TECHNISCHE UNIVERSITÄT DARMSTADT

Statistical Machine Learning

Lecture 05: Bayesian Decision Theory

Kristian Kersting

TU Darmstadt

Summer Term 2020

Today's Objectives

- Make you understand how to do an optimal decision!
- Covered Topics:
  - Bayesian Optimal Decisions
  - Classification from a Bayesian point of view
  - Risk-based Classification

Outline

1. Bayesian Decision Theory

2. Risk Minimization

3. Wrap-Up

TECHNISCHE UNIVERSITÄT DARMSTADT

Outline

1. Bayesian Decision Theory

2. Risk Minimization

3. Wrap-Up

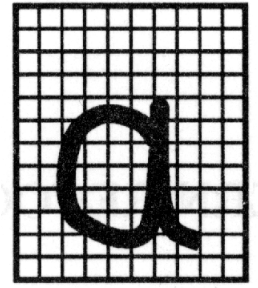
TECHNISCHE UNIVERSITÄT DARMSTADT

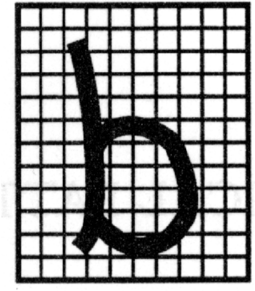
Statistical Methods

- Statistical methods in machine learning all have in common that they 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
- Today will be all about probabilities. But in future lectures, the use of probability will sometimes be much less explicit
- Nonetheless, the basic assumption about how the data is generated is always there, even if you don’t see a single probability distribution anywhere

TECHNISCHE UNIVERSITÄT DARMSTADT

Character Recognition

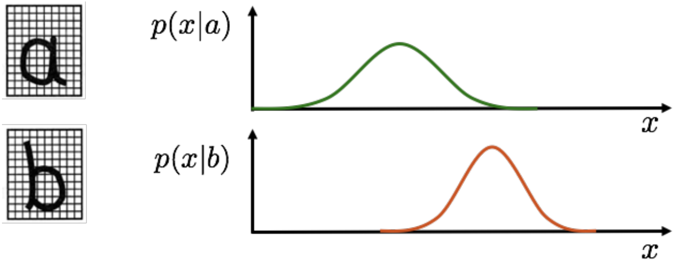




- **Goal**: classify a new letter so that the probability of a wrong classification is minimized

Class conditional probabilities

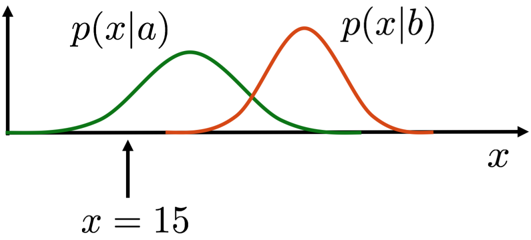
- Class conditional probabilities
  - Probability of making an observation **x** knowing that it comes from some class C_k
  - Here **x** is often a feature vector, which measures/describes properties of the data. E.g.: number of black pixels, height-width ratio, ...



TECHNISCHE UNIVERSITÄT DARMSTADT

Class conditional probabilities

- Example

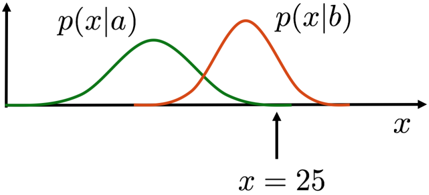


- How do we decide which class the data point belongs to?
- Here, we should decide for class **a**

TECHNISCHE UNIVERSITÄT DARMSTADT

Class conditional probabilities

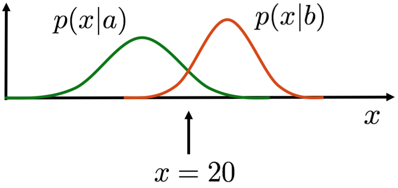
- Example



- How do we decide which class the data point belongs to?
- Since $p(\mathbf{x}|a)$ is a lot smaller than $p(\mathbf{x}|b)$ we should now decide for class **b**

Class conditional probabilities

- Example



- How do we decide which class the data point belongs to?

TECHNISCHE UNIVERSITÄT DARMSTADT

Class priors

- The *a priori* probability of a data point belonging to a particular class is called the class prior
- Example:
  - abaaababaaaabbaaaaaa
- What are $p(a)$ and $p(b)$?

$$C_1 = a \quad p(C_1) = 0.75$$

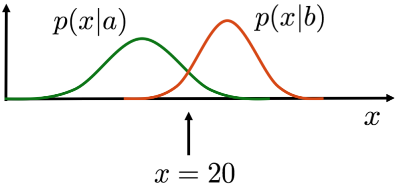
$$C_2 = b \quad p(C_2) = 0.25$$

$$\sum_k p(C_k) = 1$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Back to our problem...

- Example



- How do we decide which class the data point belongs to?
- If $p(a) = 0.75$ and $p(b) = 0.25$, we should decide for class **a**

TECHNISCHE UNIVERSITÄT DARMSTADT

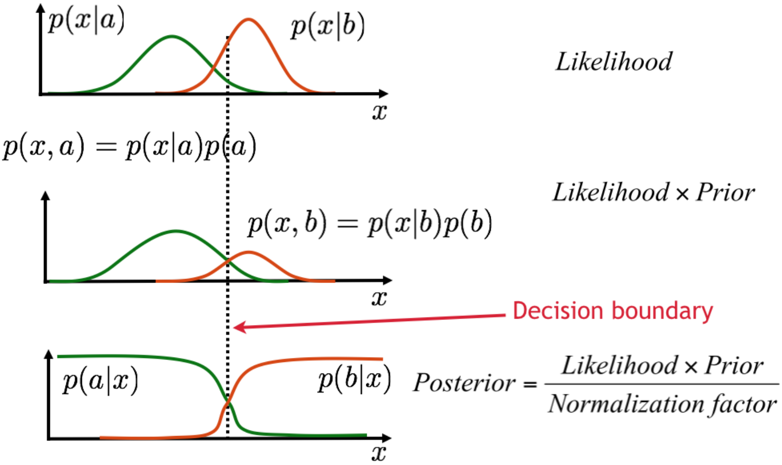
Bayesian Decision Theory

- Bayes Theorem lets us formalize the previous intuitive decision
- We want to find the a-posteriori probability (posterior) of the class C_k given the observation (feature) $\mathbf{x}$

$$p\left(C_k|\mathbf{x}\right) = \frac{p\left(\mathbf{x}|C_k\right)p\left(C_k\right)}{p\left(\mathbf{x}\right)} = \frac{p\left(\mathbf{x}|C_k\right)p\left(C_k\right)}{\sum_j p\left(\mathbf{x}|C_j\right)p\left(C_j\right)}$$

  - class prior: $p\left(C_k\right)$
  - class-conditional probability (likelihood): $p\left(\mathbf{x}|C_k\right)$
  - class posterior: $p\left(C_k|\mathbf{x}\right)$
  - normalization term: $p\left(\mathbf{x}\right)$

Bayesian Decision Theory



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

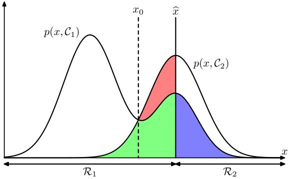
Bayesian Probabilities

- 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 misclassification rate (the probability of classifying wrongly)



$$
\begin{aligned}
p\left(\text{error}\right) &= p\left(x \in R_1, C_2\right) + p\left(x \in R_2, C_1\right) \\
&= \int_{R_1} p\left(x, C_2\right) \mathrm{d}x + \int_{R_2} p\left(x, C_1\right) \mathrm{d}x \\
&= \int_{R_1} p\left(x \middle| C_2\right) p\left(C_2\right) \mathrm{d}x + \int_{R_2} p\left(x \middle| C_1\right) p\left(C_1\right) \mathrm{d}x
\end{aligned}
$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Bayesian Decision Theory

- Decision rule: decide C_1 if $p(C_1|x) > p(C_2|x)$
- Equivalent to

$$\begin{aligned}
\frac{p(x|C_1)\,p(C_1)}{p(x)} &> \frac{p(x|C_2)\,p(C_2)}{p(x)} \\
p(x|C_1)\,p(C_1) &> p(x|C_2)\,p(C_2) \\
\frac{p(x|C_1)}{p(x|C_2)} &> \frac{p(C_2)}{p(C_1)}
\end{aligned}$$

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

TECHNISCHE UNIVERSITÄT DARMSTADT

Bayesian Decision Theory

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

- Special cases
  - If $p\left(x|C_1\right) = p\left(x|C_2\right)$, then use $p\left(C_1\right) > p\left(C_2\right)$
  - If $p\left(C_1\right) = p\left(C_2\right)$, then use $p\left(x|C_1\right) > p\left(x|C_2\right)$

More than two Classes

- Generalization to more than 2 classes:
    - Decide for class k iff it has the highest a-posteriori probability

$$p\left(C_k|x\right) > p\left(C_j|x\right) \quad \forall j \neq k$$

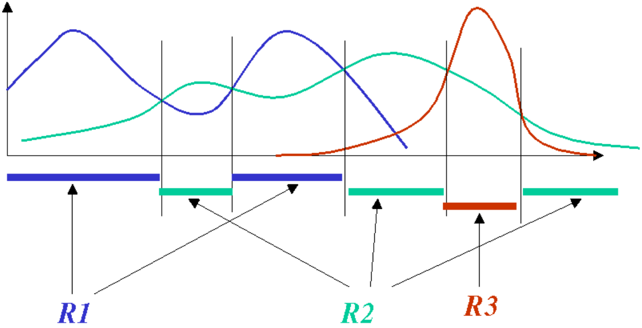
    - Equivalent to

$$\begin{aligned} p\left(x|C_k\right)p\left(C_k\right) &> p\left(x|C_j\right)p\left(C_j\right) \quad \forall j \neq k \\ \frac{p\left(x|C_k\right)}{p\left(x|C_j\right)} &> \frac{p\left(C_j\right)}{p\left(C_k\right)} \quad \forall j \neq k \end{aligned}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

More than two Classes

- Decision regions: $R_1, R_2, R_3, \ldots$



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 instance, 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 $p\left(\boldsymbol{x}|C_k\right)$

Dummy Classes

- There are also applications, where it may be advantageous to have a dummy class denoted “don’t know” or “don’t care”
  - Also called a reject option
- Not a common case though and we will not cover this in this class

Outline

1. Bayesian Decision Theory

2. Risk Minimization

3. Wrap-Up

TECHNISCHE UNIVERSITÄT DARMSTADT

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

Loss Functions

- Key idea: we have to construct a loss function in a way that expresses what we want to achieve

$$\text{loss}\,(\text{decision} = \text{healthy}|\text{patient} = \text{sick}) >>$$
$$\text{loss}\,(\text{decision} = \text{sick}|\text{patient} = \text{healthy})$$

- Possible decisions: α_i
- True classes: C_j
- Loss function: $\lambda\left(\alpha_i|C_j\right)$
- Expected loss of making a decision α_i

$$R\left(\alpha_i|x\right) = \mathbb{E}_{C_k \sim p(C_k|x)}\left[\lambda\left(\alpha_i|C_k\right)\right] = \sum_j \lambda\left(\alpha_i|C_j\right) p\left(C_j|x\right)$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Risk Minimization

- The expected loss of a decision is also called the risk of making a decision
- Instead of minimizing the Misclassification rate

$$\begin{aligned} p\left(\text{error}\right) &= p\left(x \in R_1, C_2\right) + p\left(x \in R_2, C_1\right) \\ &= \int_{R_1} p\left(x, C_2\right) \mathrm{d}x + \int_{R_2} p\left(x, C_1\right) \mathrm{d}x \\ &= \int_{R_1} p\left(x|C_2\right) p\left(C_2\right) \mathrm{d}x + \int_{R_2} p\left(x|C_1\right) p\left(C_1\right) \mathrm{d}x \end{aligned}$$

- We minimize the Overall Risk

$$R\left(\alpha_i|x\right) = \mathbb{E}_{C_k \sim p(C_k|x)}\left[\lambda\left(\alpha_i|C_k\right)\right] = \sum_j \lambda\left(\alpha_i|C_j\right) p\left(C_j|x\right)$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Risk Minimization

- 2 classes: C_1, C_2
- 2 decisions: α_1, α_2
- Loss function: $\lambda\left(\alpha_i|C_j\right) = \lambda_{ij}$
- Risk of both decisions

$$
\begin{aligned}
R\left(\alpha_1|x\right) &= \lambda_{11} p\left(C_1|x\right) + \lambda_{12} p\left(C_2|x\right) \\
R\left(\alpha_2|x\right) &= \lambda_{21} p\left(C_1|x\right) + \lambda_{22} p\left(C_2|x\right)
\end{aligned}
$$

- Goal: Create a decision rule so that overall risk is minimized
  - Decide α_1 if $R\left(\alpha_2|x\right) > R\left(\alpha_1|x\right)$

TECHNISCHE UNIVERSITÄT DARMSTADT

Risk Minimization

$$
\begin{aligned}
R(\alpha_2|x) &> R(\alpha_1|x) \\
\lambda_{21} p(C_1|x) + \lambda_{22} p(C_2|x) &> \lambda_{11} p(C_1|x) + \lambda_{12} p(C_2|x) \\
(\lambda_{21} - \lambda_{11}) p(C_1|x) &> (\lambda_{12} - \lambda_{22}) p(C_2|x)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\lambda_{21} - \lambda_{11}}{\lambda_{12} - \lambda_{22}} &> \frac{p(C_2|x)}{p(C_1|x)} = \frac{p(x|C_2)\,p(C_2)}{p(x|C_1)\,p(C_1)} \\
\frac{p(x|C_1)}{p(x|C_2)} &> \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{p(C_2)}{p(C_1)}
\end{aligned}
$$

- 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 0-1 Loss

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

- Decide α_1 if

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

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

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

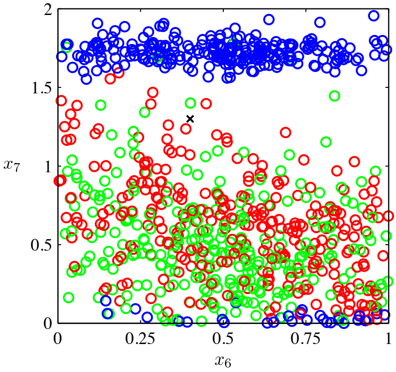
TECHNISCHE UNIVERSITÄT DARMSTADT

Bayesian Decision Theory

- Are we done with classification?
  - We have decision rules for simple and general loss functions
  - Even “Bayes optimal”
  - We can deal with 2 or more classes
  - We can deal with high dimensional feature vectors
  - We can incorporate prior knowledge on the class distribution
- What are we going to do the rest of the semester? Where is the catch?
- Where do we get the probability distributions from?

TECHNISCHE UNIVERSITÄT DARMSTADT

Training Data



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

TECHNISCHE UNIVERSITÄT DARMSTADT

Outline

1. Bayesian Decision Theory

2. Risk Minimization

3. Wrap-Up

3. Wrap-Up

You know now:

- What class conditional probabilities, class priors and class posteriors are
- What Bayesian Decision Theory is
- How to use Bayes Theorem for classification
- What misclassification rate is
- What a Bayes Optimal Classifier is
- How to generalize decision to more than 2 classes
- What risk is, and how it relates to misclassification

Self-Test Questions

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

Homework

- Reading Assignment for next lecture
  - Bishop ch. 2 (Probability Distributions), 9 (Mixture Models and EM)