Dokument Klassifikation

Versuch 4:

Dokument Klassifikation

Agenda:

1: Klassifizierung allgemein

2: der naive Bayes-Klassifizierer

3: Beispiel

4: Probleme

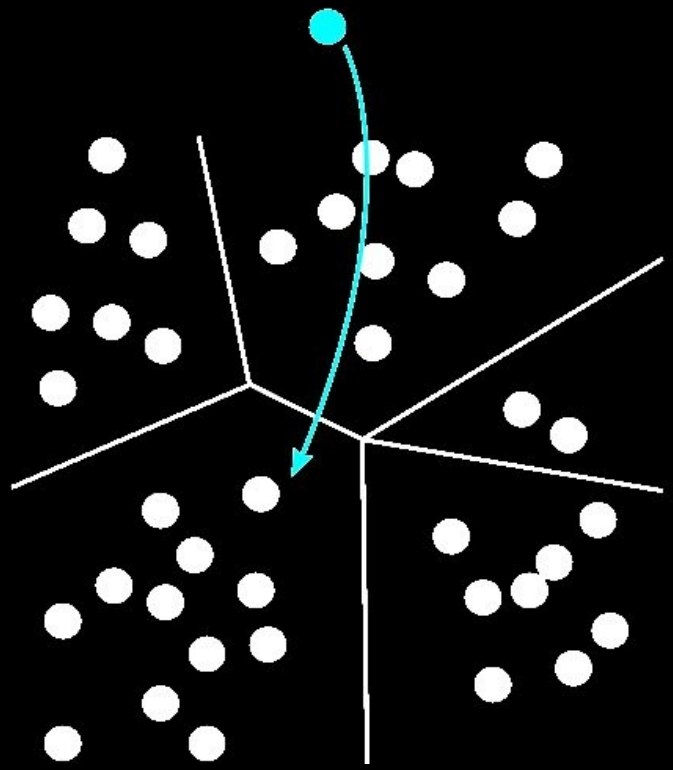
5: Fazit

6: Quellen

Versuch 4:

Dokument Klassifikation

1: Klassifizierung allgemein:



Versuch 4:

Dokument Klassifikation

1: Klassifizierung allgemein:

- Einordnung von Objekten in Kategorien

- Übersichtlichkeit, Systematik und Wissensextraktion (→ Data-Mining)

- **Bsp:**

  Kategorisierung in der **Biologie** (z.B. Tierarten)
  Kategorisierung in der **Geologie** (z.B. Böden, Klimazonen)
  Kategorisierung in der **Informatik** (z.B. Dokumente)

Versuch 4:

Dokument Klassifikation

1: Klassifizierung allgemein:

- ein Klassifizierer sortiert unsere Dokumente in **„Kategorien“**
- z.B. E-Mails in **„Spam“** oder **„nicht Spam“**
- auf großen Datenmengen möglich (Data-Mining)

Versuch 4:

Dokument Klassifikation

1: Klassifizierung allgemein:

- automatische Verfahren anhand von **Entscheidungsregeln**

Möglich: statische Verfahren

„Absender X ist *immer* Spam!“
„Absender Y ist *niemals* Spam!“

Besser: Entscheidungsregeln werden **„gelernt“** (→ KI)

Neuronalenetze (*siehe Vergangene Woche*)
Clustering (*siehe 2. Teil*)
überwachtes Lernen

Versuch 4:

Dokument Klassifikation

2: der naive Bayes-Klassifizierer:

- **Idee:**

1. Dokumente werden klassifiziert Übergeben

2. Lerne für jedes Wort die Wahrscheinlichkeit Spam zu sein

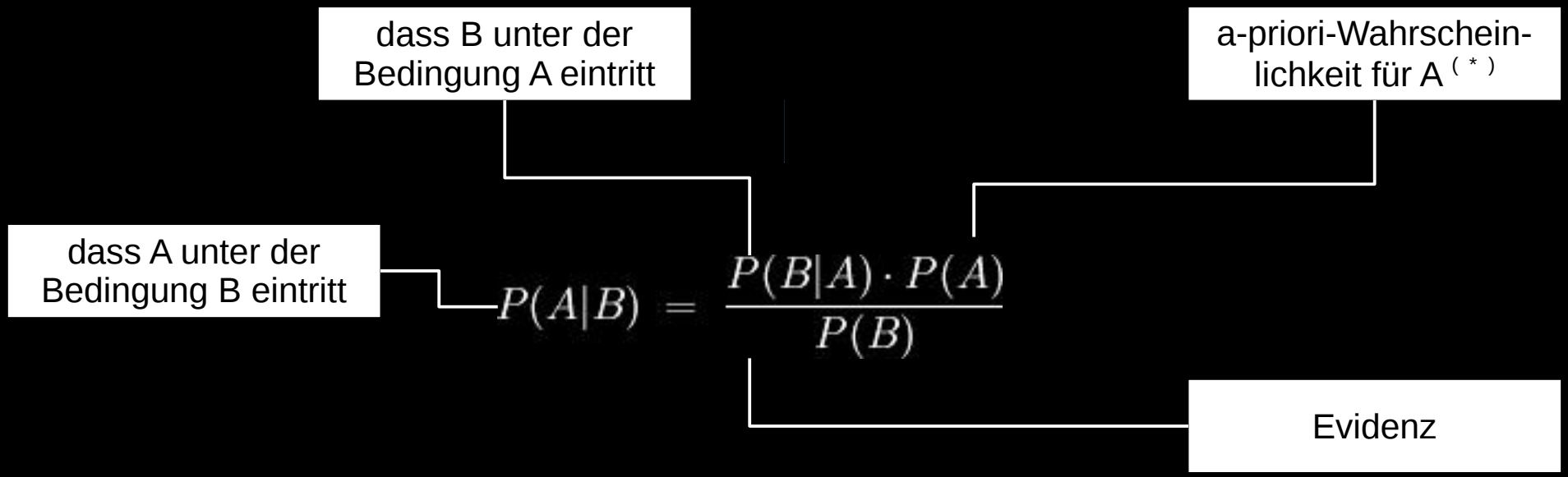
3. Lerne für ein Dokument die Wahrscheinlichkeit Spam zu sein

4. Ordne neue Dokumente in Kategorie mit max. Wahrscheinlichkeit

Versuch 4:
Dokument Klassifikation

2: der naive Bayes-Klassifizierer:

- das **Bayes-Theorem**



(*) Wahrscheinlichkeit auf Grund von Vorwissen,

z.B. beim Würfeln jede Seite P = 1/6.

Versuch 4:

Dokument Klassifikation

2: der naive Bayes-Klassifizierer:

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

A = die Klasse (Spam, nicht-Spam)
B = das Attribut (Wort)

somit ist:

P(A) = Wahrscheinlichkeit, dass diese Klasse auftritt (z.B. Spam)

P(B) = Wahrscheinlichkeit, dass dieses Attribut auftritt (z.B. Wort)

P(B|A) = Wahrscheinlichkeit, dass Attribut in Klasse fällt (z.B. Wort ist Spam)

P(A|B) = Wahrscheinlichkeit, einer best. Klasse für dieses Attribut

Versuch 4:

Dokument Klassifikation

2: der naive Bayes-Klassifizierer:

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

A = die Klasse (Spam, nicht-Spam)
B = das Attribut (Wort)

Im Falle der „E-Mail“, *ist:*

- B ein Vektor einzelner Worte
- P(B|A) ist das Produkt aus den einzelnen Wahrscheinlichkeiten:

$$p(\mathrm{B} \mid A) = p(\langle b_1 \ldots b_n \rangle \mid A_j) = \prod_{i=1}^{n} p(b_i \mid A)$$

Versuch 4:

Dokument Klassifikation

2: der naive Bayes-Klassifizierer:

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

A = die Klasse (Spam, nicht-Spam)
B = das Attribut (Wort)

Im Falle der „E-Mail“, *ist:*

- A die Zahl der Mails einer Kategorie dividiert durch die Gesamtzahl der Mails

Versuch 4:

Dokument Klassifikation

2: der naive Bayes-Klassifizierer:

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

A = die Klasse (Spam, nicht-Spam)
B = das Attribut (Wort)

Im Falle der „E-Mail“, *ist:*

- B ein Vektor von Worten

- P(B) das Produkt der einzelnen Wahrscheinlichkeiten, dass ein Wort auftritt (Anzahl des Wortes in der Mail / Anzahl aller Wörter in der Mail)

- ist in allen Kategorien gleich und muss nicht berechnet werden

Versuch 4:
Dokument Klassifikation

3: Beispiel

- wir haben gelernt:

Mail 1: gut, toll, super, nicht	→ **nicht-Spam**
Mail 2: blöd, hammer	→ **nicht-Spam**
Mail 3: gut, ok, klar, schlecht	→ **nicht-Spam**
Mail 4: nicht, blöd, schlecht	→ **Spam**
Mail 5: nicht, kaputt	→ **Spam**
Mail 6: aber, schlecht	→ **Spam**

- es kommt diese neue Mail:

Mail 7: nicht, gut, schlecht

Versuch 4:
Dokument Klassifikation

3: Beispiel

Wie wahrscheinlich fallen die Wörter in (Spam, nicht-Spam)?

nicht: **P(** nicht|Spam **)** = 2/3
P(nicht|nicht-S **)** = 1/3

gut: **P(** gut|Spam **)** = 0
P(gut|nicht-S **)** = 2/3

schl.: **P(** schlecht|Spam **)** = 2/3
P(schlecht|nicht-S **)** = 1/3

Problem: P(gut|Spam) = 0!

Es gibt 3 Spam-Mails. In 2 davon kommt „nicht“ vor.

Mail 1: gut, toll, super, nicht	→ **nicht-Spam**
Mail 2: blöd, hammer	→ **nicht-Spam**
Mail 3: gut, ok, klar, schlecht	→ **nicht-Spam**
Mail 4: nicht, blöd, schlecht	→ **Spam**
Mail 5: nicht, kaputt	→ **Spam**
Mail 6: aber, schlecht	→ **Spam**

Mail 7: nicht, gut, schlecht

Versuch 4:

Dokument Klassifikation

3: Beispiel

Lösung: Gewichtete Wahrscheinlichkeiten

→ P_G(Wort | Kategorie) = **(** 0,5 + P(Wort|Kategorie) * C_z **)** / **(** (1 + C_z) **)**

- kommt das Wort bisher nicht vor, so ist die Zugehörigkeit unentschieden
- je mehr E-Mails es mit diesem Wort gibt, mehr fließt es ein
- C_z zählt wie oft das Wort bisher auftrat

Versuch 4:
Dokument Klassifikation

3: Beispiel

nicht: $P_G($ nicht|Spam $) = (0{,}5 + 2/3 * 3)/(3 + 1) = 0{,}625$

$P_G($ nicht|nicht-S $) = 0{,}375$

gut: $P_G($ gut|Spam $) = 0{,}167$ (**!**)

$P_G($ gut|nicht-S $) = 0{,}61$

schl.: $P_G($ schlecht|Spam $) = 0{,}625$

$P_G($ schlecht|nicht-S $) = 0{,}375$

Mail 1: gut, toll, super, nicht	→ **nicht-Spam**
Mail 2: blöd, hammer	→ **nicht-Spam**
Mail 3: gut, ok, klar, schlecht	→ **nicht-Spam**
Mail 4: nicht, blöd, schlecht	→ **Spam**
Mail 5: nicht, kaputt	→ **Spam**
Mail 6: aber, schlecht	→ **Spam**

Mail 7: nicht, gut, schlecht

Versuch 4:
Dokument Klassifikation

3: Beispiel

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

P$_G$**(** nicht|Spam **)** = 0,625
P$_G$**(** nicht|nicht-S **)** = 0,375
P$_G$**(** gut|Spam **)** = 0,167 (**!**)
P$_G$**(** gut|nicht-S **)** = 0,61
P$_G$**(** schlecht|Spam **)** = 0,625
P$_G$**(** schlecht|nicht-S **)** = 0,375

P(Spam | „nicht, gut, schlecht“ **)** **=** 0,625 * 0,167 * 0,625 * 0,5 **=** **0,033**
P(nicht-S | „nicht, gut, schlecht“ **)** **=** 0,375 * 0,61 * 0,375 * 0,5 **=** **0,043**

→ die **Evidenz** P(B) kann weggelassen werden, da sie stets gleich ist!

Versuch 4:
Dokument Klassifikation

4: Probleme

- Kategorisierung kann eigentlich nur über Inhalt und Kontext erfolgen
- wir treffen eine Annahme (die höchstens teilweise stimmt)
- Worte sind nicht unabhängig zueinander (→ naive Annahme)
- Kaltstartproblem (Lösung: Gewichtung)

Versuch 4:
Dokument Klassifikation

5: Fazit

- keine exakten Wahrscheinlichkeiten
- **Aber:** die brauchen wir auch nicht, es reicht die stärkste Kategorie!
- (relativ) leicht zu implementieren
- schnelle Berechnung
- liefert in der Praxis gute Ergebnisse

Versuch 4:
Dokument Klassifikation

6: Quellen

- Dr. Johannes Maucher: **Dokument Klassifikation Skript**. 2010.

- Tobias Hetzel, Roberto Piccolantonio: **Präsentation Dokumentklassifizierung**.

- Wikipedia: **http://de.wikipedia.org/wiki/Bayes-Klassifikator**

- Wolfgang Ertel**: Einführung in die KI**. Vieweg + Teubner Verlag 2009.