TECHNISCHE UNIVERSITÄT DARMSTADT

Statistical Machine Learning

Lecture 2: Bayesian Decision Theory

Simone Schaub-Meyer & Marcus Rohrbach
Department of Computer Science
TU Darmstadt

Summer Term 2025

Today's Objectives

- Make you understand how to make an optimal decision!
- Covered Topics:
  - Classification from a Bayesian point of view
  - Bayesian Optimal Decisions
  - Risk-based Classification
  - Probability Density Estimation (First Part)

Outline

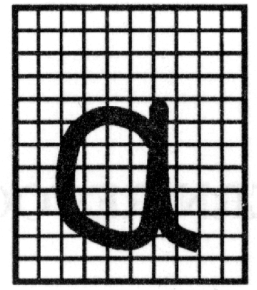
Outline

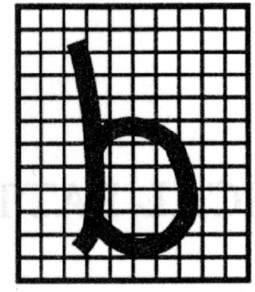
Statistical Methods

- All statistical methods in machine learning share the fundamental assumption that the data generation process is governed by the rules of probability
- The data is understood to be a set of random samples from some underlying probability distribution
- Keep in mind for future lectures: Even if we do not explicitly mention the existence of an underlying probability distribution the basic assumption about how the data is generated is always there!

TECHNISCHE UNIVERSITÄT DARMSTADT

Example: Handwritten Character Recognition





- *How to model this? Regression or Classification.*
  A: **Classification**
- **Goal**: classify a new letter such that the probability of misclassification is minimized

First concept: Class Priors

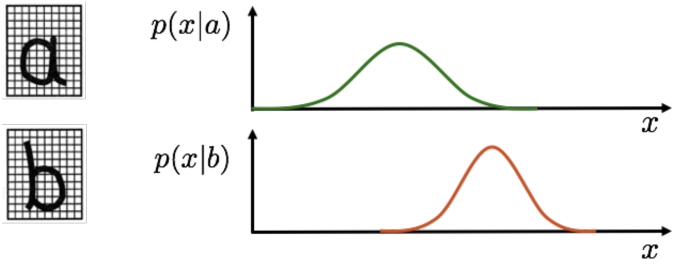
- The *a priori* probability of a data point belonging to a particular class is called the class prior
- What we can tell about the probability *before* seeing new data
- Example:
  - abaaa babaa aabba aaaaa
- *What are $p(a)$ and $p(b)$?*
  A:

$$\mathcal{C}_1 = a \quad p(\mathcal{C}_1) = 0.75$$
$$\mathcal{C}_2 = b \quad p(\mathcal{C}_2) = 0.25$$
$$\sum_k p(\mathcal{C}_k) = 1$$

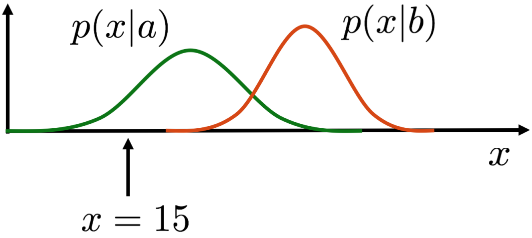
TECHNISCHE UNIVERSITÄT DARMSTADT

Second concept: Class conditional probabilities

- Probability (Likelihood) of making an observation x given that it comes from some class $\mathcal{C}_k$
- Here, x is often a feature vector, which measures/describes certain properties of the input data, e.g. number of black pixels, aspect ratio, ...

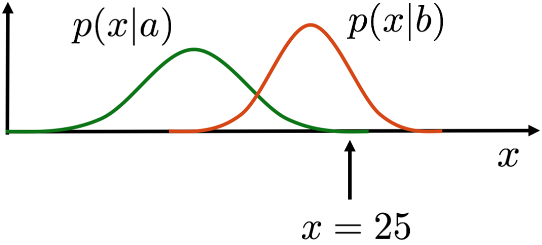


Example: Class conditional probabilities



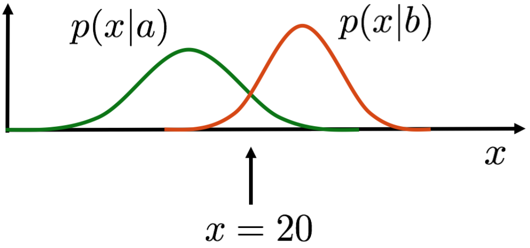
- How do we decide which class the data point belongs to?
- Since $p(x|b)$ is much smaller than $p(x|a)$, we should decide for class a.

Example: Class conditional probabilities



- How do we decide which class the data point belongs to?
- Since $p(x|a)$ is much smaller than $p(x|b)$, we should now decide for class b.

Example: Class conditional probabilities



- How do we decide which class the data point belongs to?
- Assuming the previous priors $p(a) = 0.75$ and $p(b) = 0.25$, we should decide for class a
- How can we formalize this?

TECHNISCHE UNIVERSITÄT DARMSTADT

Third Concept: Class posterior probabilities

- We want to find the *a posteriori* probability (posterior), i.e. the probability of class $\mathcal{C}_k$ given the observation x
- Bayes' Theorem

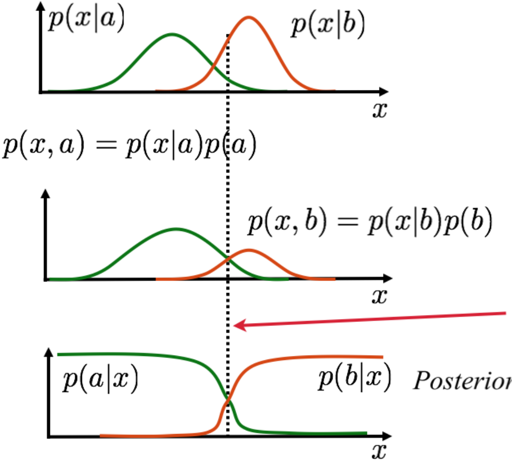
$$p\left(\mathcal{C}_k|x\right) = \frac{p\left(x|\mathcal{C}_k\right)p\left(\mathcal{C}_k\right)}{p\left(x\right)} = \frac{p\left(x|\mathcal{C}_k\right)p\left(\mathcal{C}_k\right)}{\sum_j p\left(x|\mathcal{C}_j\right)p\left(\mathcal{C}_j\right)}$$

- Interpretation

$$Posterior = \frac{Likelihood \times Prior}{Normalization\ Factor}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Bayesian Decision Theory



Likelihood

Likelihood × Prior

$$\textit{Posterior} = \frac{\textit{Likelihood} \times \textit{Prior}}{\textit{Normalization factor}}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Bayesian Decision Theory

- Why is it called this way?
  - To some extent, because it involves applying Bayes' rule
  - But this is not the whole story...
  - The real reason is that it is built on so-called Bayesian probabilities

TECHNISCHE UNIVERSITÄT DARMSTADT

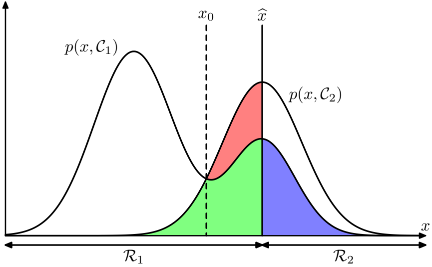
Bayesian Probabilities

- Probability is not just interpreted as a frequency of a certain event happening
- Rather, it is seen as a degree of belief in an outcome
- Only this allows us to assert a prior belief in a data point coming from a certain class
- Even though this might seem easy to accept to you now, this interpretation was quite contentious in statistics for a long time

TECHNISCHE UNIVERSITÄT DARMSTADT

Bayesian Decision Theory

- Goal: Minimize the probability of misclassification



$$
\begin{aligned}
p(\text{error}) &= p(x \in \mathcal{R}_1, \mathcal{C}_2) + p(x \in \mathcal{R}_2, \mathcal{C}_1) \\
&= \int_{\mathcal{R}_1} p(x, \mathcal{C}_2)\,\mathrm{d}x + \int_{\mathcal{R}_2} p(x, \mathcal{C}_1)\,\mathrm{d}x \\
&= \int_{\mathcal{R}_1} p(x|\mathcal{C}_2)\,p(\mathcal{C}_2)\,\mathrm{d}x + \int_{\mathcal{R}_2} p(x|\mathcal{C}_1)\,p(\mathcal{C}_1)\,\mathrm{d}x
\end{aligned}
$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Bayesian Decision Theory

- Optimal Decision rule
  - Decide for $\mathcal{C}_1$ if

$$p(\mathcal{C}_1|x) > p(\mathcal{C}_2|x)$$

  - This is equivalent to

$$\begin{aligned} \frac{p(x|\mathcal{C}_1)\,p(\mathcal{C}_1)}{p(x)} &> \frac{p(x|\mathcal{C}_2)\,p(\mathcal{C}_2)}{p(x)} \\ p(x|\mathcal{C}_1)\,p(\mathcal{C}_1) &> p(x|\mathcal{C}_2)\,p(\mathcal{C}_2) \end{aligned}$$

  - Which results in the Likelihood-Ratio Test:

$$\frac{p(x|\mathcal{C}_1)}{p(x|\mathcal{C}_2)} > \frac{p(\mathcal{C}_2)}{p(\mathcal{C}_1)}$$

  - A classifier obeying this rule is called a Bayes Optimal Classifier

TECHNISCHE UNIVERSITÄT DARMSTADT

Generalization to more than two classes

- Decide for class k whenever it has the greatest posterior probability of all classes

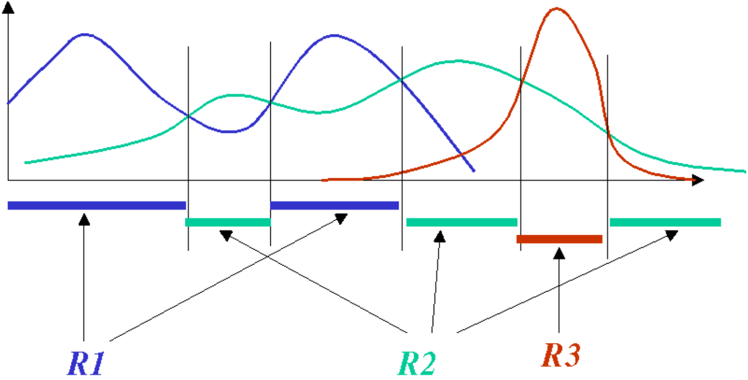
$$p\left(\mathcal{C}_k|x\right) > p\left(\mathcal{C}_j|x\right) \quad \forall j \neq k$$

- which again results in the Likelihood-Ratio Test:

$$\begin{aligned} p\left(x|\mathcal{C}_k\right) p\left(\mathcal{C}_k\right) &> p\left(x|\mathcal{C}_j\right) p\left(\mathcal{C}_j\right) \quad \forall j \neq k \\ \frac{p\left(x|\mathcal{C}_k\right)}{p\left(x|\mathcal{C}_j\right)} &> \frac{p\left(\mathcal{C}_j\right)}{p\left(\mathcal{C}_k\right)} \quad \forall j \neq k \end{aligned}$$

Visualization of the general case

- Decision regions: $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3, \ldots$



TECHNISCHE UNIVERSITÄT DARMSTADT

High Dimensional Features

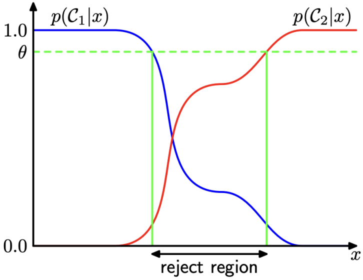
- So far we have only considered one-dimensional features, i.e., $x \in \mathbb{R}$
- We can use more features and generalize to an arbitrary D-dimensional feature space, i.e., $\boldsymbol{x} \in \mathbb{R}^D$
  - For example, in the salmon vs. sea-bass classification task

$$\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\intercal \in \mathbb{R}^2$$

  - Where x_1 is the width, and x_2 is the lightness
- The decision boundary we devised still applies to $\boldsymbol{x} \in \mathbb{R}^D$. We just need to use multivariate class-conditional densities $p\left(\boldsymbol{x}|\mathcal{C}_k\right)$

TECHNISCHE UNIVERSITÄT DARMSTADT

Reject Option



- Classification errors arise from decision regions where the largest posterior probability $p(\mathcal{C}_k|x)$ is significantly less than 1
  - Relatively high uncertainty about class memberships
  - For some applications, it may be better to reject an automatic decision entirely by introducing a decision threshold θ
  - Or use a dummy class "don't know" to which the system assigns all ambiguous cases

TECHNISCHE UNIVERSITÄT DARMSTADT

Reject Option

What are further reasons why we might want the machine learning model to abstain from making predictions?

A:

- **Insufficient training data**
- **Out-of-distribution test data**
- **Adversarial attacks**
- **Biased outputs**
- **Output not aligned with certain values**
- **High risk**
- **And many more ...**

TECHNISCHE UNIVERSITÄT DARMSTADT

Outline

2. Risk Minimization

- So far, we have tried to minimize the misclassification rate
- There are many cases when not every misclassification is equally bad
- Smoke detector
  - If there is a fire, we need to be very sure that we classify it as such
  - If there is no fire, it is ok to occasionally have a false alarm
- Medical diagnosis
  - If the patient is sick, we need to be very sure that we report them as sick
  - If they are healthy, it is ok to classify them as sick and order further testing that may help clarifying this up

Decisions with Loss Functions

- Differentiate between the possible decisions α_i and the possible true classes $\mathcal{C}_j$
- The loss may be asymmetric as in the medical diagnosis example

$$\text{loss}\left(\text{decision} = \text{healthy} \middle| \text{patient} = \text{sick}\right) >>$$
$$\text{loss}\left(\text{decision} = \text{sick} \middle| \text{patient} = \text{healthy}\right)$$

- Expected loss of making a decision α_i

$$R\left(\alpha_i|x\right) = \mathbb{E}_{\mathcal{C}_k \sim p(\mathcal{C}_k|x)}\left[\lambda\left(\alpha_i|\mathcal{C}_k\right)\right] = \sum_j \lambda\left(\alpha_i|\mathcal{C}_j\right) p\left(\mathcal{C}_j|x\right)$$

with loss function: $\lambda\left(\alpha_i|\mathcal{C}_j\right)$

Minimize the overall risk

- The expected loss of a decision is also called the risk of making a decision
- So, instead of minimizing the misclassification rate (recap)

$$
\begin{aligned}
p(\text{error}) &= p(x \in \mathcal{R}_1, \mathcal{C}_2) + p(x \in \mathcal{R}_2, \mathcal{C}_1) \\
&= \int_{\mathcal{R}_1} p(x, \mathcal{C}_2)\,\mathrm{d}x + \int_{\mathcal{R}_2} p(x, \mathcal{C}_1)\,\mathrm{d}x \\
&= \int_{\mathcal{R}_1} p(x|\mathcal{C}_2)\,p(\mathcal{C}_2)\,\mathrm{d}x + \int_{\mathcal{R}_2} p(x|\mathcal{C}_1)\,p(\mathcal{C}_1)\,\mathrm{d}x
\end{aligned}
$$

- We minimize the overall risk

$$
R(\alpha_i|x) = \mathbb{E}_{\mathcal{C}_k \sim p(\mathcal{C}_k|x)}\left[\lambda(\alpha_i|\mathcal{C}_k)\right] = \sum_j \lambda(\alpha_i|\mathcal{C}_j)\,p(\mathcal{C}_j|x)
$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Simple example

- 2 classes: $\mathcal{C}_1, \mathcal{C}_2$
- 2 decisions: α_1, α_2
- Loss function: $\lambda\left(\alpha_i|\mathcal{C}_j\right) = \lambda_{ij}$
- Expected loss (= risk R) of both decisions

$$
\begin{aligned}
R\left(\alpha_1|x\right) &= \lambda_{11} p\left(\mathcal{C}_1|x\right) + \lambda_{12} p\left(\mathcal{C}_2|x\right) \\
R\left(\alpha_2|x\right) &= \lambda_{21} p\left(\mathcal{C}_1|x\right) + \lambda_{22} p\left(\mathcal{C}_2|x\right)
\end{aligned}
$$

- Goal: Create a decision rule such that overall risk is minimized
  - *When to decide α_1?*
    A: **if** $R\left(\alpha_2|x\right) > R\left(\alpha_1|x\right)$

Risk-aware decision rule

$$
\begin{aligned}
R(\alpha_2|x) &> R(\alpha_1|x) \\
\lambda_{21} p(\mathcal{C}_1|x) + \lambda_{22} p(\mathcal{C}_2|x) &> \lambda_{11} p(\mathcal{C}_1|x) + \lambda_{12} p(\mathcal{C}_2|x) \\
(\lambda_{21} - \lambda_{11}) p(\mathcal{C}_1|x) &> (\lambda_{12} - \lambda_{22}) p(\mathcal{C}_2|x) \\
\frac{\lambda_{21} - \lambda_{11}}{\lambda_{12} - \lambda_{22}} &> \frac{p(\mathcal{C}_2|x)}{p(\mathcal{C}_1|x)} = \frac{p(x|\mathcal{C}_2)\, p(\mathcal{C}_2)}{p(x|\mathcal{C}_1)\, p(\mathcal{C}_1)} \\
\frac{p(x|\mathcal{C}_1)}{p(x|\mathcal{C}_2)} &> \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \frac{p(\mathcal{C}_2)}{p(\mathcal{C}_1)}
\end{aligned}
$$

- It is reasonable to assume that the loss of a correct decision is smaller than that of a wrong decision: $\lambda_{ij} > \lambda_{ii} \quad \forall j \neq i$

TECHNISCHE UNIVERSITÄT DARMSTADT

Risk Minimization with 0-1 Loss

- Risk-aware decision rule

$$\frac{p(x|C_1)}{p(x|C_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{p(C_2)}{p(C_1)}$$

- 0-1 loss function:

$$\lambda(\alpha_i|\mathcal{C}_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

- Decide for α_1 if

$$\frac{p(x|C_1)}{p(x|C_2)} > \frac{p(C_2)}{p(C_1)}$$

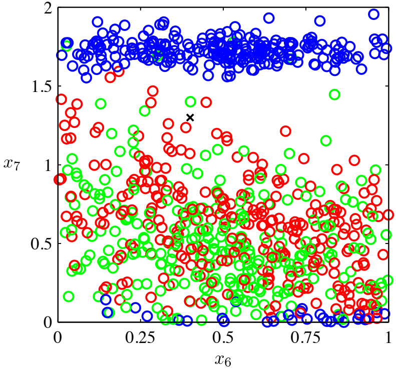
- The 0-1 loss leads to the same decision rule that minimized the misclassification rate

TECHNISCHE UNIVERSITÄT DARMSTADT

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Training Data



- How do we get the probability distributions from this so that we can classify with them?

TECHNISCHE UNIVERSITÄT DARMSTADT

Probability Density Estimation

- So far we have seen:
  - **Optimal Bayes Classification**, based on probability distributions $p(x|C_k)p(C_k)$
- The prior $p(C_k)$ is easy to deal with. We can "just count" the number of occurrences of each class in the training data
- We need to estimate/learn the **class-conditional probability density** $p(x|C_k)$
  - **Supervised training**: we know the input data points and their true labels (classes)
  - Estimate the density separately for each class C_k

TECHNISCHE UNIVERSITÄT DARMSTADT

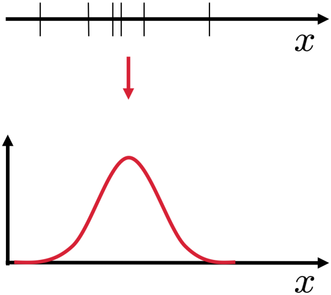
Probability Density Estimation

- **Remember:** The relationship between the outcomes of a random variable x and its probability $Pr(X = x)$ is referred to as the probability density, or simply the "density."

- Training data

$$x_1, x_2, x_3, \ldots$$

- Estimation $p(x)$



TECHNISCHE UNIVERSITÄT DARMSTADT

Types of Probability Density Estimation models

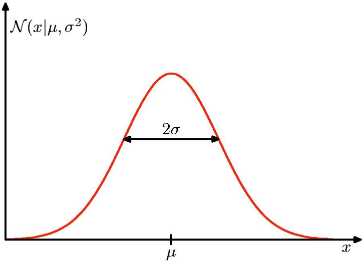
- **Parametric probability density estimation** involves selecting a common distribution and estimating the distribution parameters from data samples.
- **Non-parametric probability density estimation** involves fitting a model to the arbitrary distribution of the data, e.g., kernel density estimation – every known data point in the dataset is used as a parameter.
- **Mixture density models** are flexible models that combine parametric and non-parametric estimations.

Outline

Parametric Density Models

- Simple case: **Gaussian Distribution**

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



- Is governed by two parameters: mean and variance. If we know these parameters, we can fully describe $p(x)$

Parametric Density Models

- Notation for **parametric density models**

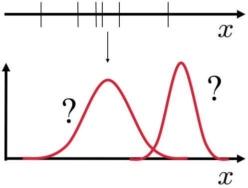
$$x \sim p(x|\theta)$$

- For the Gaussian distribution

$$\theta = (\mu, \sigma)$$
$$x \sim p(x|\mu, \sigma)$$

Parametric Density Models

- **Learning** means to estimate the parameters θ given the training data $\mathcal{D} = \{x_1, x_2, \ldots\}$



- **Likelihood** of θ is defined as the probability that the data $\mathcal{D}$ was generated from the probability density function with parameters θ

$$L(\theta) = p(\mathcal{D}|\theta)$$

Maximum Likelihood Method

- Consider a set of points $\mathcal{D} = \{x_1, \ldots, x_N\}$, we are interested in the likelihood of all data $p(\mathcal{D}|\theta)$.
- **Assumption**: the **data is i.i.d.** (independent and identically distributed)
  - The random variables x_1 and x_2 are independent if

$$P\left(x_1 \leq \alpha, x_2 \leq \beta\right) = P\left(x_1 \leq \alpha\right) P\left(x_2 \leq \beta\right) \quad \forall \alpha, \beta \in \mathbb{R}$$

  - The random variables x_1 and x_2 are identically distributed if

$$P\left(x_1 \leq \alpha\right) = P\left(x_2 \leq \alpha\right) \quad \forall \alpha \in \mathbb{R}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Maximum Likelihood Method

- Likelihood

$$L(\theta) = p(\mathcal{D}|\theta) = p(x_1, \ldots, x_N|\theta)$$

(using the i.i.d. assumption)

$$= p(x_1|\theta) \cdot \ldots \cdot p(x_n|\theta)$$

$$= \prod_{n=1}^{N} p(x_n|\theta)$$

- **Maximum Likelihood Estimation**: $\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} p(\mathcal{D}|\theta)$ seeks for the parameter $\hat{\theta}_{\mathrm{ML}}$, which best explains the data $\mathcal{D}$

- $\hat{\theta}_{\mathrm{ML}}$ is a **random variable** – it is an estimate based on the available dataset. Consequently, we are interested in its **mean value** (the **most probable value**) and its **variance**.

TECHNISCHE UNIVERSITÄT DARMSTADT

Maximum log-Likelihood Method

- It is more convenient and numerical stable to maximize the log-likelihood w.r.t. θ

$$LL(\theta) = \log L(\theta) = \log p(\mathcal{D}|\theta) = \log \prod_{n=1}^{N} p(x_n|\theta) = \sum_{n=1}^{N} \log p(x_n|\theta)$$

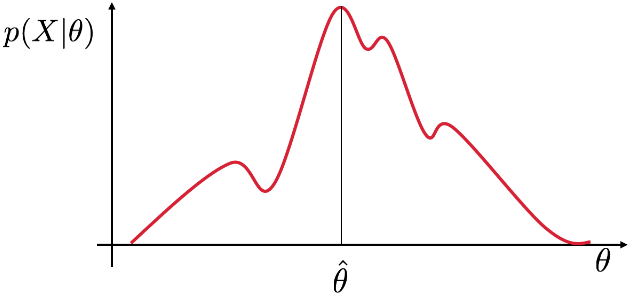
- Because the logarithm is monotonically increasing, it holds

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} \sum_{n=1}^{N} \log p(x_n|\theta) = \arg\max_{\theta} L(\theta)$$

- Maximizing a sum of terms is always easier than maximizing a product; cf., the difficulty of expressing the derivative of a long product of terms.

Likelihood Estimation

$$L(\theta) = p(\mathcal{D}|\theta) = \prod_{n=1}^{N} p(x_n|\theta)$$



Outline

5. Wrap-Up

Now, you know:

- The definition of class priors, class conditional probabilities and class posteriors
- How to use Bayes Theorem for classification
- How to calculate the probability of misclassification
- How to obtain optimal decisions using Bayes optimal classifier
- How to generalize decision making using multi-dimensional features and more than 2 classes
- The value of risk minimization and how it relates to misclassification
- Maximum Likelihood Method

Self-Test Questions

- How do we incorporate prior knowledge on the class distribution?
- How can we decide on classifying a query based on simple and general loss functions?
- What does “Bayes optimal” mean?
- How can we deal with 2 or more classes?
- How can we deal with high dimensional feature vectors?
- What are the equations for misclassification rate and risk?

Reading Assignments

To get a deeper understanding of today's topics:

- Bayesian Decision Theory: Bishop 2006, Chapter 1.5 or Murphy 2023, Chapter 5.1.1, 5.1.2
- The Bayesian idea: Lindholm 2022, Chapter 9.1

Next week:

- Probability Distributions: Bishop, Chapter 2
- Mixtures Models and EM: Bishop, Chapter 9