These are the slides of the lecture

Pattern Recognition
Winter term 2020/21
Friedrich-Alexander University of Erlangen-Nuremberg.

Erlangen, January 8, 2021
Prof. Dr.-Ing. Andreas Maier

Pattern Recognition (PR)

Prof. Dr.-Ing. Andreas Maier
Pattern Recognition Lab (CS 5), Friedrich-Alexander-Universität Erlangen-Nürnberg
Winter Term 2020/21

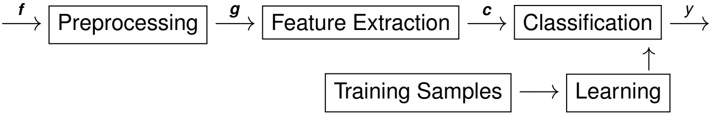
Pattern Recognition Basics

Classification of Simple Patterns

The system for the classification of simple patterns has the following generic structure

Classification of Simple Patterns

The system for the classification of simple patterns has the following generic structure



Classification of Simple Patterns (cont.)

- *Supervised learning:*
  m training samples include feature and associated class number

$$S = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), (\boldsymbol{x}_3, y_3), \ldots, (\boldsymbol{x}_m, y_m)\}$$

where $\boldsymbol{x}_i \in \mathscr{X}$ denotes the feature vector and $y_i \in Z$ denotes the class number of sample i. If nothing special is mentioned $\mathscr{X} \subseteq \mathbb{R}^d$.

Classification of Simple Patterns (cont.)

- *Supervised learning:*
  m training samples include feature and associated class number

  $$S = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), (\boldsymbol{x}_3, y_3), \ldots, (\boldsymbol{x}_m, y_m)\}$$

  where $\boldsymbol{x}_i \in \mathscr{X}$ denotes the feature vector and $y_i \in Z$ denotes the class number of sample i. If nothing special is mentioned $\mathscr{X} \subseteq \mathbb{R}^d$.

- *Unsupervised learning:*
  m training samples just include features, no class assignments and even the number of classes is (not always) known

  $$S = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_m\}$$

Bayesian Classifier

Notation:

$\boldsymbol{x} \in \mathbb{R}^d$: d-dimensional feature vector

y : class number
(usually $y \in \{0, 1\}$ or $y \in \{-1, +1\}$)

$p(y)$: prior probability of pattern class y

$p(\boldsymbol{x})$: evidence
(distribution of features in d-dimensional feature space)

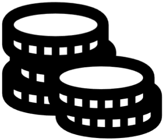
$p(\boldsymbol{x}, y)$: joint probability density function (pdf)

$p(\boldsymbol{x}|y)$: class conditional density

$p(y|\boldsymbol{x})$: posterior probability

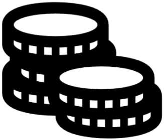
Bayesian Classifier (cont.)

	Heads	Tails
Evidence	18	33



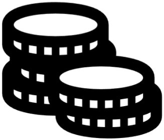
Bayesian Classifier (cont.)

	Heads	Tails	Sum
Evidence	18	33	51



Bayesian Classifier (cont.)

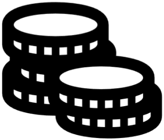
	Heads	Tails	Sum
Evidence	18	33	51



$$p(x = \text{“Heads“}) = \frac{18}{51} \approx 0{,}35$$

Bayesian Classifier (cont.)

	Heads	Tails	Sum
Evidence	18	33	51

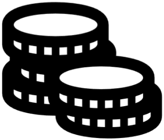


$$p(x = \text{“Heads“}) = \frac{18}{51} \approx 0{,}35$$

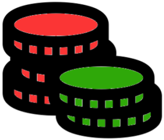
$$p(x = \text{“Tails“}) = \frac{35}{51} \approx 0{,}65$$

Bayesian Classifier (cont.)

	Heads	Tails	Sum
Evidence	18	33	51

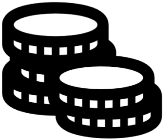


	Heads	Tails
Red coin	17	15
Green coin	1	18

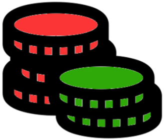


Bayesian Classifier (cont.)

	Heads	Tails	Sum
Evidence	18	33	51

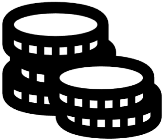


	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19

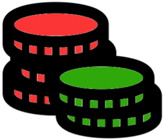


Bayesian Classifier (cont.)

	Heads	Tails	Sum
Evidence	18	33	51

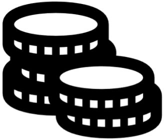


	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51



Bayesian Classifier (cont.)

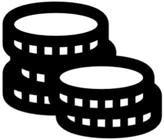
	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51



Bayesian Classifier (cont.)

$$p(x = \text{“Heads“}) = \frac{18}{51} \approx 0{,}35$$

	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51

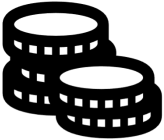


Bayesian Classifier (cont.)

$$p(x = \text{“Heads“}) = \frac{18}{51} \approx 0{,}35$$

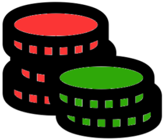
$$p(x = \text{“Tails“}) = \frac{35}{51} \approx 0{,}65$$

	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51



Bayesian Classifier (cont.)

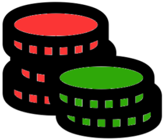
	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51



Bayesian Classifier (cont.)

$$p(y = \text{“Red coin“}) = \frac{32}{51} \approx 0{,}63$$

	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51

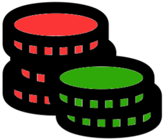


Bayesian Classifier (cont.)

$$p(y = \text{“Red coin“}) = \frac{32}{51} \approx 0{,}63$$

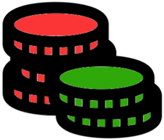
$$p(y = \text{“Green coin“}) = \frac{19}{51} \approx 0{,}37$$

	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51



Bayesian Classifier (cont.)

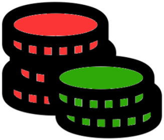
	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51



Bayesian Classifier (cont.)

$$p(x = \text{“Heads“}, y = \text{"Red coin"}) = \frac{17}{51} \approx 0{,}33$$

	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51

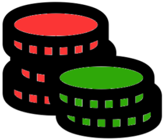


Bayesian Classifier (cont.)

$$p(x = \text{“Heads“}, y = \text{"Red coin"}) = \frac{17}{51} \approx 0{,}33$$

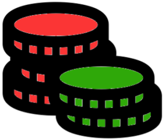
$$p(x = \text{“Tails“}, y = \text{"Green coin"}) = \frac{18}{51} \approx 0{,}35$$

	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51



Bayesian Classifier (cont.)

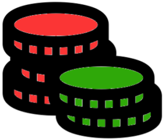
	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51



Bayesian Classifier (cont.)

$$p(x = \text{“Heads“} \mid y = \text{"Green coin"}) = \frac{1}{19} \approx 0{,}05$$

	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51

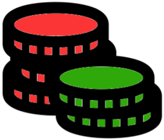


Bayesian Classifier (cont.)

$$p(x = \text{“Heads“} \mid y = \text{"Green coin"}) = \frac{1}{19} \approx 0{,}05$$

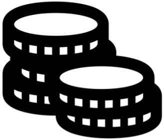
$$p(x = \text{“Tails“} \mid y = \text{"Green coin"}) = \frac{18}{19} \approx 0{,}95$$

	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51



Bayesian Classifier (cont.)

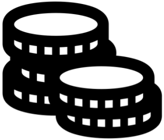
	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51



Bayesian Classifier (cont.)

$$p(y = \text{“Red coin“} \mid x = \text{“Tails“}) = \frac{15}{33} \approx 0{,}45$$

	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51

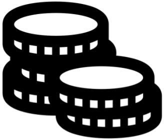


Bayesian Classifier (cont.)

$$p(y = \text{“Red coin“} \mid x = \text{“Tails“}) = \frac{15}{33} \approx 0{,}45$$

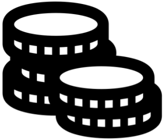
$$p(y = \text{“Green coin“} \mid x = \text{“Tails“}) = \frac{18}{33} \approx 0{,}55$$

	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51



Bayesian Classifier (cont.)

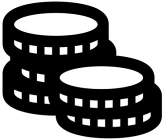
	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51



Bayesian Classifier (cont.)

$$p(y = \text{“Red coin“} \mid x = \text{"Heads"}) = \frac{17}{18} \approx 0{,}94$$

	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51

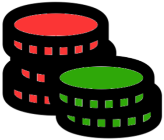


Bayesian Classifier (cont.)

$$p(y = \text{“Red coin“} \mid x = \text{"Heads"}) = \frac{17}{18} \approx 0{,}94$$

$$p(y = \text{“Green coin “} \mid x = \text{"Heads"}) = \frac{1}{18} \approx 0{,}06$$

	Heads	Tails	Sum
Red coin	17	15	32
Green coin	1	18	19
Sum	18	33	51



Bayesian Classifier (cont.)

Bayes rule:

$$\underbrace{p(\boldsymbol{x}, y)}_{\textit{joint pdf}} \quad =$$

Bayesian Classifier (cont.)

Bayes rule:

$$\underbrace{p(\boldsymbol{x}, y)}_{joint\ pdf} = \underbrace{p(y)}_{prior} \cdot \underbrace{p(\boldsymbol{x}|y)}_{class\ conditional\ pdf}$$

Bayesian Classifier (cont.)

Bayes rule:

$$\begin{aligned}\underbrace{p(\boldsymbol{x}, y)}_{joint\ pdf} &= \underbrace{p(y)}_{prior} \cdot \underbrace{p(\boldsymbol{x}|y)}_{class\ conditional\ pdf} \\ &= \underbrace{p(\boldsymbol{x})}_{evidence} \cdot \underbrace{p(y|\boldsymbol{x})}_{posterior}\end{aligned}$$

Bayesian Classifier (cont.)

Now we get the posterior as follows:

$$p(y|\boldsymbol{x}) \quad =$$

Bayesian Classifier (cont.)

Now we get the posterior as follows:

$$p(y|\boldsymbol{x}) = \frac{p(y) \cdot p(\boldsymbol{x}|y)}{p(\boldsymbol{x})}$$

Bayesian Classifier (cont.)

Now we get the posterior as follows:

$$\begin{aligned} p(y|\boldsymbol{x}) &= \frac{p(y) \cdot p(\boldsymbol{x}|y)}{p(\boldsymbol{x})} \\ &= \frac{p(y) \cdot p(\boldsymbol{x}|y)}{\sum_{y'} p(\boldsymbol{x}, y')} \end{aligned}$$

Bayesian Classifier (cont.)

Now we get the posterior as follows:

$$
\begin{aligned}
p(y|\boldsymbol{x}) &= \frac{p(y)\cdot p(\boldsymbol{x}|y)}{p(\boldsymbol{x})} \\
&= \frac{p(y)\cdot p(\boldsymbol{x}|y)}{\sum_{y'} p(\boldsymbol{x},y')} \\
&= \frac{p(y)\cdot p(\boldsymbol{x}|y)}{\sum_{y'} p(y')\cdot p(\boldsymbol{x}|y')}
\end{aligned}
$$

Bayesian Classifier (cont.)

Note:

$$p(\boldsymbol{x}) = \sum_y p(y) \cdot p(\boldsymbol{x}|y)$$

is a marginal of $p(\boldsymbol{x}, y)$.

- We get $p(\boldsymbol{x})$ by marginalizing $p(\boldsymbol{x}, y)$ over y.
- Accordingly we get $p(y)$ by marginalizing $p(\boldsymbol{x}, y)$ over $\boldsymbol{x}$, i. e.

$$p(y) = \int p(\boldsymbol{x}, y) \mathrm{d}\boldsymbol{x}$$

Did you notice: y is a discrete random variable whereas $\boldsymbol{x}$ is a continuous random vector (summation vs. integration).

Pattern Recognition Lab

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Next Time in

Pattern Recognition

Bayesian Classifier (cont.)

Now let us summarize the Bayesian decision rule:

We decide for the class y^* according to the decision rule

$$y^* \quad =$$

Bayesian Classifier (cont.)

Now let us summarize the Bayesian decision rule:

We decide for the class y^* according to the decision rule

$$y^* = \underset{y}{\operatorname{argmax}}\, p(y|\boldsymbol{x})$$

Bayesian Classifier (cont.)

Now let us summarize the Bayesian decision rule:

We decide for the class y^* according to the decision rule

$$
\begin{aligned}
y^* &= \operatorname*{argmax}_y p(y|\boldsymbol{x}) \\
&= \operatorname*{argmax}_y \frac{p(y) \cdot p(\boldsymbol{x}|y)}{p(\boldsymbol{x})}
\end{aligned}
$$

Bayesian Classifier (cont.)

Now let us summarize the Bayesian decision rule:

We decide for the class y^* according to the decision rule

$$
\begin{aligned}
y^* &= \operatorname*{argmax}_y p(y|\boldsymbol{x}) \\
&= \operatorname*{argmax}_y \frac{p(y) \cdot p(\boldsymbol{x}|y)}{p(\boldsymbol{x})} \\
&= \operatorname*{argmax}_y p(y) \cdot p(\boldsymbol{x}|y)
\end{aligned}
$$

Bayesian Classifier (cont.)

Now let us summarize the Bayesian decision rule:

We decide for the class y^* according to the decision rule

$$
\begin{aligned}
y^* &= \underset{y}{\operatorname{argmax}}\, p(y|\boldsymbol{x}) \\
&= \underset{y}{\operatorname{argmax}}\, \frac{p(y) \cdot p(\boldsymbol{x}|y)}{p(\boldsymbol{x})} \\
&= \underset{y}{\operatorname{argmax}}\, p(y) \cdot p(\boldsymbol{x}|y) \\
&= \underset{y}{\operatorname{argmax}} \{\log p(y) + \log p(\boldsymbol{x}|y)\}
\end{aligned}
$$

Bayesian Classifier (cont.)

Notes:

- The key aspect in designing a classifier is to find a good model for the posterior $p(y|\boldsymbol{x})$.
- Feature vectors $\boldsymbol{x}$ usually have fixed dimensions d in simple classification schemes,
- but $\mathscr{X}$ is not necessarily a subset of $\mathbb{R}^d$:
  features of varying dimension, sequences and sets of features

Bayesian Classifier (cont.)

- Generative modeling:
  modeling and estimation of $p(y)$ and $p(\boldsymbol{x}|y)$.

- Discriminative modeling:
  straight modeling and estimation of $p(y|\boldsymbol{x})$.

Optimality of the Bayesian Classifier

Definition

$l(y_1, y_2)$ is the loss if a feature vector belonging to class y_2 is assigned to class y_1. The $(0, 1)$-loss function is defined by

$$l(y_1, y_2) = \begin{cases} 0 & , \textit{if } y_1 = y_2 \\ 1 & , \textit{otherwise} \end{cases}$$

Optimality of the Bayesian Classifier (cont.)

The best (or optimal) decision rule according to classification loss minimizes the average loss L:

$$\mathrm{AL}(\boldsymbol{x}, y) = \sum_{y'} l(y, y') p(y' | \boldsymbol{x})$$

Optimality of the Bayesian Classifier (cont.)

Using the $(0,1)$-loss function, the class decision is based on:

$$y^* = \operatorname*{argmin}_y \mathrm{AL}(\boldsymbol{x}, y)$$

Optimality of the Bayesian Classifier (cont.)

Using the $(0, 1)$-loss function, the class decision is based on:

$$
\begin{aligned}
y^* &= \underset{y}{\operatorname{argmin}} \, \mathrm{AL}(\boldsymbol{x}, y) \\
&= \underset{y}{\operatorname{argmin}} \sum_{y'} l(y, y') \cdot p(y' | \boldsymbol{x})
\end{aligned}
$$

Optimality of the Bayesian Classifier (cont.)

Using the $(0,1)$-loss function, the class decision is based on:

$$
\begin{aligned}
y^* &= \operatorname*{argmin}_y \mathrm{AL}(\boldsymbol{x}, y) \\
&= \operatorname*{argmin}_y \sum_{y'} l(y, y') \cdot p(y' | \boldsymbol{x}) \\
&= \operatorname*{argmax}_y p(y | \boldsymbol{x})
\end{aligned}
$$

Optimality of the Bayesian Classifier (cont.)

Conclusion:

- The optimal classifier w. r. t. the (0,1)-loss function applies the Bayesian decision rule.
- This classifier is called Bayesian classifier.

⚠ The loss function is **NOT** convex.

Lessons Learned

- General structure of a classification system

Lessons Learned

- General structure of a classification system
- Supervised and unsupervised learning

Lessons Learned

- General structure of a classification system
- Supervised and unsupervised learning
- Basics on probabilities (probability, pdf, Bayes rule, etc.)

Lessons Learned

- General structure of a classification system
- Supervised and unsupervised learning
- Basics on probabilities (probability, pdf, Bayes rule, etc.)
- Optimality of Bayes classifier and the role of the loss function

Lessons Learned

- General structure of a classification system
- Supervised and unsupervised learning
- Basics on probabilities (probability, pdf, Bayes rule, etc.)
- Optimality of Bayes classifier and the role of the loss function
- Discriminative and generative approach to model a posteriori probability

Pattern Recognition Lab

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

TECHNISCHE FAKULTÄT

Next Time in

Pattern Recognition

Further Readings

- Heinrich Niemann:
  Pattern Analysis,
  Springer Series in Information Sciences 4, Springer, Berlin, 1982.
- Heinrich Niemann:
  Klassifikation von Mustern,
  Springer Verlag, Berlin, 1983.
- Richard O. Duda, Peter E. Hart, David G. Stork:
  Pattern Classification, 2nd Edition,
  John Wiley & Sons, New York, 2000.