Machine Learning (521289S) Estimation & Metrics

M.Sc. Antti Isosalo

Physiological Signal Analysis Team
Center for Machine Vision and Signal Analysis (CMVS)
University of Oulu

Recap: Bayesian Networks

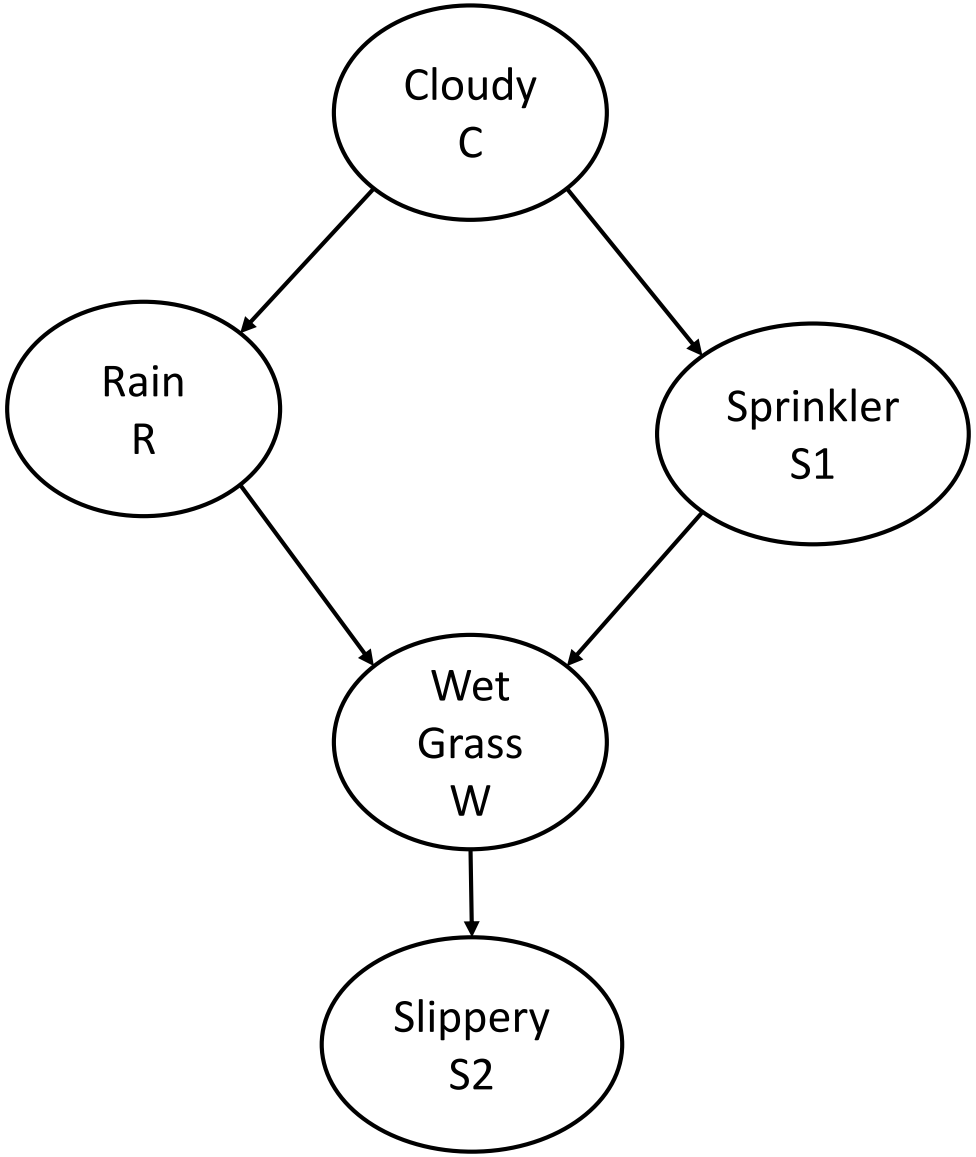
- In this case, the joint probability can be written as $P(C, R, S1, W, S2)$
- By reordering and using the conditional probability formula ($P(A,B) = P(A|B)P(B)$), we get

$$\begin{aligned} & P(S2, W, R, S1, C) \\ &= P(S2|W, R, S1, C)P(W, R, S1, C) \\ &= P(S2|W, R, S1, C)P(W|R, S1, C)P(R, S1, C) \\ &= P(S2|W, R, S1, C)P(W|R, S1, C)P(R|S1, C)P(S1, C) \\ &= P(S2|W, R, S1, C)P(W|R, S1, C)P(R|S1, C)P(S1|C)P(C) \end{aligned}$$

- Because Rain and Sprinkler nodes depend directly only on Cloudy node, and Wet Grass only depends directly on Rain and Sprinkler nodes, and Slippery node only on Wet Grass node, we get

$$\begin{aligned} & P(S2, W, R, S1, C) \\ &= P(S2|W)P(W|R, S1)P(R|C)P(S1|C)P(C) \\ &= P(C)P(R|C)P(S1|C)P(W|R, S1)P(S2|W) \end{aligned}$$

- Now we can calculate different things if we know the above probabilities.



What if we don’t know all the probabilities?

- Bayesian decision theory allows us to derive optimal statistical classifiers.
  - Optimal does not mean that we would not need to make compromises.
- These are based on probability theory and assume that a prior and a posterior probabilities are known.
- In most cases these probabilities are not known.
- There might only be data.
- One approach then becomes that we use this data in estimating the unknown probabilities and distributions.
- We can then try to apply the equations that we have learned earlier and use the resulting estimates as if they were the true values.
  - The estimation of the prior probabilities is perhaps quite straightforward, but the estimation of the class-conditional densities can be difficult.
- The problem becomes easier if we know the shape of the distribution $p(x|\omega_i)$ or if we can assume that certain distribution describes the structure of the data well.
- Then we only need to estimate the parameters of the distribution.
- The problem of parameter estimation is a classical one in statistics, and it can be approached, for example with Maximum likelihood estimation (MLE).

MLE from Bayes formula

- Suppose we have data $D = \{x_1, \dots, x_N\}$ and parameters θ.
- Here we consider θ to be a **discrete random variable**.
- The Bayes formula can then be written as

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}.$$

- NOTE: Notation $P(\theta|D)$ with capital letter P can be used here since we have **finite amount of data**.
- $P(D)$ does not depend on the method, but only data, having no effect on the magnitude of the a posterior probability $P(\theta|D)$ *in relation to a result calculated with a different* θ.
- If we assume that each combination of parameters θ is equally probable a priori wise, then $P(\theta)$ is a constant, having no effect on $P(\theta|D)$ either.
- Therefore $P(\theta|D) \propto P(D|\theta)$, and we get

$$\begin{aligned} P(D|\theta) &= P(x_1, x_2, \dots, x_N|\theta) \\ &= P(x_1|\theta)P(x_2|\theta) \dots P(x_N|\theta) \\ &= \textstyle\prod_{i=1}^{N} P(x_i|\theta), \end{aligned}$$

where we have assumed that we have **independent and identically distributed** (IID) samples with conditional independence.

- In MLE we maximize the product, the *likelihood* of θ,

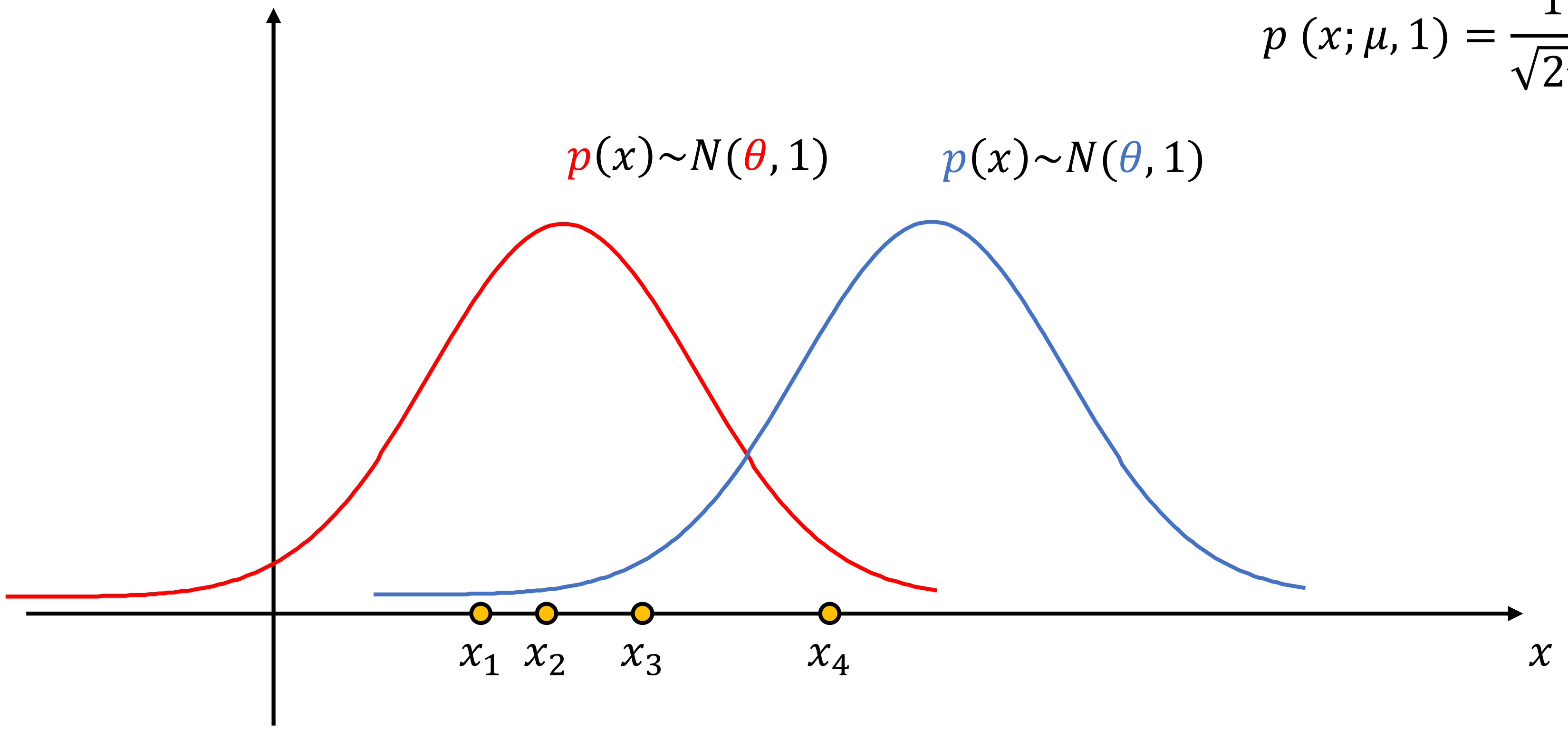
$$P(D|\theta) = \prod_{i=1}^{N} P(x_i|\theta)$$

- So, the **maximum likelihood estimate of $\boldsymbol{\theta}$** is $\hat{\theta}$ that maximizes $P(D|\theta)$.
- For calculations, log-likelihood is many times used for convenience.

MLE: How good is the fit?

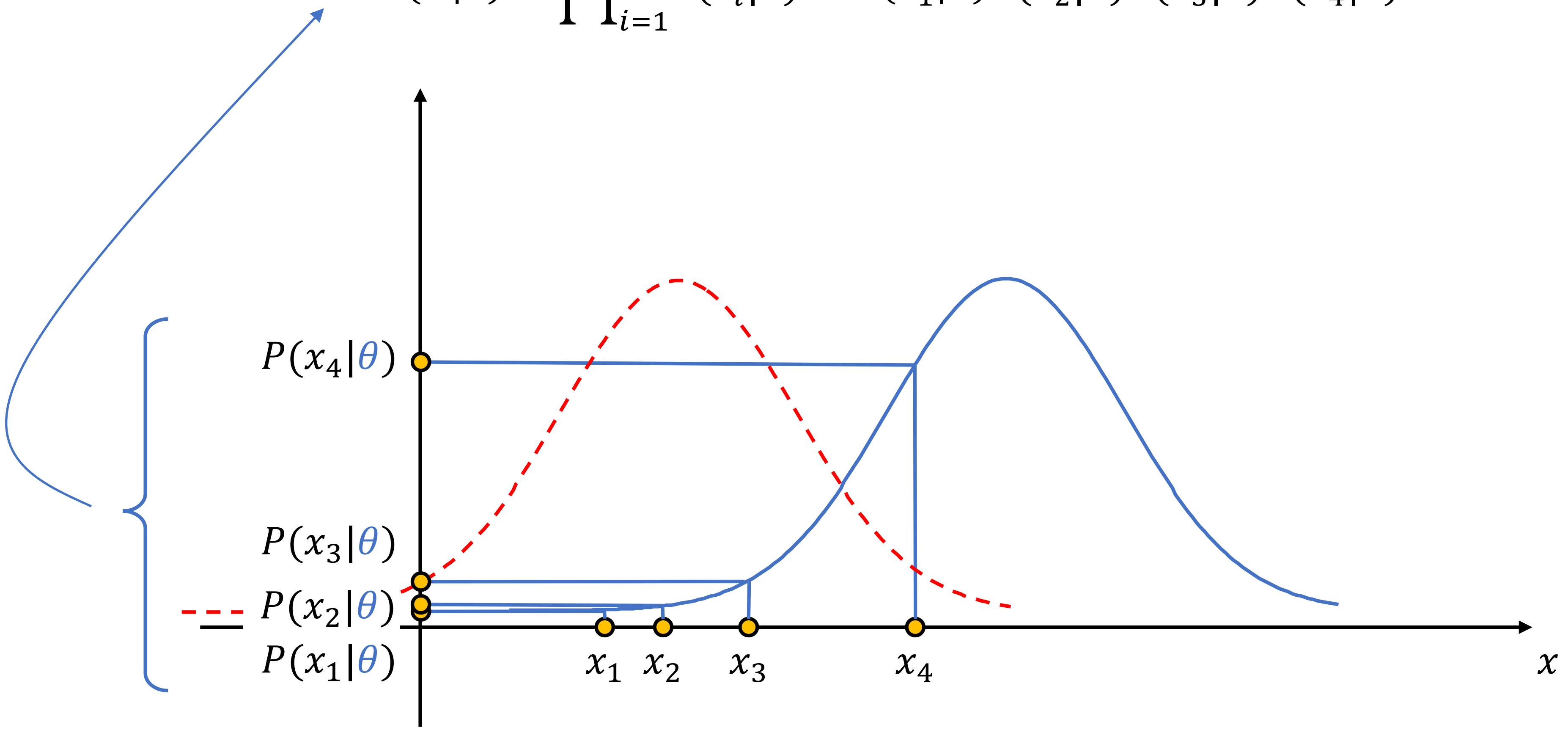
- Consider a situation where we have only had recourses to get *four* measurements: $D = \{x_1, x_2, x_3, x_4\}$

$$p\ (x; \mu, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1(x-\mu)^2}{2}}$$



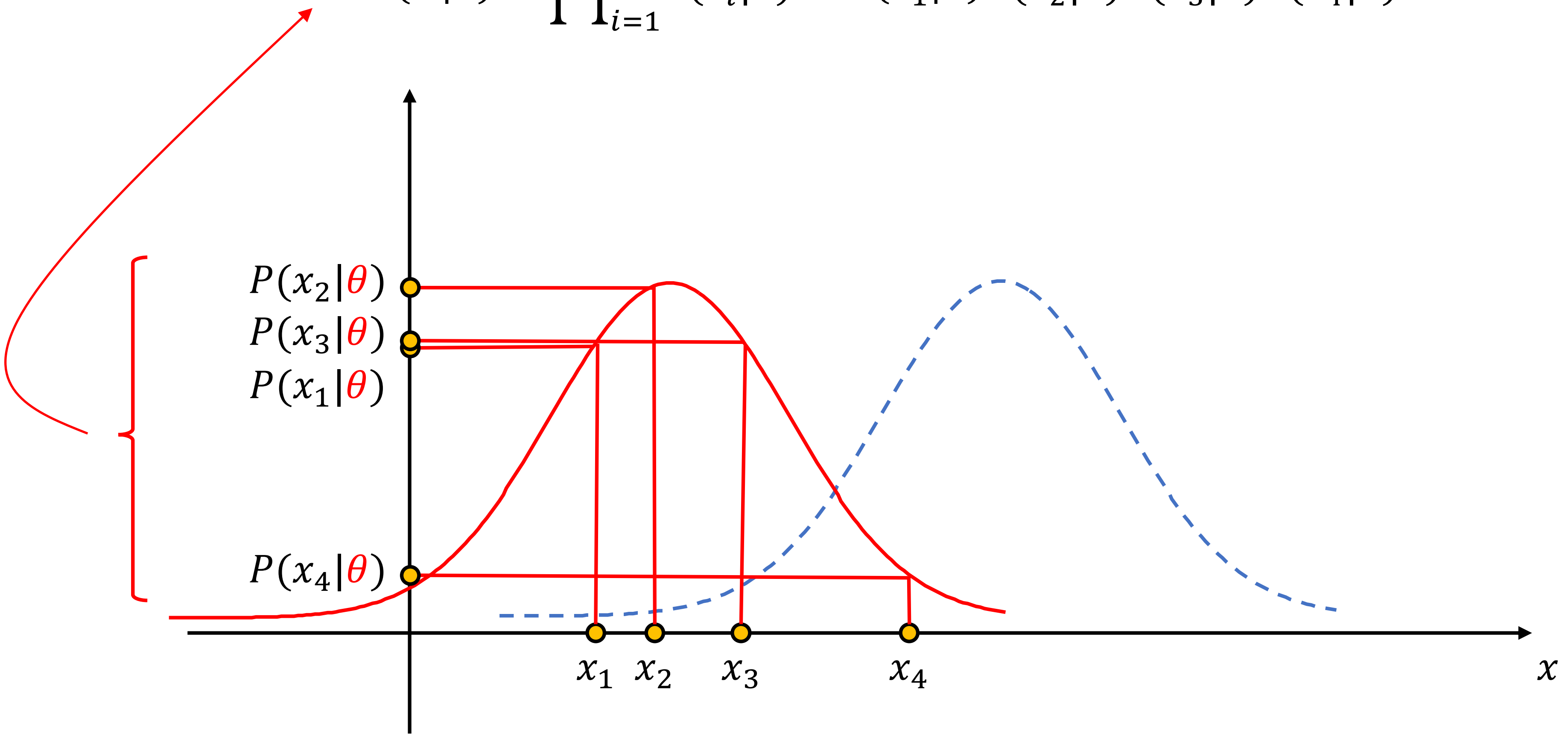
MLE: How good is the fit?

$$P(D|\theta) = \prod_{i=1}^{4} P(x_i|\theta) = P(x_1|\theta)P(x_2|\theta)P(x_3|\theta)P(x_4|\theta)$$



MLE: How good is the fit?

$$P(D|\theta) = \prod_{i=1}^{4} P(x_i|\theta) = P(x_1|\theta)P(x_2|\theta)P(x_3|\theta)P(x_4|\theta)$$



What if we don’t know the shapes of the distributions?

- If we don’t know the shapes of the distributions in the data, we can use nonparametric methods for density estimation.
  - We might have, e.g. data that has several peaks, several local maxima.
- These nonparametric methods do not require any assumptions of the distributions.
- Two common ways of estimating the density at certain point $\boldsymbol{x}$ are
  - k_n-nearest-neighbour method
  - Parzen window method
- With Parzen and k_n-nearest-neighbour methods we need some kind of metric to measure distance between samples within different classes

Metrics

- When we measure the distance between two vectors, we need to define a suitable *metric*, giving a scalar distance between the two vectors
    - In practical applications we measure, for example, similarity of two patterns (represented by vectors, say *feature vectors* in this context).

- A metric must have four properties: for all vectors $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$
- non-negativity: $D(\boldsymbol{a}, \boldsymbol{b}) \geq 0$
- reflexivity: $D(\boldsymbol{a}, \boldsymbol{b}) = 0$ if and only if $\boldsymbol{a} = \boldsymbol{b}$
- symmetry: $D(\boldsymbol{a}, \boldsymbol{b}) = D(\boldsymbol{b}, \boldsymbol{a})$
- triangle inequality: $D(\boldsymbol{a}, \boldsymbol{b}) + D(\boldsymbol{b}, \boldsymbol{c}) \geq D(\boldsymbol{a}, \boldsymbol{c})$