Communication Technologies 1 (CT1)

Machine Learning

Bayesian Classification for Activity Recognition

M.Sc. Michel Morold

Lecture in SS 2019

09.05.2019

Contents

Introduction – Sensor-based Activity Recognition

Sensor Data

Segmentation → Feature Extraction → Model creation → Model evaluation

Segmentation / Features (JH)

Bayesian Classification (MM)
Decision Tree Classifier (CA)
Clustering/ KNN (MB)
Hidden Markov Models (IK)

Introduction – Machine Learning

ComTec

Supervised

- Bayesian statistics
- Decision trees
- Artificial neural network
- Support vector machines
- Hidden Markov models
- Etc…

Unsupervised

- Data clustering
- Self-organizing map
- Artificial neural network
- Expectation-maximization algorithm
- Etc…

Reinforcement

- Monte Carlo Method
- Q-learning
- Temporal difference learning
- Learning Automata
- Etc…

Introduction – Machine Learning

ComTec

- Supervised Classification
  - Basic task in data analysis and pattern recognition
  - Construction of a classifier from a *training set* with preclassified instances (Building a model)
    - A *training set* consists of instances (set of attributes and a corresponding class label)
  - The classifier then assigns previously unseen instances from a *test set* to class labels
    - A *test set* (instances without class labels) is used to evaluate the classifier’s performance

Introduction – Machine Learning

- Bayesian Classification
  - Supervised Machine Learning based on the Bayes' Theorem
  - Statistical method for classification
  - Bayesian classifiers are able to deal with issues of uncertainty and noise
  - The most famous example for Bayesian Classification is the Naive Bayes classifier which represents a simplified Bayesian Network

Probability theory

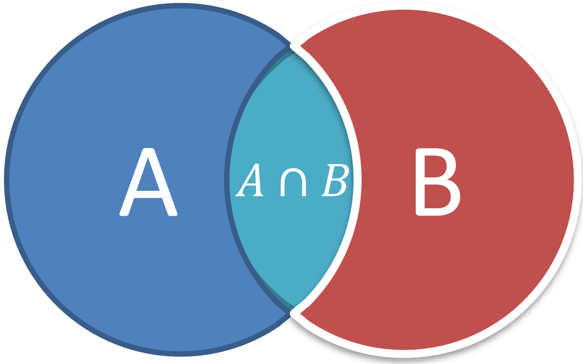
ComTec

- Given two events $A, B \subseteq \Omega$ we define
  - $p(A)$ as probability that A will occur
  - $p(B)$ as probability that B will occur
  - $p(A \cap B)$ as probability that A **and** B will occur
  - $p(A|B)$ as probability of A given that B already occured
  - $p(B|A)$ as probability of B given that A already occured

Probability theory

- Given two events A, B, with $p(A) > 0$ and $p(B) > 0$, then

$$p(A \cap B) = p(A) \cdot p(B|A) = p(B) \cdot p(A|B)$$



Probability theory

ComTec

- Example: We have a bag of 5 old and 10 new batteries. Now you randomly pick two batteries. What is the probability that you pick two new Batteries?
  - Let $A =$ "$First\ battery\ is\ new$" and $\mathrm{B} =$ "$Second\ battery\ is\ new$"
  - Calculate $p(A \cap B) = p(A) \cdot p(B|A)$

$$p(A \cap B) = p(A) \cdot p(B|A) = \frac{10}{15} \cdot \frac{9}{14} \approx 0.43$$

Probability theory

ComTec

- Given two events A, B, with $p(A) > 0$ and $p(B) > 0$, A and B are statistically independent when

$$p(B|A) = p(B) \qquad \text{with } p(A) > 0$$

$$p(A|B) = p(A) \qquad \text{with } p(B) > 0$$

$$p(A \cap B) = p(A) \cdot p(B)$$

Probability theory

ComTec

- A random variable X is a function, whose possible values are numerical outcomes of a random phenomenon and which maps every elementary event ω to a real number

$$
\begin{aligned}
X\colon\ \Omega &\to \mathbb{R} \\
\omega &\longmapsto X(\omega)
\end{aligned}
$$

Probability theory

ComTec

- A **discrete** random variable can take only a finite number of distinct values

- For every value x_i, the probability that the outcome of a discrete random variable X is equal to x_i is

$$p_i = \mathrm{P}(\mathrm{X} = x_i)$$

Probability theory

- A **continuous** random variable is continuous if its cumulative distribution function $F_x(x) = P(X \leq x)$ is continuous everywhere

- The cumulative distribution function can be expressed as the integral of its probability density function (f_x):

$$\mathrm{F}_x(x) = \int_{-\infty}^{x} f_x(t)dt$$

Probability theory

- The probability that X lies in the interval $[a, b]$ with $a < b$ is

$$P(a \leq X \leq b) = F_x(b) - F_x(a) = \int_a^b f(t)dt$$

Bayes Theorem

ComTec

- The Bayes theorem describes the calculation of conditioned probabilities
- It is named after **Thomas Bayes** (1701 – 1761), an English statistician, philosopher and Presbyterian minister who studied at the University of Edinburgh

[1]

Bayes Theorem: Definition

ComTec

- Given two events A, B, with $p(B) > 0$, the probability of A given that B already occurred, can be calculated by

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)}$$

- $p(A|B)$ is the posterior probability
- $p(B|A)$ is the likelihood function
- $p(A)$ and $p(B)$ are prior probabilities

Bayes Theorem

- Let $A_1, A_2, \dots A_n$ be a partition of the probability space, and $B \subseteq \Omega$. The probability for one of these events given that B already occurred is

$$p\left(A_j \middle| B\right) = \frac{p\left(B \middle| A_j\right) \cdot p(A_j)}{\sum_{k=1}^{n} p(B|A_k) \cdot p(A_k)}$$

- This rule also called *law of total probability*

Bayes Theorem – Example

Assume the following facts

- 25 % of all your e-mails are spam
- The probability, that a spam e-mail contains the word "lottery" is 19 %
- The probability, that a non-spam e-mail contains the word "lottery" is 1%

Question: Assume you receive an e-mail x, which contains the word "lottery", what is the probability that this e-mail is spam?

=> $p(\text{"}x \text{ is spam"} \mid \text{"}x \text{ contains the word 'lottery'"})$?

Bayes Theorem – Example

ComTec

A = "x is spam"
B = "x contains the word 'lottery' "

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)}$$

We do not know $p(B)$ directly, but we can use the law of total probability:

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B|A) \cdot p(A) + p(B|\bar{A}) \cdot p(\bar{A})}$$

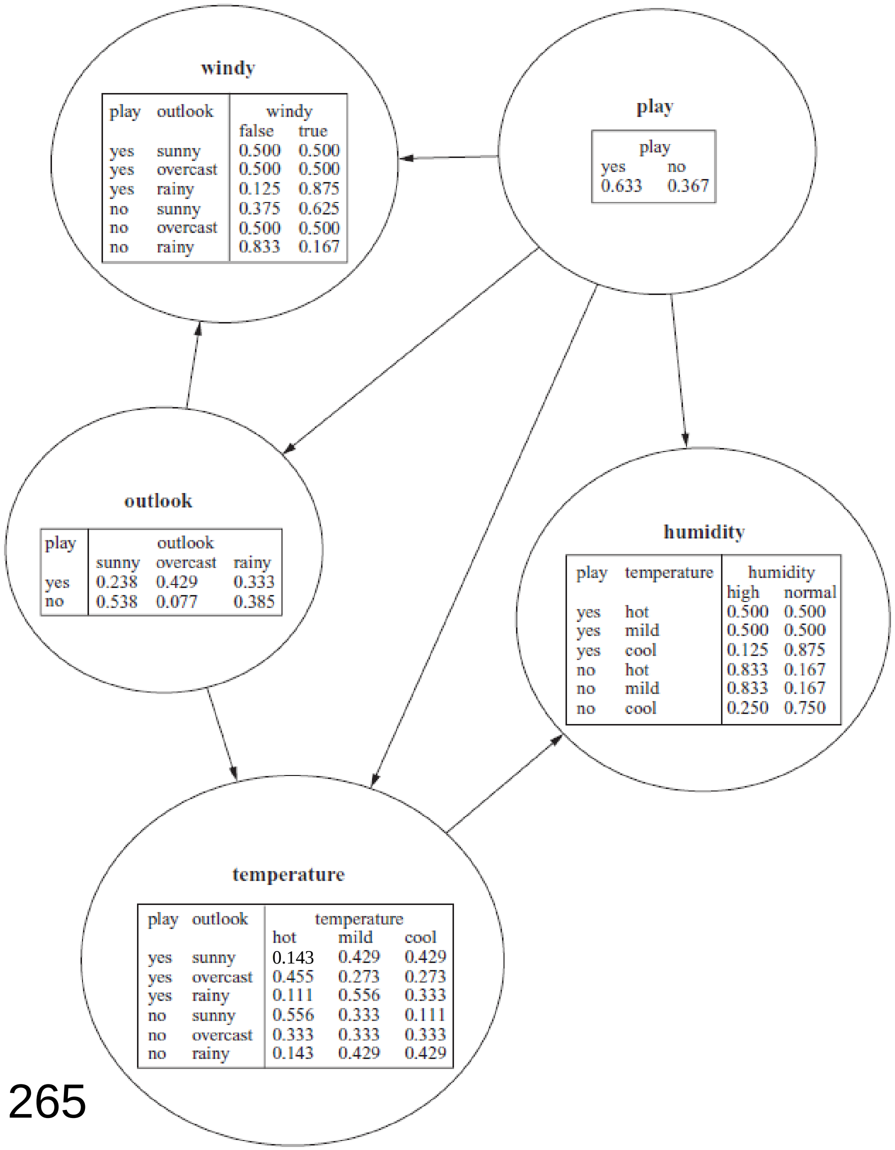
$$= \frac{0.19 \cdot 0.25}{(0.19 \cdot 0.25) + (0.01 \cdot 0.75)} = 0.864 = 86.4\ \%$$

Bayesian Networks

- Represented by a directed acyclic graph (DAG)
- Nodes are random variables $X_1, \dots, X_n$
  - Each node contains a conditional probability table
- Directed edges describe conditional dependencies
- Each random variable is conditionally independent of all its nondescendants given its parents

$$p(X_1, \dots, X_n) = \prod_{i=1}^{n} p(X_i | X_{i-1}, \dots, X_1) = \prod_{i=1}^{n} p(X_i | X_i' s\ parents)$$

Bayesian Networks – Example



[4] p. 265

Consider an instance E

- $outlook = rainy$
- $temperature = cool$
- $humidity = high$
- $windy = true$

What is the probability for $play = no$ and $play = yes$ given the evidence E above?

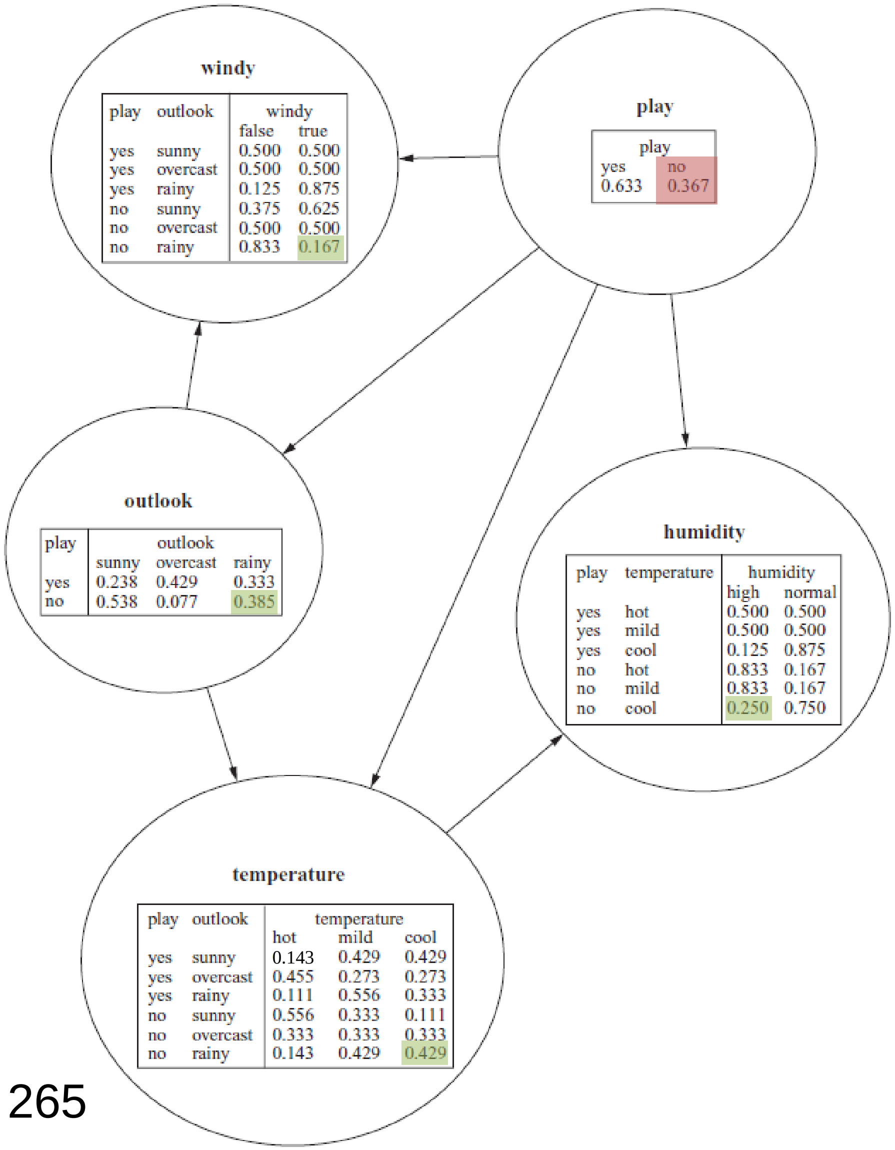
Bayesian Networks – Example

ComTec

Let $E = (outlook = rainy, temp = cool, hum = high, windy = true)$

$$p(play = \boldsymbol{no} \mid E) = \frac{p(E|play = no) \cdot p(play = no)}{p(E)}$$

Bayesian Networks – Example



$$
\begin{aligned}
E \;&= (outlook = rainy,\\
&temp = cool,\\
&hum = high,\\
&windy = true)
\end{aligned}
$$

$$\frac{p(E|play = no) \cdot p(play = no)}{p(E)}$$

$$\frac{(0.385 \cdot 0.429 \cdot 0.250 \cdot 0.167) \cdot 0.367}{p(E)}$$

[4] p. 265

Bayesian Networks – Example

$$E = (outlook = rainy, temp = cool, hum = high, windy = true)$$

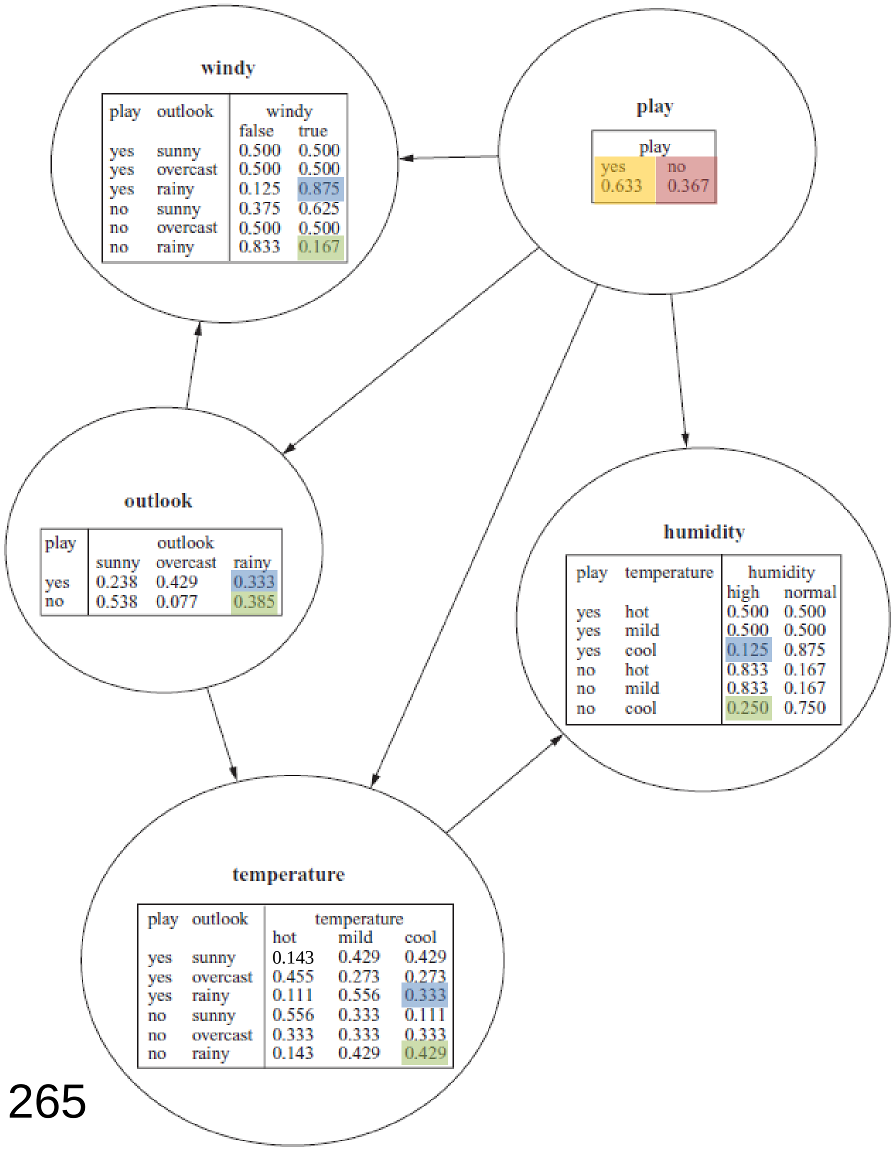
$$p(play = \boldsymbol{no} \mid E) = \frac{p(E \mid play = no) \cdot p(play = no)}{p(E)}$$

$$= \frac{(0.385 \cdot 0.429 \cdot 0.250 \cdot 0.167) \cdot 0.367}{p(E)}$$

$$= \frac{(0.385 \cdot 0.429 \cdot 0.250 \cdot 0.167) \cdot 0.367}{p(E \mid play = no) \cdot p(play = no) + p(E \mid play = yes) \cdot p(play = yes)}$$

Bayesian Networks – Example

ComTec



windy

play	outlook	windy: false	windy: true
yes	sunny	0.500	0.500
yes	overcast	0.500	0.500
yes	rainy	0.125	0.875
no	sunny	0.375	0.625
no	overcast	0.500	0.500
no	rainy	0.833	0.167

play

play: yes	play: no
0.633	0.367

outlook

play	outlook: sunny	outlook: overcast	outlook: rainy
yes	0.238	0.429	0.333
no	0.538	0.077	0.385

humidity

play	temperature	humidity: high	humidity: normal
yes	hot	0.500	0.500
yes	mild	0.500	0.500
yes	cool	0.125	0.875
no	hot	0.833	0.167
no	mild	0.833	0.167
no	cool	0.250	0.750

temperature

play	outlook	temperature: hot	temperature: mild	temperature: cool
yes	sunny	0.143	0.429	0.429
yes	overcast	0.455	0.273	0.273
yes	rainy	0.111	0.556	0.333
no	sunny	0.556	0.333	0.111
no	overcast	0.333	0.333	0.333
no	rainy	0.143	0.429	0.429

$$\frac{(0.385 \cdot 0.429 \cdot 0.250 \cdot 0.167) \cdot 0.367}{(0.385 \cdot 0.429 \cdot 0.250 \cdot 0.167) \cdot 0.367 + (0.333 \cdot 0.333 \cdot 0.125 \cdot 0.875) \cdot 0.633}$$

[4] p. 265

Bayesian Networks – Example

ComTec

Let $E = (outlook = rainy, temp = cool, hum = high, windy = true)$

$$p(play = \boldsymbol{no} \mid E) = \frac{p(E \mid play = no) \cdot p(play = no)}{p(E)}$$

$$= \frac{(0.385 \cdot 0.429 \cdot 0.250 \cdot 0.167) \cdot 0.367}{p(E)}$$

$$= \frac{(0.385 \cdot 0.429 \cdot 0.250 \cdot 0.167) \cdot 0.367}{p(E \mid play = no) \cdot p(play = no) + p(E \mid play = yes) \cdot p(play = yes)}$$

$$= \frac{(0.385 \cdot 0.429 \cdot 0.250 \cdot 0.167) \cdot 0.367}{(0.385 \cdot 0.429 \cdot 0.250 \cdot 0.167) \cdot 0.367 + (0.333 \cdot 0.333 \cdot 0.125 \cdot 0.875) \cdot 0.633}$$

$$\approx 0.245 = 24.5\ \%$$

Bayesian Networks – Example

Let $E = (outlook = rainy, temp = cool, hum = high, windy = true)$

$$p(play = \boldsymbol{yes} \mid E) = \frac{p(E \mid play = yes) \cdot p(play = yes)}{p(E)}$$

Bayesian Networks – Example

ComTec

Let $E \ = (outlook = rainy, temp = cool, hum = high, windy = true)$

$$p(play = \boldsymbol{yes} \mid E) = \frac{p(E|\, play = yes) \cdot p(play = yes)}{p(E)}$$

$$= \frac{(0.333 \cdot 0.333 \cdot 0.125 \cdot 0.875) \cdot 0{,}633}{p(E)}$$

Bayesian Networks – Example

Let $E = (outlook = rainy, temp = cool, hum = high, windy = true)$

$$p(play = \boldsymbol{yes} \mid E) = \frac{p(E \mid play = yes) \cdot p(play = yes)}{p(E)}$$

$$= \frac{(0.333 \cdot 0.333 \cdot 0.125 \cdot 0.875) \cdot 0{,}633}{p(E)}$$

$$= \frac{(0.333 \cdot 0.333 \cdot 0.125 \cdot 0.875) \cdot 0{,}633}{p(E \mid play = no) \cdot p(play = no) + p(E \mid play = yes) \cdot p(play = yes)}$$

Bayesian Networks – Example

ComTec

Let $E = (outlook = rainy, temp = cool, hum = high, windy = true)$

$$p(play = \boldsymbol{yes} \mid E) = \frac{p(E \mid play = yes) \cdot p(play = yes)}{p(E)}$$

$$= \frac{(0.333 \cdot 0.333 \cdot 0.125 \cdot 0.875) \cdot 0{,}633}{p(E)}$$

$$= \frac{(0.333 \cdot 0.333 \cdot 0.125 \cdot 0.875) \cdot 0{,}633}{p(E \mid play = no) \cdot p(play = no) + p(E \mid play = yes) \cdot p(play = yes)}$$

$$= \frac{(0.333 \cdot 0.333 \cdot 0.125 \cdot 0.875) \cdot 0{,}633}{(0.385 \cdot 0.429 \cdot 0.250 \cdot 0.167) \cdot 0.367 + (0.333 \cdot 0.333 \cdot 0.125 \cdot 0.875) \cdot 0.633}$$

$$\approx 0.755 = 75.5\ \%$$

Bayesian Networks – Learning

- Given a training set, the problem of learning a Bayesian network is to find a network that best matches the training set
- A common approach is to introduce a **scoring function**, which evaluates a network with regard to the training data
- Then search for the best network according to this scoring function
  - Determine nodes and edges
  - Calculate probability tables for each node

Naive Bayes

- Developed for supervised machine learning
- Assumptions
  - All predictive attributes are statistically independent given the class
  - No hidden or latent attribute influences the prediction process
- Able to deal with missing attribute values

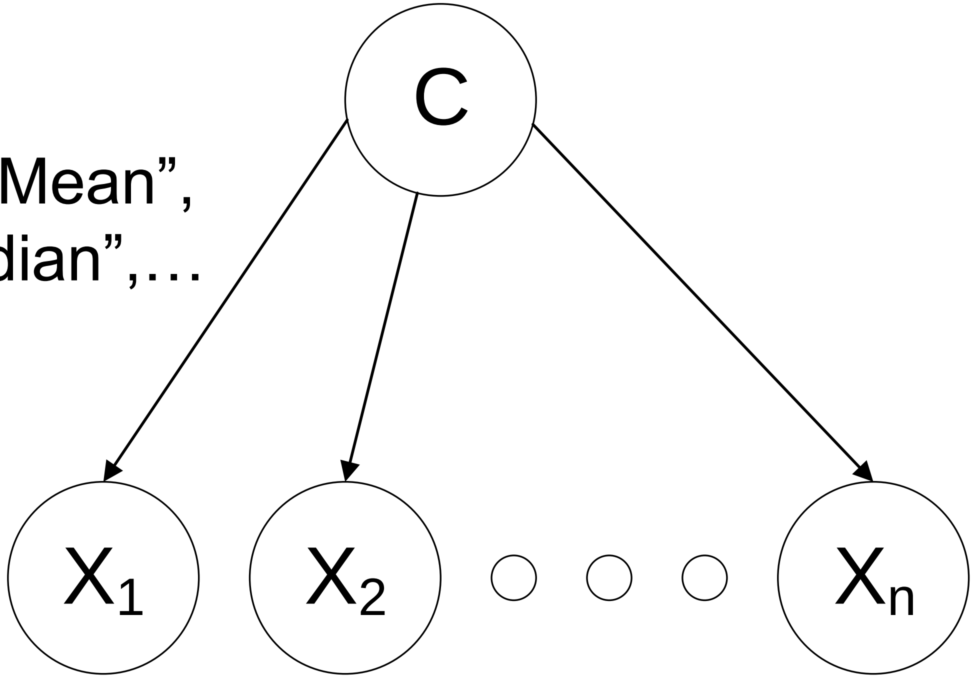
Naive Bayes

- Common Applications for Naive Bayes
  - Document / Text Classification (e.g. Spam Filter)
  - Activity Recognition

- Despite the simplified approach, the Naive Bayes Classifier is competitive with more sophisticated classifiers on real-world datasets (Langley et al. [6])

Naive Bayes as Bayesian Network

- A Bayesian network for the Naive Bayes classifier consists of n+1 nodes
- $X_1 \dots X_n$ attributes
  - Examples:
    "X-axis Accelerometer Mean",
    "Z-axis Gyroscope Median",…
- C class
  - Possible values:
    "Running", "Sitting",
    "Standing"…



Naive Bayes

ComTec

- Given a test case $\boldsymbol{x}$ to classify, Bayes‘ rule is used to calculate the probability of each class given the attribute vector and then predicts the most probable class c:

$$p(C = c|\boldsymbol{X} = \boldsymbol{x}) = \frac{p(C = c) \cdot p(\boldsymbol{X} = \boldsymbol{x}|C = c)}{p(\boldsymbol{X} = \boldsymbol{x})}$$

C	random variable denoting the class of an instance
c	a particular class label
X	random variable denoting the observed attribute value
x	a particular observed attribute value
$C = c$	represents that C equals class c
$\boldsymbol{X} = \boldsymbol{x}$	represents $X_1 = x_1 \wedge X_2 = x_2 \wedge \cdots \wedge X_k = x_k$

Naive Bayes

ComTec

- Given a test case $\boldsymbol{x}$ to classify, Bayes‘ rule is used to calculate the probability of each class given the attribute vector:

$$p(C = c|\boldsymbol{X} = \boldsymbol{x}) = \frac{p(C = c) \cdot p(\boldsymbol{X} = \boldsymbol{x}|C = c)}{p(\boldsymbol{X} = \boldsymbol{x})}$$

- $p(C = c)$ is the probability that the class is c, which can be calculated by

$$p(C = c) = \frac{\#\ of\ instances\ with\ class\ c\ in\ training\ set}{\#\ of\ all\ instances\ in\ training\ set}$$

Naive Bayes

ComTec

- Given a test case $\boldsymbol{x}$ to classify, Bayes‘ rule is used to calculate the probability of each class given the attribute vector:

$$p(C = c|\boldsymbol{X} = \boldsymbol{x}) = \frac{p(C = c) \cdot p(\boldsymbol{X} = \boldsymbol{x}|C = c)}{p(\boldsymbol{X} = \boldsymbol{x})}$$

- As $C_1, C_2, \dots, C_n$ is a partition of the probability space, we can write

$$p(\boldsymbol{X} = \boldsymbol{x}) = \sum_i p(C_i = c_i) \cdot p(\boldsymbol{X} = \boldsymbol{x}|\, C_i = c_i)$$

Naive Bayes

ComTec

- Given a test case $\boldsymbol{x}$ to classify, Bayes‘ rule is used to calculate the probability of each class given the attribute vector:

$$p(C = c|\boldsymbol{X} = \boldsymbol{x}) = \frac{p(C = c) \cdot p(\boldsymbol{X} = \boldsymbol{x}|C = c)}{\sum_i p(C_i = c_i) \cdot p(\boldsymbol{X} = \boldsymbol{x}|C_i = c_i)}$$

- As $\boldsymbol{X} = \boldsymbol{x}$ equals $X_1 = x_1 \land X_2 = x_2 \land \cdots \land X_k = x_k$ and all attributes are assumed to be conditionally independent, we can write

$$p(\boldsymbol{X} = \boldsymbol{x}|C = c) = \prod_i p(X_i = x_i|C = c)$$

Naive Bayes

ComTec

- Given a test case $\boldsymbol{x}$ to classify, Bayes' rule is used to calculate the probability of each class given the attribute vector:

$$p(C = c|\boldsymbol{X} = \boldsymbol{x}) = \frac{p(C = c) \cdot p(\boldsymbol{X} = \boldsymbol{x}|C = c)}{\sum_i p(C_i = c_i) \cdot p(\boldsymbol{X} = \boldsymbol{x}|C_i = c_i)}$$

- Next question: How does the Naive Bayes algorithm calculate $p(\boldsymbol{X} = \boldsymbol{x}|C = c)$?

Naive Bayes – discrete / numeric attributes

ComTec

- For each **discrete** attribute, $p(X = x|C = c)$ is modelled as probability, that the attribute X will take on the particular value x, when the class is c

- **Numeric** attributes are modeled by a continuous probability distribution of the range of the attribute’s values

Naive Bayes – numeric attributes

- The Naive Bayes classifier makes the assumption, that within each class the values of numeric attributes are **normally distributed**
- This distribution is modelled over the attribute's **mean μ** and **standard deviation *σ*** as probability density function:

$$p(X = x|C = c) = g(x;\ \mu_c, \sigma_c)$$

$$g(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Naive Bayes – numeric attributes

ComTec

Keep in mind that this formula

$$p(X = x|C = c) = g(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

is not strictly correct, as the probability that a real-valued random variable equals any value is 0. Therefore, we write

$$p(x \leq X \leq x + \Delta) = \int_x^{x+\Delta} g(x; \mu, \sigma)\, dx$$

$$\lim_{\Delta \to 0} \frac{p(x \leq X \leq x + \Delta)}{\Delta} = g(x; \mu, \sigma)$$

For a very small Δ,

$$p(X = x) \approx g(x; \mu, \sigma) \cdot \Delta$$

This factor Δ then appears in the numerator and cancels out when the normalization is done.

Naive Bayes – numeric attributes

- Training: For each continuous attribute X, the mean and standard deviation is calculated given the class

$$p(X = x | C = c) = g(x;\ \mu_c, \sigma_c)$$

- Let $\{x_1, \ldots, x_n\}$ be all values for a continuous attribute X with class c, then calculate mean μ_c and standard deviation σ_c:

$$\mu_c = \frac{1}{n} \sum_i x_i$$

$$\sigma_c = \frac{1}{n-1} \sum_i (x_i - \mu_c)^2$$

Naive Bayes – Example

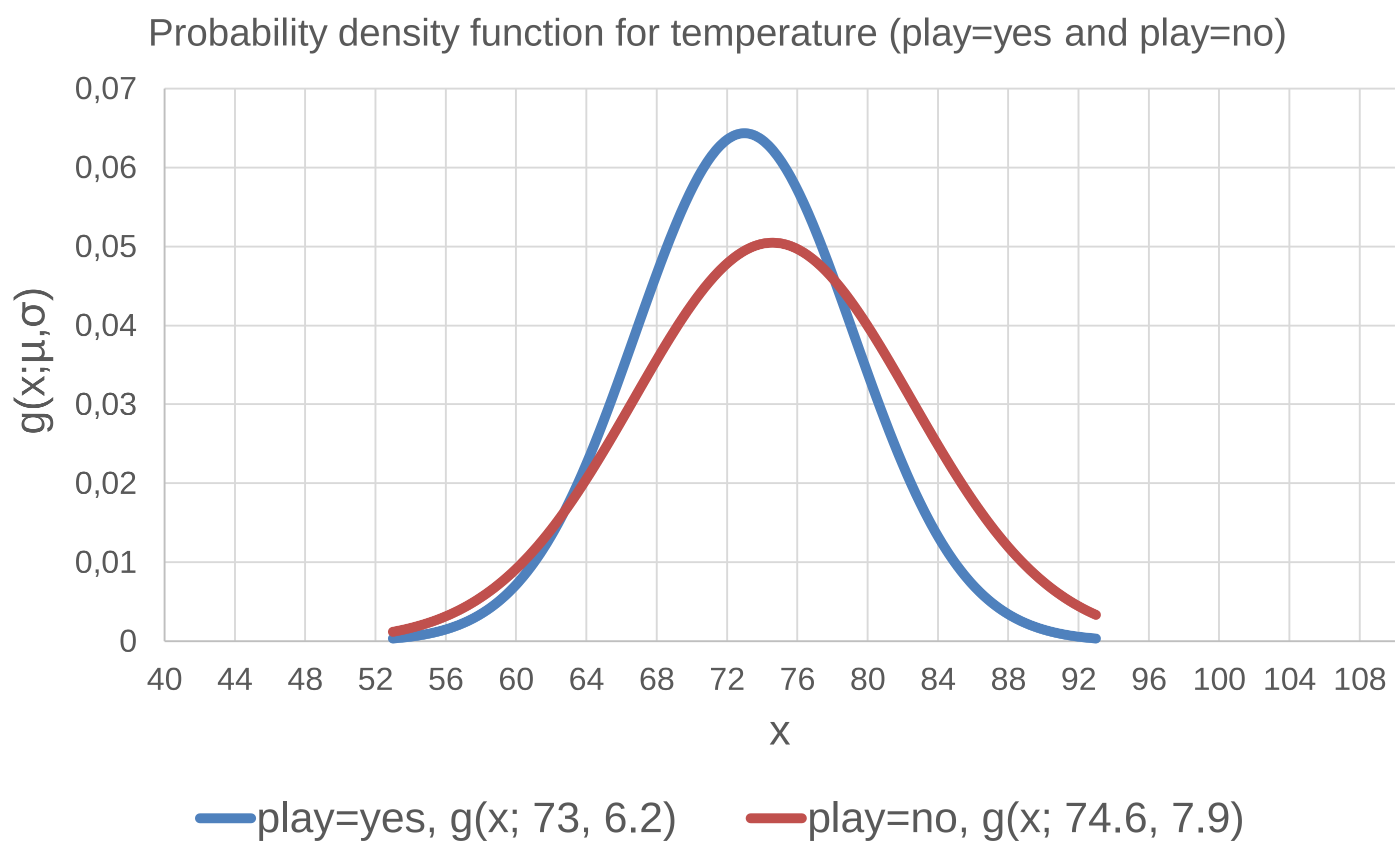
- Assume the following data set

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	mean	73	74.6	mean	79.1	86.2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	std. dev.	6.2	7.9	std. dev.	10.2	9.7	true	3/9	3/5		
rainy	3/9	2/5											

Naive Bayes – Example

ComTec

$$g(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Naive Bayes – Example

ComTec

- Now calculate the probability for classes $play = yes$ and $play = no$ if $outlook = sunny, temperature = 66, \ humidity = 90, \ windy = true$
- Predict the most probable class (“yes” or “no”) by calculating for each class:

$$p(C = c|\boldsymbol{X} = \boldsymbol{x}) = \frac{p(C = c) \cdot p(\boldsymbol{X} = \boldsymbol{x}|C = c)}{\sum_i p(C_i = c_i) \cdot p(\boldsymbol{X} = \boldsymbol{x}|C_i = c_i)}$$

Naive Bayes – Example

ComTec

- Calculate $p(C = c|\boldsymbol{X} = \boldsymbol{x})$ for $C = play$, $c = yes\ (and\ no)$ and $\boldsymbol{X} = \boldsymbol{x}$: $outlook = sunny,\ temperature = 66,\ humidity = 90,\ windy = true$

$$p(play = yes|\boldsymbol{X} = \boldsymbol{x})$$

$$= \frac{p(play = yes) \cdot p(\boldsymbol{X} = \boldsymbol{x}|play = yes)}{p(play = yes) \cdot p(\boldsymbol{X} = \boldsymbol{x}|play = yes) + p(play = no) \cdot p(\boldsymbol{X} = \boldsymbol{x}|play = no)}$$

Naive Bayes – Example

ComTec

- Assume the following data set

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	mean	73	74.6	mean	79.1	86.2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	std. dev.	6.2	7.9	std. dev.	10.2	9.7	true	3/9	3/5		
rainy	3/9	2/5											

p(play=no)

p(play=yes)

Naive Bayes – Example

ComTec

- Calculate $p(C = c|\boldsymbol{X} = \boldsymbol{x})$ for $C = play$, $c = yes$ $(and\ no)$ and $\boldsymbol{X} = \boldsymbol{x}$: $outlook = sunny,\ temperature = 66,\ humidity = 90,\ windy = true$

$$p(play = yes|\boldsymbol{X} = \boldsymbol{x})$$

$$= \frac{p(play = yes) \cdot p(\boldsymbol{X} = \boldsymbol{x}|play = yes)}{p(play = yes) \cdot p(\boldsymbol{X} = \boldsymbol{x}|play = yes) + p(play = no) \cdot p(\boldsymbol{X} = \boldsymbol{x}|play = no)}$$

$$= \frac{\frac{9}{14} \cdot p(\boldsymbol{X} = \boldsymbol{x}|play = yes)}{\frac{9}{14} \cdot p(\boldsymbol{X} = \boldsymbol{x}|play = yes) + \frac{5}{14} \cdot p(\boldsymbol{X} = \boldsymbol{x}|play = no)}$$

- Next: Calculate $p(\boldsymbol{X} = \boldsymbol{x}|play = yes)$ and $p(\boldsymbol{X} = \boldsymbol{x}|play = no)$

Naive Bayes – Example

- Now, calculate $p(\boldsymbol{X} = \boldsymbol{x} | play = yes) = \prod_i p(X_i = x_i | play = yes)$:

$$p(outlook = sunny \mid play = yes) = \frac{2}{9}$$

$$p(temperature = 66 \mid play = yes) = g(66; 73, 6.2) = \frac{1}{\sqrt{2\pi} \cdot 6.2} e^{-\frac{(66-73)^2}{2\cdot(6.2)^2}} = 0.034$$

$$p(humidity = 90 \mid play \ = \ yes) = g(90; 79.1, 10.2) = \frac{1}{\sqrt{2\pi} \cdot 10.2} e^{-\frac{(90-79.1)^2}{2\cdot(10.2)^2}} = 0.0221$$

$$p(windy = true \mid play = yes) = \frac{3}{9}$$

- ➢ $p(\boldsymbol{X} = \boldsymbol{x} | play = yes) = \frac{2}{9} \cdot 0.034 \cdot 0.0221 \cdot \frac{3}{9}$

- ➢ Analog: $p(\boldsymbol{X} = \boldsymbol{x} | play = no) = \frac{3}{5} \cdot 0.0279 \cdot 0.0381 \cdot \frac{3}{5}$

Naive Bayes – Example

ComTec

- Calculate $p(C = c|\boldsymbol{X} = \boldsymbol{x})$ for $C = play$, $c = yes$ $(and\ no)$ and $\boldsymbol{X} = \boldsymbol{x}$: $outlook = sunny,\ temperature = 66,\ humidity = 90,\ windy = true$

$$p(play = yes|\boldsymbol{X} = \boldsymbol{x})$$

$$= \frac{p(play = yes) \cdot p(\boldsymbol{X} = \boldsymbol{x}|play = yes)}{p(play = yes) \cdot p(\boldsymbol{X} = \boldsymbol{x}|play = yes) + p(play = no) \cdot p(\boldsymbol{X} = \boldsymbol{x}|play = no)}$$

$$= \frac{\frac{9}{14} \cdot p(\boldsymbol{X} = \boldsymbol{x}|play = yes)}{\frac{9}{14} \cdot p(\boldsymbol{X} = \boldsymbol{x}|play = yes) + \frac{5}{14} \cdot p(\boldsymbol{X} = \boldsymbol{x}|play = no)}$$

$$= \frac{\frac{9}{14} \cdot \frac{2}{9} \cdot 0.034 \cdot 0.0221 \cdot \frac{3}{9}}{\frac{9}{14} \cdot \frac{2}{9} \cdot 0.034 \cdot 0.0221 \cdot \frac{3}{9} + \frac{5}{14} \cdot \frac{3}{5} \cdot 0.0279 \cdot 0.0381 \cdot \frac{3}{5}} = 0.208 = 20.8\%$$

Naive Bayes – Example

- Calculate $p(C = c|\boldsymbol{X} = \boldsymbol{x})$ for $C = play$ and $\boldsymbol{X} = \boldsymbol{x}$: $outlook = sunny,$ $temperature = 66,\ humidity = 90,\ windy = true$

$p(play = yes|\boldsymbol{X} = \boldsymbol{x}) = 20.8\ \%$

$p(play = no|\boldsymbol{X} = \boldsymbol{x}) = 79.2\ \%$

➢ The Naive Bayes Classifier would predict $play = no$ as the most probable class in this case

Extension: Flexible Naive Bayes

ComTec

- The Flexible Naive Bayes learning algorithm uses a different kernel density estimation for continuous attributes:

$$p(X = x|C = c) = \frac{1}{n}\sum_i g(x;\ \mu_i, \sigma_c)$$

with $\mu_i = x_i$ and $\sigma_c = \frac{1}{\sqrt{n_c}}$

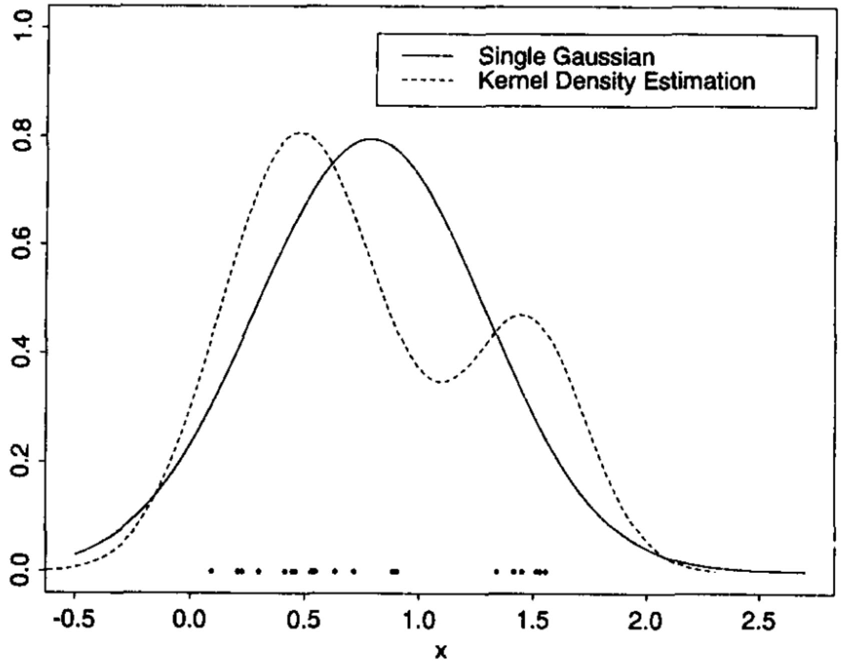
(n_c : number of training instances in c)

- Flexible Naive Bayes stores every continuous attribute and performs n evaluations, one per observed value of X in class c

Flexible Naive Bayes – Example

ComTec

- Example Gaussian vs. Kernel method to estimate density of a continuous variable



[3]

Flexible Naive Bayes

ComTec

- Algorithmic complexity given n training instances and k features

Operation	Naive Bayes		Flexible Naive Bayes	
	Time	Space	Time	Space
Training of n instances	$O(nk)$	$O(k)$	$O(nk)$	$O(nk)$
Test on m instances	$O(mk)$		$O(mnk)$	

Flexible Naive Bayes

- Flexible Naive Bayes
  - Using a kernel density estimation leads to an increase in storage and computational complexity compared to Naive Bayes
  - Flexible Naive Bayes is able to perform better in domains which violate the normality assumption (see [3])

Summary

ComTec

- Bayesian Classification
  - Bayesian classification algorithms are using the Bayes' Theorem for expressing conditional dependencies and are can deal with issues of noise and uncertainties
  - Bayesian Networks represent conditional dependencies in a acyclic directed graph
  - The Naive Bayes classifier can be represented by a simplified Bayesian Network
  - Although the Naive Bayes classifier assumes conditional independence between the inputs and that each continuous attribute is normally distributed, it is competitive with other state-of-the-art classifiers

Sources

ComTec

[1] https://commons.wikimedia.org/wiki/File:Thomas_Bayes.gif#/media/File:Thomas_Bayes.gif (last checked: 16.05.2018)

[2] G. Teschl and S. Teschl, *Mathematik für Informatiker: Bd. 2: Analysis und Statistik*, 3., überarb. Aufl. Berlin ;Heidelberg: Springer Vieweg, 2014.

[3] G. H. John and P. Langley, “Estimating continuous distributions in Bayesian classifiers,” in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 1995, pp. 338–345.

[4] I. H. Witten, E. Frank, and M. A. Hall, *Data mining: practical machine learning tools and techniques*, 3rd ed. Burlington, MA: Morgan Kaufmann, 2011.

[5] N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian Network Classifiers,” *Mach. Learn.*, vol. 29, no. 2–3, pp. 131–163, Nov. 1997.

[6] P. Langley, W. Iba and, and K. Thompson, “An Analysis of Bayesian Classifiers,” in *Proceedings of the Tenth National Conference on Artificial Intelligence*, San Jose, California, 1992, pp. 223–228.