive Bayes Classifier

- In contrast with optimal Bayes, naïve Bayes places conditional assumption!
  - known to work well, even the assumption is not true!
  - a single cause directly influence a number of conditionally independent effects

$$P(\text{cause}, \text{effect}_1, \text{effect}_2, \dots, \text{effect}_n) = P(\text{cause}) \prod_{i=1}^n P(\text{effect}_i | \text{cause})$$

- Along with decision trees, neural networks, nearest neighbors, a widely used learning approaches
- **When to use?**
  - moderate or large training set available
  - variables describing instances are (class-conditionally) independent
- **Successful applications**
  - classifying text documents
  - diagnosis

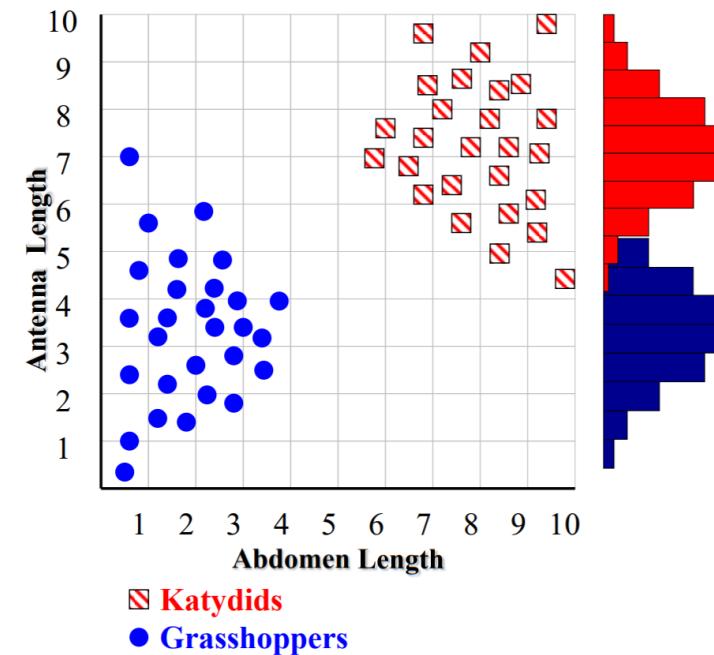
# Recall: Bayes classifier

- Assume target function  $f: X \longrightarrow \Sigma$ 
  - where each instance  $\mathbf{x}$  is described by features  $x_1, x_2, \dots, x_m$
  - the most probable value of  $f(\mathbf{x})$  is

$$h_{MAP} = \operatorname{argmax}_h p(h | \mathbf{x})$$

$$h_{MAP} = \operatorname{argmax}_h \frac{p(\mathbf{x} | h) \cdot p(h)}{p(\mathbf{x})}$$

$$h_{MAP} = \operatorname{argmax}_h p(\mathbf{x} | h) \cdot p(h)$$



# Naïve assumption

- Naive Bayes assumption:

$$p(h \mid \mathbf{x}) = p(h \mid y_1 = x_1, y_2 = x_2, \dots, y_m = x_m) = \prod_{j=1}^m p(h \mid y_j = x_j)$$

- ... yielding  
the naïve Bayes classifier:

$$h_{MAP} = \operatorname{argmax}_h p(h) \prod_{j=1}^m p(y_j = x_j \mid h)$$

- for each target value  $h$ : estimate  $p(h)$
- for each attribute value  $x_j$ : estimate  $p(x_j \mid h)$
- are frequentist views the only possibility to estimate these probabilities?

$$h_{MAP} = \operatorname{argmax}_h \hat{p}(h) \prod_{j=1}^m \hat{p}(y_j = x_j \mid h)$$

# Naïve Bayes: example

## Goal

- learn NB classifier and classify  $\mathbf{x}_{new} = (\text{age} \leq 30, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{rating} = \text{fair})$

| age    | income | student | credit rating | buys     |          |
|--------|--------|---------|---------------|----------|----------|
|        |        |         |               | computer | computer |
| 30     | high   | no      | fair          | no       | no       |
| 30     | high   | no      | excellent     | no       | no       |
| 31..40 | high   | no      | fair          | yes      | yes      |
| >40    | medium | no      | fair          | yes      | yes      |
| >40    | low    | yes     | fair          | yes      | yes      |
| >40    | low    | yes     | excellent     | no       | no       |
| 31..40 | low    | yes     | excellent     | yes      | yes      |
| 30     | medium | no      | fair          | no       | no       |
| 30     | low    | yes     | fair          | yes      | yes      |
| >40    | medium | yes     | fair          | yes      | yes      |
| 30     | medium | yes     | excellent     | yes      | yes      |
| 31..40 | medium | no      | excellent     | yes      | yes      |
| 31..40 | high   | yes     | fair          | yes      | yes      |
| >40    | medium | no      | excellent     | no       | no       |

# Naïve Bayes: example

- $p(\text{buys\_computer}=\text{yes})=9/14$  and  $p(\text{buys\_computer}=\text{no})=5/14$
- Class-conditional distributions  $p(x | h)$ 
  - $p(\text{age} \leq 30 | \text{buys\_computer}=\text{yes}) = 2/9 = 0.22$
  - $p(\text{age} \leq 30 | \text{buys\_computer}=\text{no}) = 3/5 = 0.6$
  - $p(\text{income}=\text{medium} | \text{buys\_computer}=\text{yes}) = 4/9 = 0.44$
  - $p(\text{income}=\text{medium} | \text{buys\_computer}=\text{no}) = 2/5 = 0.4$
  - $p(\text{student}=\text{yes} | \text{buys\_computer}=\text{yes}) = 6/9 = 0.67$
  - $p(\text{student}=\text{yes} | \text{buys\_computer}=\text{no}) = 1/5 = 0.2$
  - $p(\text{credit\_rating}=\text{fair} | \text{buys\_computer}=\text{yes}) = 6/9 = 0.67$
  - $p(\text{credit\_rating}=\text{fair} | \text{buys\_computer}=\text{no}) = 2/5 = 0.4$
- $\mathbf{x}_{\text{new}} = (\text{age} \leq 30, \text{income} = \text{medium}, \text{student}=\text{yes}, \text{rating}=\text{fair})$ 
  - $p(\mathbf{x} | c_1): p(\mathbf{x} | \text{buys\_computer}=\text{yes}) = 0.22 \times 0.44 \times 0.67 \times 0.67 = 0.043$
  - $p(\mathbf{x} | c_2): p(\mathbf{x} | \text{buys\_computer}=\text{no}) = 0.6 \times 0.4 \times 0.2 \times 0.4 = 0.019$
  - $p(\mathbf{x} | c_1) \times p(c_1): p(\mathbf{x} | \text{buys\_computer}=\text{yes}) \times p(\text{buys\_computer}=\text{yes}) = 0.028$
  - $p(\mathbf{x} | c_2) \times p(c_2): p(\mathbf{x} | \text{buys\_computer}=\text{no}) \times p(\text{buys\_computer}=\text{no}) = 0.007$
- $\mathbf{x}$  belongs to class  $\text{buys\_computer}=\text{yes}$ ,  $p(c_1 | \mathbf{x}) = 0.028/(0.028+0.007)$

# Estimating probabilities in small samples

- We have estimated probabilities based on the number  $n_a$  of times an event  $a$  occurs over all occurrences  $n$ 
  - however  $n_a$  estimate can be poor when  $n$  is small
  - **problem:** what if none of the training instances with a target  $h$  have the value  $a$ ?
    - $n_a$  is 0 which will rewrite the joint probability  $p(\mathbf{x} | h)$  as 0, irrespective of other features

- **Solution?**

- when  $n_a$  is very small:

$$\text{from } p(y_j = a | h) = \frac{n_a}{n} \quad \text{to} \quad \hat{p}(y_j = a | h) = \frac{n_a + d \cdot p(h)}{n + d}$$

- $n$  is number of training examples
    - $n_a$  number of  $a$  occurrences from observations with target  $h$
    - $p(h)$  is the prior estimate
    - $d$  is the weight given to the prior: the number of “virtual” examples, generally  $d \ll n$

# Naïve Bayes: comments

- **Advantages**

- easy to implement
  - good results obtained in most of the cases

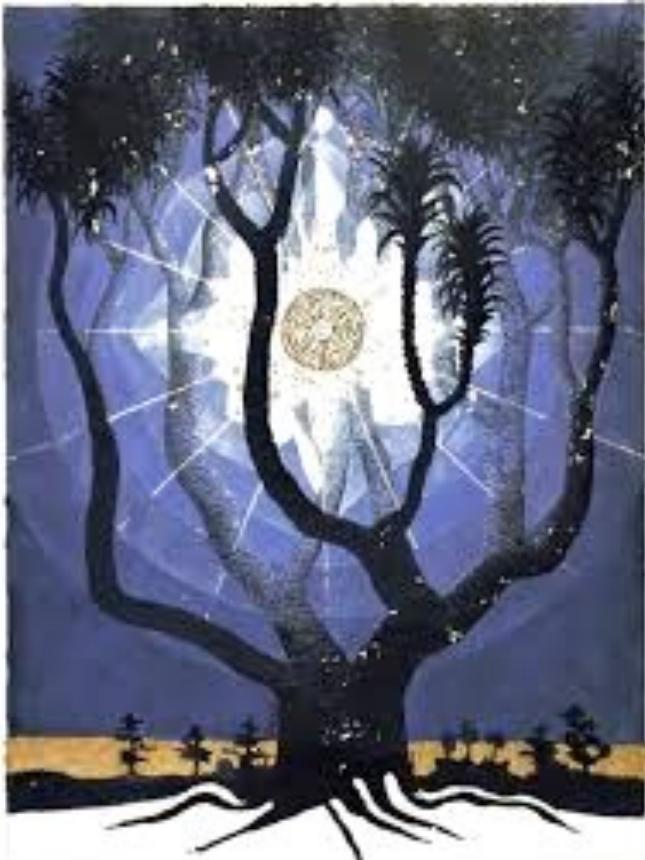
- **Disadvantages**

- class conditional independence *assumption*
  - loss of accuracy
  - dependencies exist among variables (e.g. symptoms)

- How to deal with these dependencies?

- **Bayesian Belief Networks**
  - relaxes independence assumption towards sets of variables

# Outline



- **Probability theory**
  - prior and posterior probability
  - maximum a posterior (MAP) and maximum likelihood (ML)
- **Bayes optimal classifier**
  - Bayesian learning in discrete spaces
  - Bayesian learning in numeric data spaces
- **Naïve Bayes classifier**
  - conditional independence
  - classification with naïve Bayes
  - estimating probabilities in small samples

# Thank You



[miguel.j.couceiro@tecnico.ulisboa.pt](mailto:miguel.j.couceiro@tecnico.ulisboa.pt)

[andreas.wichert@tecnico.ulisboa.pt](mailto:andreas.wichert@tecnico.ulisboa.pt)