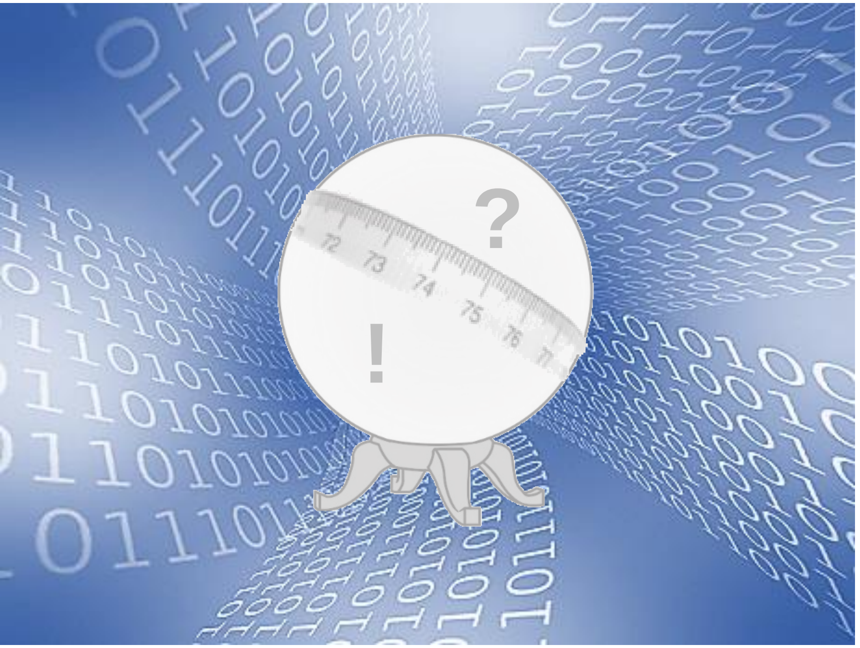
Prof. Dr. Bastian Amberg
School of Business & Economics
Department of Information Systems



Business Intelligence

14 Evidence and Probabilities

(another variant of modeling)

Prof. Dr. Bastian Amberg
(summer term 2024)
3.7.2024

Schedule

		Wed., 10:00-12:00		Fr., 14:00-16:00 (Start at 14:30)		Self-study		
	W1	17.4.	(Meta-)Introduction	19.4.		Python-Basics	Chap. 1	
	W2	24.4.	Data Warehouse – Overview & OLAP	26.4.	*[Blockveranstaltung SE Prof. Gersch]*		Chap. 2	
Basics	W3	1.5.		3.5.			Chap. 3	
	W4	8.5.	Data Warehouse Modeling I & II	10.5.	Data Mining Introduction			
	W5	15.5.	CRISP-DM, Project understanding	17.5.	Python-Basics-Online Exercise	Python-Analytics	Chap. 1	
	W6	22.5.	Data Understanding, Data Visualization I	24.5.	*No lectures, but bonus tasks 1.) Co-Create your exam 2.) Earn bonus points for the exam*		Chap. 2	
	W7	29.5.	Data Visualization II	31.5.				
Main Part	W8	5.6.	Data Preparation	7.6.	Predictive Modeling I (10:00 -12:00)	BI-Project	Start	
	W9	12.6.	Predictive Modeling II	14.6.	Python-Analytics-Online Exercise		|	
	W10	19.6.	*Guest Lecture Dr. Ionescu*	21.6.	Fitting a Model		|	Case Study
	W11	26.6.	How to avoid overfitting	28.6.	What is a good Model?		|	
	W12	3.7.	Project status update Evidence and Probabilities	5.7.	Similarity (and Clusters) From Machine to Deep Learning I		|	
Deep-ening	W13	10.7.		12.7.	From Machine to Deep Learning II		|	
	W14	17.7.	Project presentation	19.7.	Project presentation		End	
Ref.					*Klausur 1.Termin, 31.7.'24 Klausur 2.Termin, 2.10.'24*	Projektbericht		

Last Lesson

- ✓ **Confusion Matrix**

Actual Values

$$\text{Predicted}\quad \begin{array}{c|cc} & \textbf{p} & \textbf{n} \\ \hline \textbf{Y} & \text{True positives} & \text{False positives} \\ \textbf{N} & \text{False negatives} & \text{True negatives} \end{array}$$

Evaluation metrics, for example

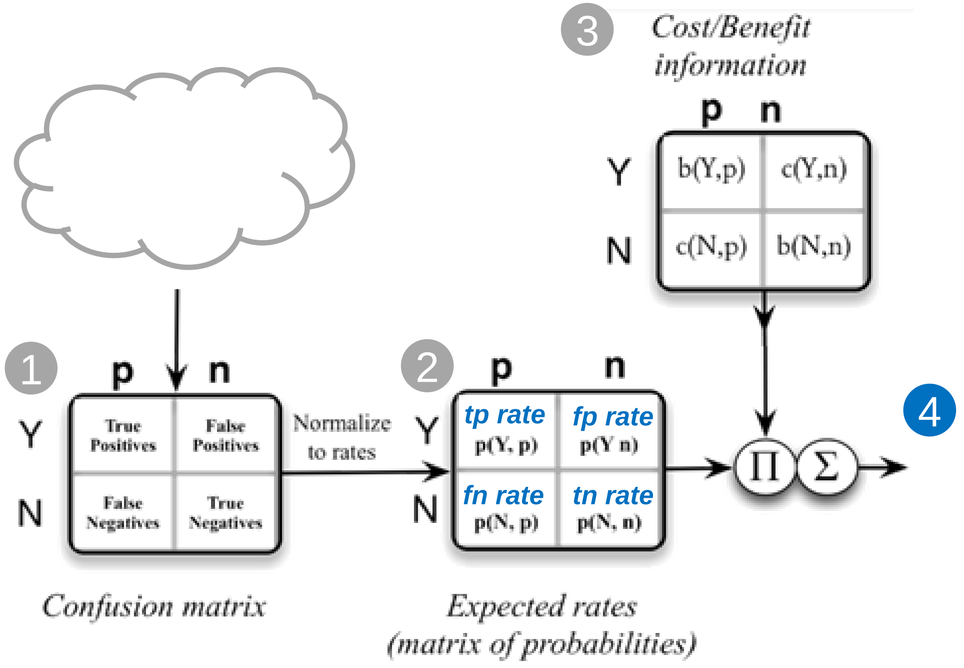
Accuracy (count of correct decisions): $\frac{TP+TN}{P+N}$

True positive rate / Recall / Specificity : $\frac{TP}{TP+FN}$

Precision: $\frac{TP}{TP+FP}$

F-measure (harmonic mean): $2 \cdot \frac{precision \cdot recall}{precision+recall}$

- ✓ **The expected value framework**



4 $$EP = p(p) \cdot [p(Y|p) \cdot b(Y,p) + p(N|p) \cdot c(N,p)] + p(n) \cdot [p(N|n) \cdot b(N,n) + p(Y|n) \cdot c(Y,n)]$$

Example

1

	p	n
Y	56	7
N	5	42

2

$P = 61,$
$p(\mathbf{p}) = 0.55,$
$tp\ rate = 56/61 = 0.92,$
$fn\ rate = 5/61 = 0.08$

$N = 49,$
$p(\mathbf{n}) = 0.45,$
$fp\ rate = 7/49 = 0.14,$
$tn\ rate = 42/49 = 0.86$

3

	p	n
Y	99	-1
N	0	0

4 $= 0.55 \cdot [0.92 \cdot b(Y,p) + 0.08 \cdot c(N,p)] +$
….

This expected value means that….

- ✓ **What is the appropriate baseline for comparison?**

Agenda

Two other variants of modeling

Naive Bayes classifier

k-Nearest Neighbors

Evidence and Probabilities

Similarity

Next Lesson

Bayes‘ Rule

- Introduction
- Applying Bayes‘ rule to data science
- Naive Bayes
- Advantages and Disadvantages of Naive Bayes
- Example

Similarity and Distance

- Nearest Neighbors
- Example

Ref.

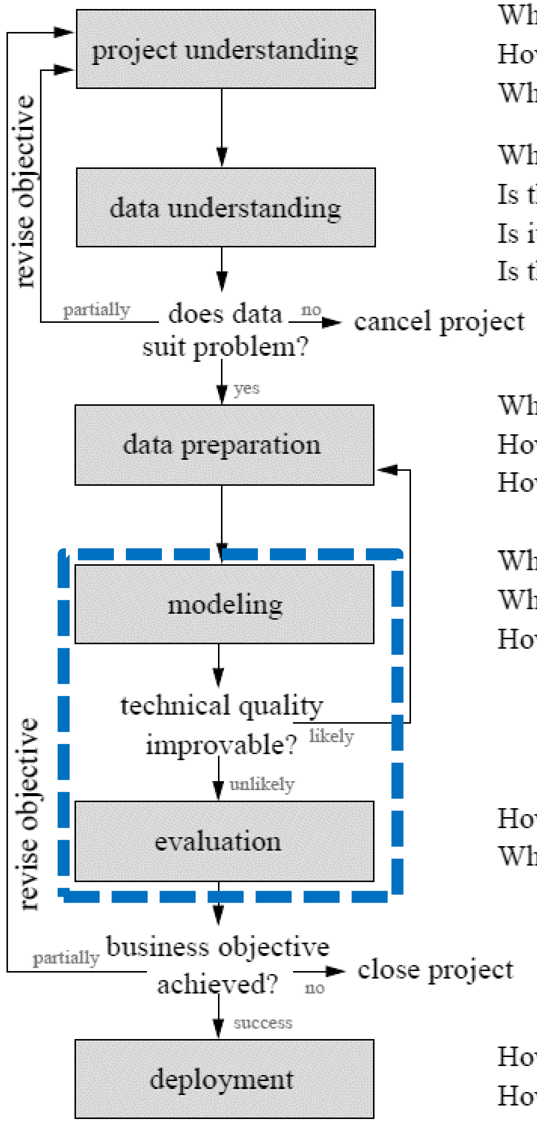
CRISP-DM

Cross
Industry
Standard
Process for
Data
Mining

Iteration as
a rule

Process of data
exploration

Implementation of the
KDD Process



What exactly is the problem, the expected benefit?
How would a solution look like?
What is known about the domain?

What data do we have available?
Is the data relevant to the problem?
Is it valid? Does it reflect our expectations?
Is the data quality, quantity, recency sufficient?

Which data should we concentrate on?
How is the data best transformed for modeling?
How may we increase the data quality?

What kind of model architecture suits the problem best?
What is the best technique/method to get the model?
How good does the model perform technically?

How good is the model in terms of project requirements?
What have we learned from the project?

How is the model best deployed?
How do we know that the model is still valid?

Ref. Wirth / Hipp (2000), Azevedo (2008)

Introductory example

Modeling techniques so far: Discriminative
What is the best way to distinguish target values?

Now: *Generative*
Which class most likely generated this example?

Analyze data instances as **evidence** for or against different values of the target

Idea:
use historical data to **estimate both the direction and the strength of the evidence**

Combine the evidence to estimate the resulting likelihood of class membership

Example:
Target online displays to consumers based on webpages they have visited in the past

Run a targeted campaign for, e.g., a luxury hotel
Target variable: *will the consumer book a hotel room within one week after having seen the advertisement?*
Cookies allow for observing which consumers book rooms

A consumer is characterized by the set of websites we have observed her to have visited before (using cookies)

We assume that some of these websites are more likely to be visited by good prospects for the luxury hotel

Problem:
we do not have the resources to estimate the evidence potential for each site manually

Another example: a similar problem is spam detection

Ref.

Combining evidence probabilistically

What is the probability $p(C)$ that if you show an ad to any customer, she will book a room given some evidence E (such as the websites visited by a *particular* customer)?

→ $p(C|E)$

Problem:
for any particular collection of evidence E,
we may not have seen enough cases / seen it at all!

Idea:
Consider the different pieces of evidence separately, and then combine evidence

Ref.

Reminder: statistical (in)dependence

If the events A and B are statistically independent ($p(B) = p(B|A)$), then we can compute the probability that both A and B occur as $p(AB) = p(A) * p(B)$.

Example: rolling a fair dice

The general formular for combining probabilities that takes care of dependencies between events is $p(AB) = p(A) * p(B|A)$

-> Given that you know A, what is the probability of B

Remember:
Alternative Expected profit computation

$$EP = \boldsymbol{p(Y,\mathrm{p})} \cdot b(Y,\mathrm{p}) + p(N,\mathrm{p}) \cdot c(N,\mathrm{p}) + p(N,\mathrm{n}) \cdot b(N,\mathrm{n}) + p(Y,\mathrm{n}) \cdot c(Y,\mathrm{n})$$

$$EP = \boldsymbol{p(Y|\mathrm{p}) \cdot p(\mathrm{p})} \cdot b(Y,\mathrm{p}) + p(N|\mathrm{p}) \cdot p(\mathrm{p}) \cdot c(N,\mathrm{p}) + p(N|\mathrm{n}) \cdot p(\mathrm{n}) \cdot b(N,\mathrm{n}) + p(Y|\mathrm{n}) \cdot p(\mathrm{n}) \cdot c(Y,\mathrm{n})$$

$$\dots$$

Ref.

Bayes‘ rule

Note that in $p(AB) = p(A)p(B|A)$ the order of A and B is rather arbitrary, one could also write $p(AB) = p(B)p(A|B)$

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

Let B be some **hypothesis H** that we are interested in assessing the likelihood of, and A some evidence E that we have observed:

$$\boxed{p(H|E) = \frac{p(E|H)p(H)}{p(E)}}$$

(**Bayes’ rule**)

Bayes‘ rule says that we can compute the probability of our hypothesis H given some evidence E by instead looking at the probability of the evidence given the hypothesis as well as the unconditional probability of the hypothesis and the evidence.

Example: medical diagnosis

Hypothesis $H = \text{measles}$, Evidence $E = red\ spots$

In order to directly estimate $p(measles|red\ spots)$, we would need to think through all the different reasons a person might exhibit red spots and what proportion of them would be measles.

Instead: $p(E|H)$ is the prob. that one has red spots given that one has measles. $p(H)$ is simply the prob. that someone has measles, and $p(E)$ that someone has red spots. This is by far easier to assess.

Image: Thomas Bayes (Wikimedia)

Exercise - Bayes' rule

Assume that the probability of having cancer is 0.05—meaning that 5% of people have cancer.
Now, assume that the probability of being a smoker is 0.10—meaning that 10% of people are smokers.
and assume that 20% of people with cancer are smokers.

Calculate the Likelihood of a person to have cancer, given the observation of being a smoker.

$$p(H|E) = \frac{p(E|H)p(H)}{p(E)}$$

2 Min.

Ref. https://towardsdatascience.com/what-is-bayes-rule-bb6598d8a2fd

Applying Bayes‘ rule to data science

A lot of DM methods are based on Bayes‘ rule

Bayes‘ rule for classification of the probability that the target variable C takes on the class of interest c after taking the evidence E (feature values) into account:

$$p(C = c|E) = \frac{p(E|C = c)p(C = c)}{p(E)}$$

- $p(C = c)$ is the „prior“ probability of the class, i.e., the probability we would assign to the class before seeing any evidence [e.g., the prevalence of c in the population = percentage of all examples that are of class c]
- $p(E|C = c)$ is the likelihood of seeing the evidence E [the percentage of examples of class c that have E]
- $p(E)$ is the likelihood of the evidence [occurence of E]

- Estimating these values, we could use $p(C = c|E)$ as an estimate of class probability
- Alternatively, we could use the values as a score to rank instances

Drawback:
if E is a usual vector of attribute values,
we would require knowing the full joint probability of the example

This is difficult to measure

We may never see a specific example in the training data that matches a given E in our test data

➡ „Naive“ Bayes:
Make a particular **assumption of independence**!

Naive Bayes (1/2)

Conditional independence:
use the class of the example as condition

This allows for easy combination of probabilities:

$$p(AB|C) = p(A|C) \cdot p(B|C)$$

In other words: we assume that the attributes are conditionally independent and ignore its order, i.e.

$$p(E|c) = p(e_1|c) \cdot p(e_2|c) \cdot \dots \cdot p(e_k|c)$$

Each of the $p(e_i|c)$ terms can be computed directly from the data (count up the prop. we see e_i in c)

$$p(c|E) = \frac{p(e_1|c) \cdot p(e_2|c) \cdot \dots \cdot p(e_k|c) \cdot p(c)}{p(E)}$$ (Naïve Bayes)

Bayes' rule for classification

$$p(c|E) = \frac{p(E|c)p(c)}{p(E)}$$

Ref.

Naive Bayes (2/2)

$$p(c|E) = \frac{p(e_1|c) \cdot p(e_2|c) \cdot \ldots \cdot p(e_k|c) \cdot p(c)}{p(E)}$$ (Naïve Bayes)

Naive Bayes classifies a new example by estimating the probability that the example belongs to each class and reports the class with highest probability

Note that the denominator $p(E)$ never actually has to be calculated

- We can focus on the numerator for comparison of different classes c, because the denominator is always the same
- If we need probability estimates, the probabilities will add up to one, so we can derive it from the other quantities

$$p(c_0|E) = \frac{p(e_1|c_0) \cdot p(e_2|c_0) \cdot \ldots \cdot p(e_k|c_0) \cdot p(c_0)}{p(e_1|c_0) \cdot p(e_2|c_0) \cdot \ldots \cdot p(e_k|c_0) \cdot p(c_0) + p(e_1|c_1) \cdot p(e_2|c_1) \cdot \ldots \cdot p(e_k|c_1) \cdot p(c_1)}$$

Ref.

Naive Bayes classifier

Example

How does a naïve Bayes classifier classify the object $\boxed{(t, l, y)}$?

We need to calculate

$p(m/E) = ?$ $L(\text{Sex} = m | \text{Height} = t, \text{Weight} = l, \text{long\_hair} = y)$

$$= P(\text{Height} = t|\text{Sex} = m) \cdot P(\text{Weight} = l|\text{Sex} = m) \cdot P(\text{long\_hair} = y|\text{Sex} = m) \cdot P(\text{Sex} = m)$$

and

$p(f/E) = ?$ $L(\text{Sex} = f | \text{Height} = t, \text{Weight} = l, \text{Long\_hair} = y)$

$$= P(\text{Height} = t|\text{Sex} = f) \cdot P(\text{Weight} = l|\text{Sex} = f) \cdot P(\text{Long\_hair} = y|\text{Sex} = f) \cdot P(\text{Sex} = f).$$

Available data

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

We do not need to calculate p(E), i.e., p(t,l,y). Why?

Ref.

Naive Bayes classifier

Example – Male?

$P(\text{Height} = t | \text{Sex} = m)$?

$P(\text{Height} = t | \text{Sex} = m)$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

$P(\text{Height} = t | \text{Sex} = m) = 2/4 = 1/2$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

calculate P(Weight...) and P(Long_hair...) in the same way

$P(\text{Weight} = l | \text{Sex} = m) = 0/4 = 0$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

$P(\text{Long\_hair} = y | \text{Sex} = m) = 0/4 = 0$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

$P(\text{Sex} = m) = 4/10 = 2/5$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

$$L(\text{Sex} = m | \text{Height} = t, \text{Weight} = l, \text{Long\_hair} = y) = \frac{1}{2} \cdot 0 \cdot 0 \cdot \frac{2}{5} = 0$$

Nenner nicht berücksichtigt, da Vergleich zweier Klassen

Naive Bayes classifier

Example – Female?

$P(\text{Height} = t | \text{Sex} = f)$?

$P(\text{Height} = t | \text{Sex} = f)$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

$P(\text{Height} = t | \text{Sex} = f) = 1/6$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

calculate P(Weight...) and P(Long_hair...) in the same way

$P(\text{Weight} = l | \text{Sex} = f) = 3/6 = 1/2$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	g	n	n	m

$P(\text{Long\_hair} = y | \text{Sex} = f) = 4/6 = 2/3$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

$P(\text{Sex} = f) = 6/10 = 3/5$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

$$L(\text{Sex} = f | \ \text{Height} = t, \ \text{Weight} = l, \ \text{Long\_hair} = y)$$

$$= \frac{1}{6} \cdot \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{5} = 1/30$$

$$> 0$$

$$= L(\text{Sex} = m | \ \text{Height} = t, \ \text{Weight} = l, \ \text{Long\_hair} = y)$$

Classification of (t, l, y): female (f)

(Dis)Advantages of Naive Bayes

Naive Bayes

- is a simple classifier, although it takes all the feature evidence into account
- is very **efficient** in terms of storage space and computational time
- performs surprisingly well for classification
- is an „**incremental learner**“
- Works also well with unstructured data, such as texts

Note that the independence assumption does not hurt classification performance very much

- To some extent, we double the evidence
- Tends to make the probability estimates **more extreme** in the correct direction

Don‘t use the probability estimates themselves!

- But ranking is fine

Excursus

Lift typically measures the relative increase of likelihood from a model above the baseline

In other words, it measures **how much more prevalent** the positive class is in the selected subpopulation over the prevalence in the population as a whole.

In the context of classification:
Lift is the amount by which a classifier concentrates the positive examples above the negative examples

Recommended for further reading: A Model of Evidence “Lift“, Provost and Fawcett, Chap.9, pp. 244 - 247

Kosinski et al. (2013):Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, doi: 10.1073/pnas.1218772110.

For example, predicting intelligence from Facebook likes ☺

Ref.

Agenda

Two other variants of modeling

Naive Bayes classifier

Evidence and Probabilities

Bayes‘ Rule

- Introduction
- Applying Bayes‘ rule to data science
- Naive Bayes
- Advantages and Disadvantages of Naive Bayes
- Example

k-Nearest Neighbors

Similarity

Next Lesson

Similarity and Distance

- Nearest Neighbors
- Example

Ref.

Similarity and distance
→ Nearest neighbors

Next Lesson

Similarity is at the core of many DM methods

If two things are **similar** in some ways,

they often share other characteristics as well

Image: www.welt.de

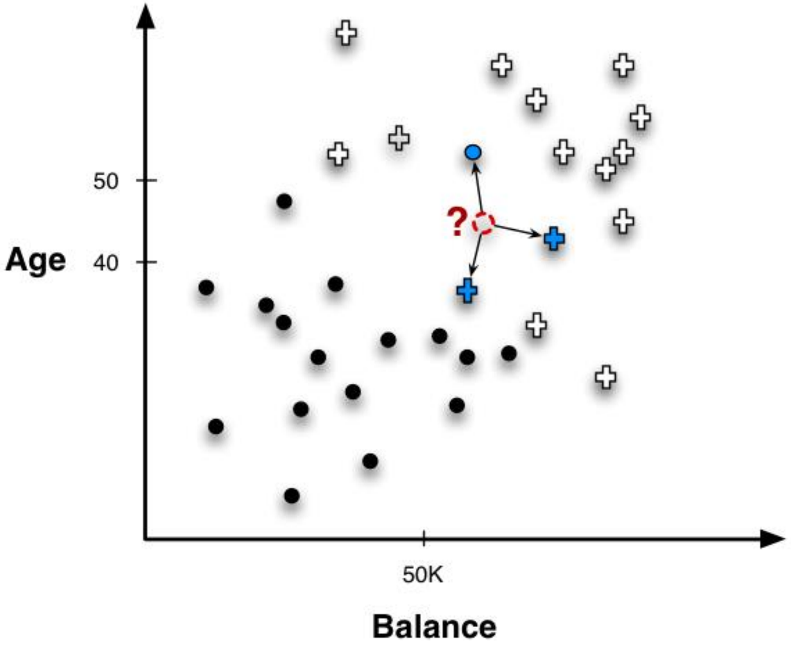
Some business cases

- Use similarity for classification and regression
- Group similar items together into clusters
- Provide recommendations to people (Amazon, Netflix)
- Reasoning from similar cases
(medicine, law)

Ref.

Nearest neighbors for predictive modeling

k-Nearest Neighbors



Look for the nearest neighbors and derive target class for new example

Fragen?

- ✓ Bayes’ Rule
  - ✓ Applying Bayes’ rule to data science
  - ✓ Naïve Bayes
  - ✓ Advantages and Disadvantages of Naïve Bayes

- ○ Similarity and distance
  - ○ Nearest Neighbors

Recommended reading

Evidence and Probabilities

Provost, F., Fawcett, T.	Data Science for Business, Chapter 9
Berthold et al.	Guide to Intelligent Data Analysis, Chapter 8.2

Recommended for further reading: A Model of Evidence "Lift", Provost and Fawcett, Chap.9, pp. 244 – 247

Kosinski et al. (2013): Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences,* doi: 10.1073/pnas.1218772110.

Similarity

Provost, F., Fawcett, T.	Data Science for Business, Chapter 6
Berthold et al.	Guide to Intelligent Data Analysis, Chapter 7, 9.1
Hand, D.	Principles of Data Mining, Chapter 10

Ref.