*** Applied Machine Learning Fundamentals ***
Bayesian Decision Theory

Daniel Wehner, M.Sc.

SAP SE / DHBW Mannheim

Winter term 2023/2024

Find all slides on GitHub (DaWe1992/Applied_ML_Fundamentals)

Lecture Overview

Agenda for this Unit

Section:

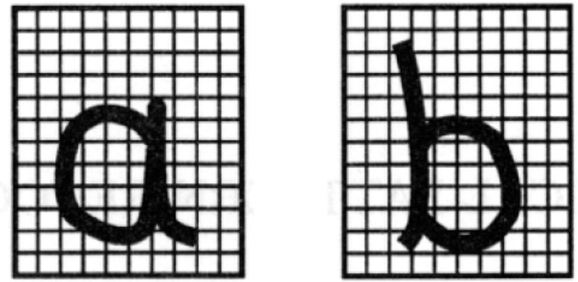
Bayesian Decision Theory

Statistical Methods

- Statistical methods assume that the process that 'generates' the data is governed by the **rules of probability**
- The data is understood to be a set of **random samples** from some underlying **probability distribution**
- This is the reason for the name **statistical machine learning**

The basic assumption about how the data is generated is always there, even if you don't see a single probability distribution!

Running Example: Optical Character Recognition (OCR)



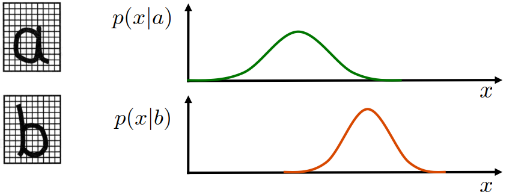
Goal: Classify a new letter so that the probability of a wrong classification is minimized

Class Conditional Probabilities

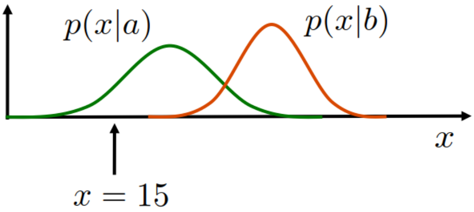
- First concept: **Class conditional probabilities**
- Probability of $\boldsymbol{x}$ given a specific class $\mathcal{C}_k$ is formally written as:

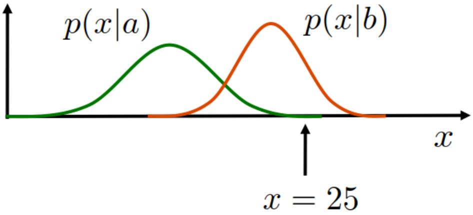
$$p(\boldsymbol{x}|\mathcal{C}_k) \in [0, 1] \tag{1}$$

- $\boldsymbol{x} \in \mathbb{R}^m$ is a feature vector, e. g. # black pixels, height-width ratio, ...



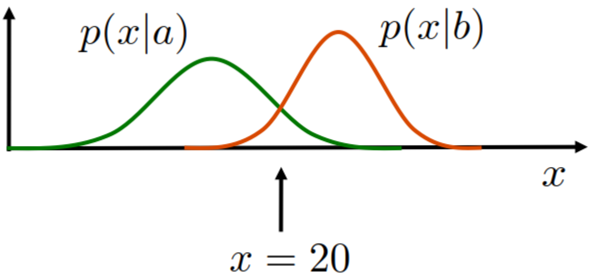
Class Conditional Probabilities (Ctd.)





If $x = 15$ we would predict class a, since $p(15|a) > p(15|b)$.
If $x = 25$ we would output class b, since $p(25|b) > p(25|a)$.

Class Conditional Probabilities (Ctd.)



We have a problem!

- **Which class should be chosen now?**
- The conditional probabilities are the same... ☠

Class Prior Probabilities

- Second concept: **Class priors**
- The prior probability of a data point belonging to a particular class $\mathcal{C}_k$

$$\mathcal{C}_1 \equiv a \qquad p(\mathcal{C}_1) = 0.75$$
$$\mathcal{C}_2 \equiv b \qquad p(\mathcal{C}_2) = 0.25$$

- By definition:
  - $0 \leqslant p(\mathcal{C}_k) \leqslant 1, \ \forall k$
  - The sum of all probabilities equals one: $\sum_{k=1}^{|\mathcal{C}|} p(\mathcal{C}_k) = 1$
- **The class prior is equivalent to a prior belief in the class label**

How would you decide now?

How to get the Prior Probabilities?

Count Count's advice:

Simply count the number of instances in each class!

Important

Bayes' Theorem

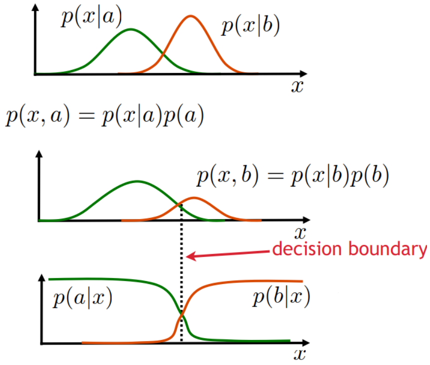
- What we actually want to compute: $p(\mathcal{C}_k|\boldsymbol{x}) \Rightarrow$ **Posterior probability**
- We can compute it by applying **Bayes' theorem**
- This is one of the **most important formulas (!!!)**

$$\overbrace{p(\mathcal{C}_k|\boldsymbol{x})}^{\text{Class posterior}} = \frac{\overbrace{p(\boldsymbol{x}|\mathcal{C}_k)}^{\text{Class cond.}} \cdot \overbrace{p(\mathcal{C}_k)}^{\text{Class prior}}}{\underbrace{p(\boldsymbol{x})}_{\text{Normalization term}}} = \frac{p(\boldsymbol{x}|\mathcal{C}_k) \cdot p(\mathcal{C}_k)}{\sum_{j=1}^{|\mathcal{C}|} p(\boldsymbol{x}|\mathcal{C}_j) \cdot p(\mathcal{C}_j)} \tag{2}$$

Calculation of the Posterior Probability

- By applying Bayes' theorem we can compute the posterior
- Simply plug ❶ and ❷ into Bayes' theorem
  1. Class prior probabilities
  2. Class conditional probabilities

We get the final **decision boundary**



a Priori vs. a Posteriori

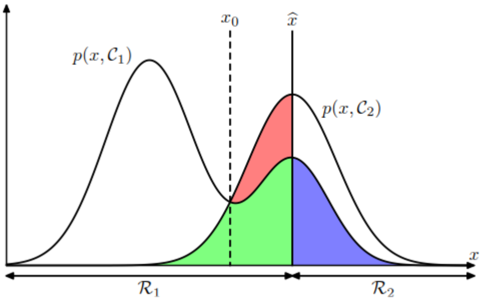
a Priori

A **belief** or conclusion **based on assumptions** or reasoning of some sort rather than actual experience or empirical evidence. Before actually encountering, experiencing, or observing a fact.

a Posteriori

A fact, belief, or argument that is **based on actual experience**, experiment, or observation.

Error Minimization



$$p(error) = p(x \in \mathcal{R}_1, \mathcal{C}_2) + p(x \in \mathcal{R}_2, \mathcal{C}_1)$$

$$= \overbrace{\int_{\mathcal{R}_1} p(x|\mathcal{C}_2) \cdot p(\mathcal{C}_2)\,\mathrm{d}x}^{\text{red + green area}} + \underbrace{\int_{\mathcal{R}_2} p(x|\mathcal{C}_1) \cdot p(\mathcal{C}_1)\,\mathrm{d}x}_{\text{blue area}}$$

Important

Bayes' optimal Classifier

- Decision rule:
  - Decide $\mathcal{C}_1$, if $p(\mathcal{C}_1|\boldsymbol{x}) > p(\mathcal{C}_2|\boldsymbol{x})$
  - This is equivalent to: *(we don't need the normalization)*

$$p(\boldsymbol{x}|\mathcal{C}_1) \cdot p(\mathcal{C}_1) > p(\boldsymbol{x}|\mathcal{C}_2) \cdot p(\mathcal{C}_2) \tag{3}$$

  - Which is in turn equivalent to:

$$\frac{p(\boldsymbol{x}|\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)} > \frac{p(\mathcal{C}_2)}{p(\mathcal{C}_1)} \tag{4}$$

- A classifier obeying this rule is called **Bayes' optimal Classifier**

Section:

(Multinomial) Naïve Bayes

A naïve Assumption

- We want to compute $p(\mathcal{C}_k|\boldsymbol{x})$. Recall Bayes' theorem:

Our first classification algorithm!

$$p(\mathcal{C}_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_k) \cdot p(\mathcal{C}_k)}{p(\boldsymbol{x})} \tag{5}$$

- Assumptions:
  - All features x_j are pairwise conditionally independent ($\Rightarrow$ **naïve**)

$$p(\boldsymbol{x}|\mathcal{C}_k) = p(x_1|\mathcal{C}_k) \cdot p(x_2|\mathcal{C}_k, x_1) \cdot p(x_3|\mathcal{C}_k, x_1, x_2) \cdot \ldots = \prod_{j=1}^{m} p(x_j|\mathcal{C}_k) \tag{6}$$

  - $p(\boldsymbol{x})$ is constant w. r. t. class label $\Rightarrow$ It is omitted

How to get the most probable Class?

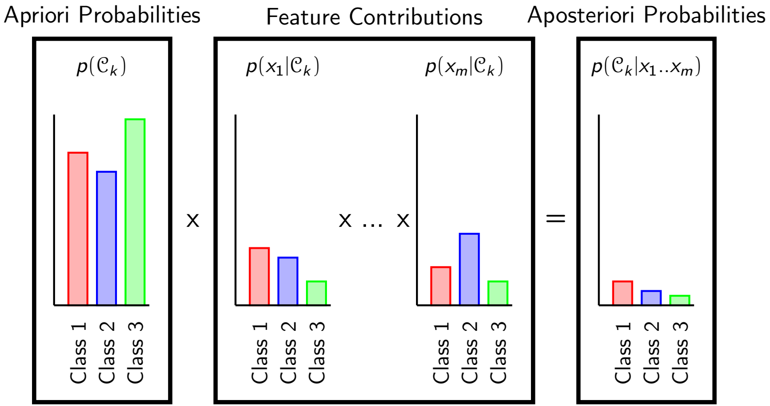
- **Given:**
  - New instance $\boldsymbol{x} = \langle x_1, x_2, \ldots, x_m \rangle$ to be classified
  - Finite set of κ classes $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_\kappa\}$
  - Labeled training data ($\Rightarrow$ supervised learning)
- **Wanted:** Most probable class $\mathcal{C}_{MAP}$ (maximum aposteriori) for $\boldsymbol{x}$:

$$\mathcal{C}_{MAP} = \underset{\mathcal{C}_k \in \{\mathcal{C}_1, \ldots, \mathcal{C}_\kappa\}}{\arg\max} \ \widehat{p}(\mathcal{C}_k | \boldsymbol{x}) \tag{7}$$

$$= \underset{\mathcal{C}_k \in \{\mathcal{C}_1, \ldots, \mathcal{C}_\kappa\}}{\arg\max} \ \widehat{p}(\mathcal{C}_k) \prod_{j=1}^{m} \widehat{p}(x_j | \mathcal{C}_k) \tag{8}$$

$\widehat{p}$ denotes an **approximated** probability

How to get the most probable Class? (Ctd.)



Example Data Set

Outlook	Temperature	Humidity	Wind	PlayGolf
sunny	hot	high	weak	**no**
sunny	hot	high	strong	**no**
overcast	hot	high	weak	**yes**
rainy	mild	high	weak	**yes**
rainy	cool	normal	weak	**yes**
rainy	cool	normal	strong	**no**
overcast	cool	normal	strong	**yes**
sunny	mild	high	weak	**no**
sunny	cool	normal	weak	**yes**
rainy	mild	normal	weak	**yes**
sunny	mild	normal	strong	**yes**
overcast	mild	high	strong	**yes**
overcast	hot	normal	weak	**yes**
rainy	mild	high	strong	**no**
sunny	cool	high	strong	???

How to estimate the Probabilities?

- How to estimate the probabilities $\widehat{p}(\mathcal{C}_k)$ and $\widehat{p}(x_j|\mathcal{C}_k)$?
- **Solution**: Simply count the occurrences

$$\widehat{p}(\mathcal{C}_k) = \frac{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)} = \mathcal{C}_k\}}{n} \tag{9}$$

$$\widehat{p}(x_j = v|\mathcal{C}_k) = \frac{\sum_{i=1}^{n} \mathbb{1}\{x_j^{(i)} = v \wedge y^{(i)} = \mathcal{C}_k\}}{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)} = \mathcal{C}_k\}} \tag{10}$$

- $\mathbb{1}\{bool\}$ is the indicator function
  (returns 1, if *bool* is true, 0 otherwise. E. g.: $\mathbb{1}\{1+1=2\}=1$, $\mathbb{1}\{3=2\}=0$)

Let's compute some Probabilities

- New instance $\boldsymbol{x} = \langle sunny, cool, high, strong \rangle$
- What is its class?
- Let's compute some of the probabilities needed:

$$
\begin{aligned}
\widehat{p}(Golf = yes) &= 9/14 = 0.64 \\
\widehat{p}(Golf = no) &= 5/14 = 0.36 \\
\widehat{p}(Outlook = sunny | Golf = yes) &= 2/9 = 0.22 \\
\widehat{p}(Outlook = sunny | Golf = no) &= 3/5 = 0.60
\end{aligned}
$$

$$\ldots$$

Class Prediction

$$
\begin{aligned}
\widehat{p}(yes|\mathbf{x}) &= \overbrace{\widehat{p}(sunny|yes)}^{=0.22} \cdot \overbrace{\widehat{p}(cool|yes) \cdot \widehat{p}(high|yes) \cdot \widehat{p}(strong|yes)}^{\text{calculate probabilities accordingly}} \cdot \overbrace{\widehat{p}(yes)}^{=0.64} \\
&= 0.0053
\end{aligned}
$$

$$
\begin{aligned}
\widehat{p}(no|\mathbf{x}) &= \underbrace{\widehat{p}(sunny|no)}_{=0.60} \cdot \underbrace{\widehat{p}(cool|no) \cdot \widehat{p}(high|no) \cdot \widehat{p}(strong|no)}_{\text{calculate probabilities accordingly}} \cdot \underbrace{\widehat{p}(no)}_{=0.36} \\
&= 0.0206
\end{aligned}
$$

Classification: $\mathcal{C}_{MAP} = no$ (no golf today...)

Scaling the Output

- **But wait!** These probabilities don't sum up to one!?!?
  - This is because we dropped the normalization term $p(\boldsymbol{x})$
  - Scaling can fix this:

$$\widehat{p}(yes|\boldsymbol{x})_{norm} = \frac{0.0053}{0.0053 + 0.0206} = 0.205$$

$$\widehat{p}(no|\boldsymbol{x})_{norm} = \frac{0.0206}{0.0053 + 0.0206} = 0.795$$

- Scaling does **not** change the prediction

Laplace Smoothing

- **Problem:** A feature value $v^\star$ in the test data not seen during training
- $\widehat{p}(v^\star|\mathcal{C}_k) = 0$: The whole product becomes zero...
- **Solution**: Laplace smoothing

$$\widehat{p}(\mathcal{C}_k) = \frac{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)} = \mathcal{C}_k\} + 1}{n + \kappa} \tag{11}$$

$$\widehat{p}(x_j = v|\mathcal{C}_k) = \frac{\sum_{i=1}^{n} \mathbb{1}\{x_j^{(i)} = v \wedge y^{(i)} = \mathcal{C}_k\} + 1}{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)} = \mathcal{C}_k\} + \kappa} \tag{12}$$

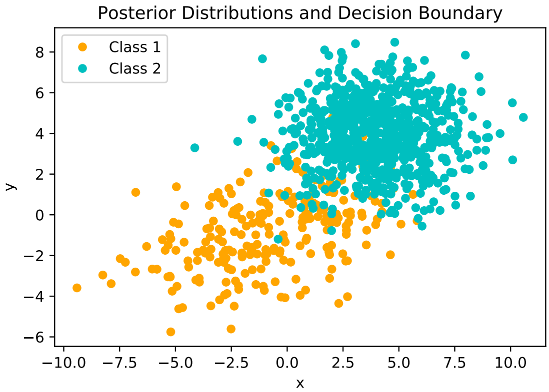
Section:
Gaussian Naïve Bayes

Handling of continuous Data
Maximum Likelihood Estimation (MLE)
Generative vs. Discriminative Models

Handling of continuous Data

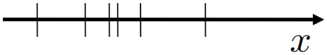
- We have learned about Bayes' optimal classifiers which classify data based on the probability distribution $p(\boldsymbol{x}|\mathcal{C}_k) \cdot p(\mathcal{C}_k)$
- Multinomial naïve Bayes can only be used for **discrete data**
- **How to get these probabilities in the continuous case?**
    - The prior $p(\mathcal{C}_k)$ is still easy to compute
    - The estimation of class conditional probabilities $p(\boldsymbol{x}|\mathcal{C}_k)$ is more complicated
    - Assume labeled data; estimate the density separately for each class $\mathcal{C}_k$
- NB: For ease of notation: $p(x) \equiv p(x|\mathcal{C}_k)$

Training Data Example

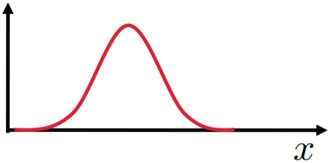


General Approach

- Given some (continuous) training data $\boldsymbol{X} = \{x^{(i)}\}_{i=1}^{n}$ (where all $x^{(i)}$ belong to the same class):



- Estimate $p(x)$ using a fixed parametric form:



Example: Gaussian Distribution

- One common case is the **Gaussian distribution**:

$$p(x|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{13}$$

- Notation for parametric models:
  - $p(x|\boldsymbol{\theta})$
  - In the case of a Gaussian: $\boldsymbol{\theta} = \{\mu, \sigma^2\}$, where $\mu \equiv$ mean, and $\sigma^2 \equiv$ variance

Learning the Parameters

- Learning means estimating the parameters $\boldsymbol{\theta}$ given the data $\boldsymbol{X}$
- Likelihood of the parameters $\boldsymbol{\theta}$:
  - Is defined as the probability that $\boldsymbol{X}$ was generated by a probability density function (pdf) with parameters $\boldsymbol{\theta}$

$$\mathcal{L}(\boldsymbol{\theta}) = p(\boldsymbol{X}|\boldsymbol{\theta}) \tag{14}$$

  - We want to **maximize** the likelihood

$\Rightarrow$ Maximum likelihood estimation (MLE)

A fundamental Assumption

- How to compute $\mathcal{L}(\boldsymbol{\theta})$?
- The data is assumed to be i. i. d. (independent and identically distributed):
  - Two random variables x_1 and x_2 are independent, if

$$P(x_1 \leqslant \alpha, x_2 \leqslant \beta) = P(x_1 \leqslant \alpha) \cdot P(x_2 \leqslant \beta) \qquad \forall \alpha, \beta \in \mathbb{R} \tag{15}$$

  - Two random variables x_1 and x_2 are identically distributed, if

$$P(x_1 \leqslant \alpha) = P(x_2 \leqslant \alpha) \qquad \forall \alpha \in \mathbb{R} \tag{16}$$

Important

Computation of the Likelihood

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

$$= p(x^{(1)}, x^{(2)}, \dots, x^{(n)}|\boldsymbol{\theta})$$

data is independent:

$$= p(x^{(1)}|\boldsymbol{\theta}) \cdot p(x^{(2)}|\boldsymbol{\theta}) \cdot \ldots \cdot p(x^{(n)}|\boldsymbol{\theta})$$

data is identically distributed:

$$= \prod_{i=1}^{n} p(x^{(i)}|\boldsymbol{\theta}) \tag{17}$$

What is the problem here?

Computation of the Likelihood (Ctd.)

- **Problem:** Large n might cause arithmetic underflows! **(why?)**
- Transform the likelihood using the logarithm $\Rightarrow$ log-likelihood

Why is this an allowed transformation?

$$\begin{aligned} \mathcal{LL}(\boldsymbol{\theta}) &= \log \mathcal{L}(\boldsymbol{\theta}) \\ &= \log \prod_{i=1}^{n} p(x^{(i)}|\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \log p(x^{(i)}|\boldsymbol{\theta}) \end{aligned} \tag{18}$$

$\log \Pi = \Sigma \log$

Maximum Likelihood of a Gaussian

- $\boldsymbol{\theta} = \{\mu, \sigma^2\}$

$$\mathcal{LL}(\{\mu, \sigma^2\}) = \sum_{i=1}^{n} \log \mathcal{N}(x^{(i)} | \mu, \sigma^2) \tag{19}$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x^{(i)} - \mu)^2}{2\sigma^2}\right\} \tag{20}$$

- Find μ_{ml} and σ^2_{ml} which maximize the log-likelihood:

$$\mu_{ml}, \sigma^2_{ml} = \underset{\mu, \sigma^2}{\arg\max}\, \mathcal{LL}(\boldsymbol{\theta})$$

Maximum Likelihood of a Gaussian (Ctd.)

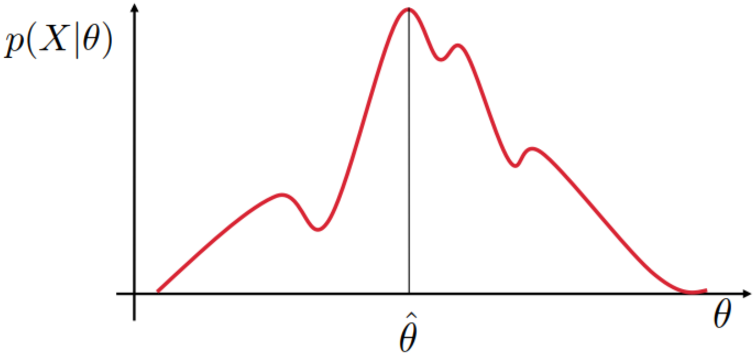
- Compute the partial derivatives with respect to the parameters $\boldsymbol{\theta}$
- Derivative w. r. t. μ:

$$\nabla_{\mu} \mathcal{L}\mathcal{L}(\boldsymbol{\theta}) = \nabla_{\mu} \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x^{(i)}-\mu)^2}{2\sigma^2}\right\} = \sum_{i=1}^{n} \frac{x^{(i)}-\mu}{\sigma^2}$$

- Set derivative to zero and solve:

$$\sum_{i=1}^{n} (x^{(i)}-\mu) \stackrel{!}{=} 0 \Leftrightarrow n \cdot \mu = \sum_{i=1}^{n} x^{(i)} \Leftrightarrow \mu = \frac{1}{n} \sum_{i=1}^{n} x^{(i)}$$

Maximization of the Likelihood



Important

We can classify!

Looks familiar?

- Maximum likelihood parameters:

$$\mu_{ml} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)} \qquad \sigma_{ml}^2 = \frac{1}{n}\sum_{i=1}^{n} (x^{(i)} - \mu_{ml})^2$$

- Now we can use Bayes' rule to predict class labels
  - We have the priors...
  - ...and the class conditionals
- Also, the **decision boundary** can be computed

Multivariate Case

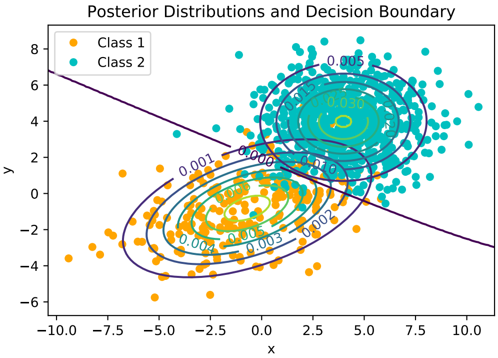
- The solution above is for 1-D data; what if we have more dimensions?
- **Multivariate Gaussian distribution**:

$$\mathcal{N}_D(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{21}$$

- Luckily, the derivations don't change:

$$\boldsymbol{\mu}_{ml} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}^{(i)} \qquad \boldsymbol{\Sigma}_{ml} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{ml})(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{ml})^\intercal \tag{22}$$

Gaussian naïve Bayes – Final Model



$$\boxed{p(\mathcal{C}_k|\boldsymbol{x}) = \mathcal{N}_D(\boldsymbol{x}|\boldsymbol{\mu}_{\mathcal{C}_k}, \boldsymbol{\Sigma}_{\mathcal{C}_k}) \cdot p(\mathcal{C}_k)}$$

NB: $\mathcal{N}_D(\boldsymbol{x}|\boldsymbol{\mu}_{\mathcal{C}_k}, \boldsymbol{\Sigma}_{\mathcal{C}_k})$ denotes the Gaussian distribution estimated for class $\mathcal{C}_k$ (using MLE). $p(\mathcal{C}_k)$ is the prior probability of class $\mathcal{C}_k$ (as in the discrete case).

Generative vs. Discriminative Models

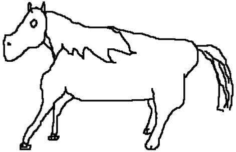
Generative Model

The artist

A **generative** algorithm models **how** the data was generated. **It models the respective probability distributions.**

Discriminative Model

The lousy painter



A **discriminative** algorithm does not care about how the data was generated. **It only knows how to distinguish the classes.**

Section:
Wrap-Up

Summary
Self-Test Questions
Lecture Outlook

Summary

- Important concepts: **Class conditional probabilities** and **class priors**
- Use **Bayes' theorem** to get the **class posteriors**
- **Bayes' optimal classifier:** Decide for the most probable class
- Naïve Bayes assumes all features to be **pairwise conditionally independent**
- We can use **parametric models** to estimate the density of the data. They assume a certain **parametric form**, e. g. a Gaussian distribution
- This allows us to work with **continuous features**

Important

Self-Test Questions

1. What are class conditional probabilities?
2. What does *Bayes' optimal* mean?
3. How can we incorporate prior knowledge about the class distribution into the classification?
4. What is the naïve assumption which naïve Bayes makes? When might this be a problem?
5. Explain what maximum aposteriori is!
6. What is maximum likelihood estimation? How can you get the maximum likelihood estimate for a Gaussian distribution?

What's next...?

Unit I	Machine Learning Introduction
Unit II	Mathematical Foundations
Unit III	Bayesian Decision Theory
Unit IV	**Regression**
Unit V	Classification I
Unit VI	Evaluation
Unit VII	Classification II
Unit VIII	Clustering
Unit IX	Dimensionality Reduction

Thank you very much for the attention!

Topic: *** Applied Machine Learning Fundamentals *** Bayesian Decision Theory
Term: Winter term 2023/2024

Contact:
Daniel Wehner, M.Sc.
SAP SE / DHBW Mannheim
daniel.wehner@sap.com

Do you have any questions?