These are the slides of the lecture

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

Please feel free to reuse any of the figures and slides, as long as you keep a reference to the source of these slides at `https://lme.tf.fau.de/teaching/` acknowledging the authors Niemann, Hornegger, Hahn, Steidl, Nöth, Seitz, Rodriguez, Das and Maier.

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

Naïve Bayes

Naïve Bayes and Statistical Independency

Naïve Bayes is

- still widely (and successfully) used
- often outperforming much more advanced classifiers
- appropriate in the presence of high dimensional features (curse of dimensionality)
- also called "Idiot's Bayes"

Naïve Bayes and Statistical Independency (cont.)

For the class dependent pdf we can do the following factorization:

$$p(\boldsymbol{x}|y) \quad = \quad p(x_1, x_2, \ldots, x_d|y)$$

Naïve Bayes and Statistical Independency (cont.)

For the class dependent pdf we can do the following factorization:

$$\begin{aligned} p(\boldsymbol{x}|y) &= p(x_1, x_2, \dots, x_d|y) \\ &= p(x_1|y)p(x_2, x_3, \dots, x_d|y, x_1) \end{aligned}$$

Naïve Bayes and Statistical Independency (cont.)

For the class dependent pdf we can do the following factorization:

$$
\begin{aligned}
p(\boldsymbol{x}|y) &= p(x_1, x_2, \dots, x_d|y) \\
&= p(x_1|y)p(x_2, x_3, \dots, x_d|y, x_1) \\
&= p(x_1|y)p(x_2|y, x_1)p(x_3, x_4, \dots, x_d|y, x_1, x_2)
\end{aligned}
$$

Naïve Bayes and Statistical Independency (cont.)

For the class dependent pdf we can do the following factorization:

$$
\begin{aligned}
p(\boldsymbol{x}|y) &= p(x_1, x_2, \ldots, x_d|y) \\
&= p(x_1|y)p(x_2, x_3, \ldots, x_d|y, x_1) \\
&= p(x_1|y)p(x_2|y, x_1)p(x_3, x_4, \ldots, x_d|y, x_1, x_2) \\
&= p(x_1|y)\prod_{i=2}^{d} p(x_i|y, x_1, \ldots, x_{i-1})
\end{aligned}
$$

Naïve Bayes and Statistical Independency (cont.)

- The Naïve Bayes classifier makes a very strong – so to call naïve – independency assumption.
- All d components of the feature vector $\boldsymbol{x}$ are assumed to be mutually independent.

Naïve Bayes and Statistical Independency (cont.)

- The Naïve Bayes classifier makes a very strong – so to call naïve – independency assumption.
- All d components of the feature vector $\boldsymbol{x}$ are assumed to be mutually independent.
- This independency assumption implies:

$$p(\boldsymbol{x}|y) = \prod_{i=1}^{d} p(x_i|y)$$

Naïve Bayes and Statistical Independency (cont.)

The decision rule of naïve Bayes reads as follows:

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

Naïve Bayes and Statistical Independency (cont.)

The decision rule of naïve Bayes reads as follows:

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

Naïve Bayes and Statistical Independency (cont.)

The decision rule of naïve Bayes reads as follows:

$$
\begin{aligned}
y^* &= \operatorname*{argmax}_y p(y|\boldsymbol{x}) \\
&= \operatorname*{argmax}_y p(y)p(\boldsymbol{x}|y) \\
&= \operatorname*{argmax}_y p(y)\prod_{i=1}^{d} p(x_i|y)
\end{aligned}
$$

An Example: Naïve Bayes and Gaussians

Example

Assume the 100–dimensional feature vector $\boldsymbol{x} \in \mathbb{R}^{100}$ belonging to class y is normally distributed and all components are *mutually dependent*:

$$\begin{aligned} \boldsymbol{\mu}_y &\in \mathbb{R}^{100} \\ \boldsymbol{\Sigma} &= \boldsymbol{\Sigma}^T \in \mathbb{R}^{100\times 100} \end{aligned}$$

The total number of parameters to be estimated for each class is

An Example: Naïve Bayes and Gaussians

Example

Assume the 100–dimensional feature vector $\boldsymbol{x} \in \mathbb{R}^{100}$ belonging to class y is normally distributed and all components are *mutually dependent*:

$$\begin{aligned} \boldsymbol{\mu}_y &\in \mathbb{R}^{100} \\ \boldsymbol{\Sigma} &= \boldsymbol{\Sigma}^T \in \mathbb{R}^{100\times 100} \end{aligned}$$

The total number of parameters to be estimated for each class is

$$100 + 100 \cdot (100+1)/2 = 5150.$$

An Example: Naïve Bayes and Gaussians (cont.)

Example cont.

Assume the 100–dimensional feature vector $\boldsymbol{x} \in \mathbb{R}^{100}$ belonging to class y is normally distributed and all components are *mutually independent*.

$$p(\boldsymbol{x}|y) = \prod_{i=1}^{100} p(x_i|y) \quad = \quad \prod_{i=1}^{100} \mathcal{N}(x_i; \mu_i, \sigma_i^2).$$

For each component $i = \{1, 2, 3, \ldots, 100\}$ we have to estimate mean $\mu_i \in \mathbb{R}$ and variance $\sigma_i^2 \in \mathbb{R}$. The total number of parameters to be estimated for each class is

An Example: Naïve Bayes and Gaussians (cont.)

Example cont.

Assume the 100–dimensional feature vector $\boldsymbol{x} \in \mathbb{R}^{100}$ belonging to class y is normally distributed and all components are *mutually independent*.

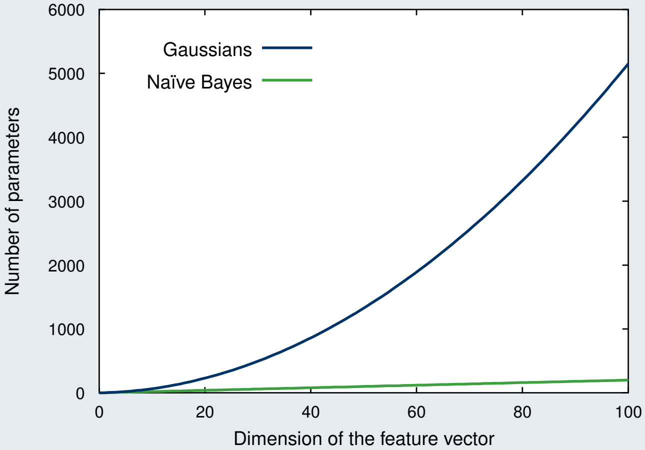
$$p(\boldsymbol{x}|y) = \prod_{i=1}^{100} p(x_i|y) \quad = \quad \prod_{i=1}^{100} \mathcal{N}(x_i; \mu_i, \sigma_i^2).$$

For each component $i = \{1, 2, 3, \ldots, 100\}$ we have to estimate mean $\mu_i \in \mathbb{R}$ and variance $\sigma_i^2 \in \mathbb{R}$. The total number of parameters to be estimated for each class is

$$100 + 100 = 200.$$

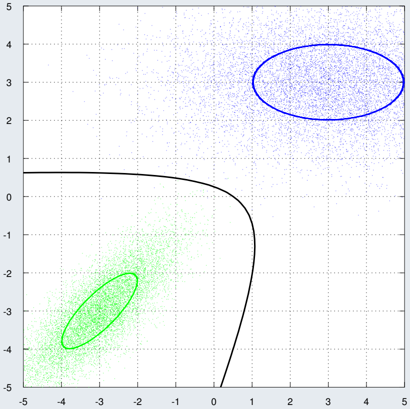
An Example: Naïve Bayes and Gaussians (cont.)

Example cont.



An Example: Naïve Bayes and Gaussians (cont.)

Example cont.



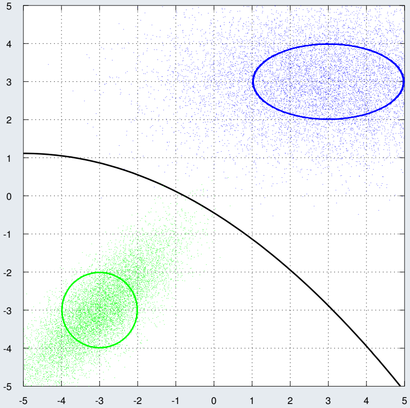
$p(y = 0) = 0.5$

$p(y = 1) = 0.5$

Fig.: Quadratic decision boundary that considers statistical dependency

An Example: Naïve Bayes and Gaussians (cont.)

Example cont.



$p(y=0) = 0.5$

$p(y=1) = 0.5$

Fig.: Quadratic decision boundary assuming independency of x_1 and x_2

Naïve Bayes

Let us consider the logit transform

$$\log \frac{p(y=0|\boldsymbol{x})}{p(y=1|\boldsymbol{x})} \quad =$$

Naïve Bayes

Let us consider the logit transform

$$\log \frac{p(y=0|\boldsymbol{x})}{p(y=1|\boldsymbol{x})} \quad = \quad \log \frac{p(y=0)p(\boldsymbol{x}|y=0)}{p(y=1)p(\boldsymbol{x}|y=1)}$$

Naïve Bayes

Let us consider the logit transform

$$\log \frac{p(y=0|\boldsymbol{x})}{p(y=1|\boldsymbol{x})} = \log \frac{p(y=0)p(\boldsymbol{x}|y=0)}{p(y=1)p(\boldsymbol{x}|y=1)}$$

$$= \log \frac{p(y=0)}{p(y=1)} + \log \frac{p(\boldsymbol{x}|y=0)}{p(\boldsymbol{x}|y=1)}$$

Naïve Bayes

Let us consider the logit transform

$$
\begin{aligned}
\log \frac{p(y=0|\boldsymbol{x})}{p(y=1|\boldsymbol{x})} &= \log \frac{p(y=0)p(\boldsymbol{x}|y=0)}{p(y=1)p(\boldsymbol{x}|y=1)} \\
&= \log \frac{p(y=0)}{p(y=1)} + \log \frac{p(\boldsymbol{x}|y=0)}{p(\boldsymbol{x}|y=1)} \\
&= \log \frac{p(y=0)}{p(y=1)} + \log \frac{\prod_{i=1}^{d} p(x_i|y=0)}{\prod_{i=1}^{d} p(x_i|y=1)}
\end{aligned}
$$

Naïve Bayes

Let us consider the logit transform

$$\begin{aligned}
\log \frac{p(y=0|\boldsymbol{x})}{p(y=1|\boldsymbol{x})} &= \log \frac{p(y=0)p(\boldsymbol{x}|y=0)}{p(y=1)p(\boldsymbol{x}|y=1)} \\
&= \log \frac{p(y=0)}{p(y=1)} + \log \frac{p(\boldsymbol{x}|y=0)}{p(\boldsymbol{x}|y=1)} \\
&= \log \frac{p(y=0)}{p(y=1)} + \log \frac{\prod_{i=1}^{d} p(x_i|y=0)}{\prod_{i=1}^{d} p(x_i|y=1)} \\
&= \underbrace{\alpha_0 + \sum_{i=1}^{d} \alpha_{0,i}(x_i)}_{\text{generalized additive model}}
\end{aligned}$$

Naïve Bayes (cont.)

Is there anything between Bayes and Naïve Bayes?

Naïve Bayes (cont.)

There are multiple techniques to beat the curse of dimensionality, for example:

- Reduction of the parameter space
  - Introduction of independency assumptions (from complete dependency to mutual independency)
  - Parameter tying
- Reduction of the dimension of the feature vectors

Naïve Bayes (cont.)

First order dependency

$$p(\boldsymbol{x}|y) \quad = \quad p(x_1, x_2, \ldots, x_d|y)$$

Naïve Bayes (cont.)

First order dependency

$$
\begin{aligned}
p(\boldsymbol{x}|y) &= p(x_1, x_2, \ldots, x_d|y) \\
&= p(x_1|y)p(x_2, x_3, \ldots, x_d|y, x_1)
\end{aligned}
$$

Naïve Bayes (cont.)

First order dependency

$$
\begin{aligned}
p(\boldsymbol{x}|y) &= p(x_1, x_2, \ldots, x_d|y) \\
&= p(x_1|y)p(x_2, x_3, \ldots, x_d|y, x_1) \\
&= p(x_1|y)p(x_2|y, x_1)p(x_3, x_4, \ldots, x_d|y, x_1, x_2)
\end{aligned}
$$

Naïve Bayes (cont.)

First order dependency

$$
\begin{aligned}
p(\boldsymbol{x}|y) &= p(x_1, x_2, \dots, x_d|y) \\
&= p(x_1|y)p(x_2, x_3, \dots, x_d|y, x_1) \\
&= p(x_1|y)p(x_2|y, x_1)p(x_3, x_4, \dots, x_d|y, x_1, x_2) \\
&= p(x_1|y)\prod_{i=2}^{d} p(x_i|y, x_{i-1})
\end{aligned}
$$

Naïve Bayes (cont.)

Example

First order dependency in a Gaussian random vector can be identified through the covariance matrix $\boldsymbol{\Sigma}$. It has the following structure:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{1,1} & \sigma_{2,1} & 0 & 0 & \cdots & 0 & 0 \\ \sigma_{2,1} & \sigma_{2,2} & \sigma_{3,2} & 0 & \cdots & 0 & 0 \\ 0 & \sigma_{3,2} & \sigma_{3,3} & \sigma_{4,3} & \cdots & 0 & 0 \\ 0 & 0 & \sigma_{4,3} & \sigma_{4,4} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \sigma_{d,d-1} \\ 0 & 0 & 0 & 0 & \cdots & \sigma_{d,d-1} & \sigma_{d,d} \end{pmatrix}$$

Naïve Bayes (cont.)

Example

First order dependency in Gaussian random vector with tied diagonal elements, i. e. $\sigma_{i,i} = \sigma$:

$$\Sigma = \begin{pmatrix} \sigma & \sigma_{2,1} & 0 & 0 & \cdots & 0 & 0 \\ \sigma_{2,1} & \sigma & \sigma_{3,2} & 0 & \cdots & 0 & 0 \\ 0 & \sigma_{3,2} & \sigma & \sigma_{4,3} & \cdots & 0 & 0 \\ 0 & 0 & \sigma_{4,3} & \sigma & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \sigma_{d,d-1} \\ 0 & 0 & 0 & 0 & \cdots & \sigma_{d,d-1} & \sigma \end{pmatrix}$$

Lessons Learned

- Naïve Bayes is rather successful.
- Naïve Bayes does not require a huge set of training data.
- Statistical dependency vs. dimension of the search space.
- Naïve Bayes: give it a try!

Pattern Recognition Lab

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

TECHNISCHE FAKULTÄT

Next Time in

Pattern Recognition

Further Readings

- Brian D. Ripley:
  Pattern Recognition and Neural Networks,
  Cambridge University Press, Cambridge, 1996.

- Christopher M. Bishop:
  Pattern Recognition and Machine Learning,
  Springer, New York, 2006

Comprehensive Questions

- What is the assumption of Naïve Bayes?
- How does the assumption affect the class dependent pdf?
- What is the structure of the covariance matrix of normal-distributed classes in Naïve Bayes?
- How can Naïve Bayes be extended to first-order statistical dependencies?