Human Centered Multimedia
Institute of Computer Science

Universität Augsburg University

Multimodale Analyse

Wiederholung von Klassifikation

Elisabeth André
Stephan Hammer
Chi-Tai Dang

HCM

Human Centered Multimedia
Institute of Computer Science
Augsburg University
Universitätsstraße. 6a
86159 Augsburg, Germany

- Bisher haben wir erfahren, wie man eine Vielzahl von **Merkmalen** bei den unterschiedlichen Daten **berechet**.
- Außerdem haben wir gelernt, wie man aus dieser Menge die geeigneten **Merkmale** zur Klassifikation **auswählt**.

- Wir werden nun Verfahren kennenlernen die genutzt werden um **Daten** in bestimmte **Klassen einzuordnen**
- Innerhalb dieser Klassen befinden sich dann diejenigen Daten, deren Merkmale sich mehr oder weniger ähneln.

  → Diesen Vorgang nennt man **Klassifikation**

Multimodale Analyse
Klassifikation

- Anhand von Beispielen erlernt das System eine Zuordnung von Objekten in vorhandene **Klassen**

Inputs → Lerner → Outputs

Beispiele Datenpaare

 ↔ **Merkel**

 ↔ **Merkel**

 ↔ **Kretschmann**

 ↔ **Kretschmann**

...

 **?**

- **Was kann Klassifikation?**
  - Automatisches Sortieren von Äpfeln und Birnen in Güteklassen
  - Einteilung von Kunden in Kundengruppen (z.B. *„Top-Kunden“*, *„Normal-Kunden“, „Problemfälle“* und *„potentielle Wechsler“*)
  - Unterscheidung von Museumsbesuchern anhand ihrer Navigationsmuster (*„Schmetterling“*, *„Grashüpfer“*, etc.)
  - Automatische Handschrift- oder Zeichenerkennung: Varianten eines Buchstabens werden der gleichen Klasse zugeordnet
  - Klassifikation von Gesichtern anhand von biometrischen Merkmalen oder sogar Identifikation von Gesichtern (*„FBI“*)
  - Erkennung von Emotionsklassen anhand von Sprache, Gesten, Körperhaltung, Gesichtsmimik und biosensorischen Signalen.

Sensoreingabe

Mustererkennung

Feature Extraktion

Klassifikator

Vorhersage

Ärger

Freude

Human Centered Multimedia
Institute of Computer Science

Universität Augsburg University

Multimodale Analyse

Grundlagen der Wahrscheinlichkeitstheorie

Elisabeth André
Stephan Hammer
Chi-Tai Dang

Human Centered Multimedia
Institute of Computer Science
Augsburg University
Universitätsstraße. 6a
86159 Augsburg, Germany

- Eine **Zufallsgröße** ist eine Größe, deren Wert wir nicht exakt kennen bzw. vorhersagen können. Wir können den möglichen Werten nur bestimmte **Wahrscheinlichkeiten** für ihr Auftreten zuordnen.

- Zufallsgrößen sind bei uns **Merkmale** und **Klassen**

- Der Fall, dass eine Zufallsgröße einen bestimmten Wert annimmt nennen wir **Ereignis** oder auch **Beobachtung.**

- Die **Wahrscheinlichkeit** eines Ereignisses A wird bezeichnet als *P(A)* (*P* steht dabei für **Probability**).

- Wahrscheinlichkeiten erfüllen die folgenden **Axiome**:
  - Ist *A* ein Ereignis, dann gilt für die Wahrscheinlichkeit von *A*:

$$0 \le P(A) \le 1$$

  - Ist *A* ein Ereignis, dann gilt für das Gegenereignis von *A*:

$$P(\overline{A}) = 1 - P(A)$$

  - Sind *A* und *B* Ereignisse, die nicht gleichzeitig eintreten können:

$$P(A \cup B) = P(A) + P(B)$$

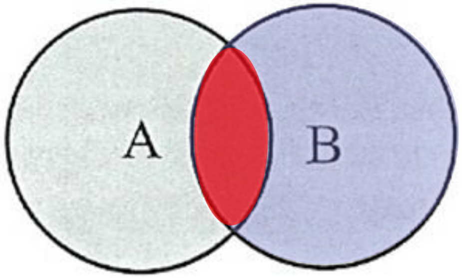
  - Sind *A* und *B* Ereignisse, die auch gleichzeitig eintreten können:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Angenommen wir wollen die Wahrscheinlichkeit für ein Ereignis *A* bestimmen unter der Bedingung, dass wir schon wissen dass ein anderes Ereignis *B* eintritt.

- Die **bedingte Wahrscheinlichkeit** eines Ereignisses *A* gegeben, dass ein Ereigniss *B* gilt, wird berechnet durch:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad wenn\ P(B) > 0$$



Universität Augsburg University

- Durch Umformung der Definitionsformel der bedingten Wahrscheinlichkeit entsteht der **Multiplikationssatz**:

$$P(A \cap B) = P(A | B) P(B)$$

- Die Verallgemeinerung des **Multiplikationssatz** lautet:

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P\left(A_i \,\middle|\, \bigcap_{j=1}^{i-1} A_j\right)$$

- Vgl. Kettenregel:

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1) P(A_2 | A_1) P(A_3 | A_1 \cap A_2) \ldots P\left(A_n \,\middle|\, \bigcap_{i=1}^{n-1} A_i\right)$$

- Kennt man nur die bedingten Wahrscheinlichkeiten von einem Ereignis *A*, sowie die Wahrscheinlichkeiten der bedingenden Ereignisse so kann man die sogenannte **totale Wahrscheinlichkeit** berechnen:

$$P(A) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B})$$

- Verallgemeinert lautet die **totale Wahrscheinlichkeit**:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

wenn die Ereignisse $B_1, B_2, \ldots, B_n$ eine Partition des Wahrscheinlichkeitsraums bilden und $P(B_i) > 0$ gilt.

- Der **Satz von Bayes** ergibt sich direkt aus der Definition der bedingten Wahrscheinlichkeit und der Definition des Multiplikationssatzes durch einfache Umformung:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}, \quad \text{wenn } P(A), P(B) > 0$$

- Der Nenner *P(B)* dieser Formel kann dabei mit Hilfe des Satzes der totalen Wahrscheinlichkeit berechnet werden.

- Die Allgemeine Form des **Satz von Bayes** lautet dann:

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)}$$

- Wenn zwei Ereignisse *A* und *B* **unabhängig** sind, dann soll die totale Wahrscheinlichkeit von *A* niemals davon beeinflusst werden, ob wir schon wissen, dass *B* eintritt:

$$P(A) = P(A \mid B)$$

- Daraus ergibt sich dann für die Wahrscheinlichkeit dass beide Ereignisse eintreten die folgende Umformung:

$$P(A) = P(A \mid B) = \frac{P(A \cap B)}{P(B)} \Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$$

Human Centered Multimedia
Institute of Computer Science

Universität Augsburg University

Multimodale Analyse

Bayessche Statistik und Naive Bayes Klassifikation

Elisabeth André
Stephan Hammer
Chi-Tai Dang

HCM

Human Centered Multimedia
Institute of Computer Science
Augsburg University
Universitätsstraße. 6a
86159 Augsburg, Germany

- In der **Bayesschen Statistik** werden die Regeln für die Klassifikation von Beobachtungen mit dem **Satz von Bayes** als bedingte Wahrscheinlichkeiten formuliert.

- Seien $h_1, h_2, \ldots, h_n \in H$ Hypothesen die nicht gleichzeitig auftreten können, von denen aber immer eine eintritt, dann kann man folgende Fragestellungen beantworten

Was ist die **wahrscheinlichste Hypothese** $h_i \in H$, also die **wahrscheinlichste Klassifikation** einer neuen Dateninstanz bei bekannten Trainingsdaten?

Die Bayessche Statistik
Wichtige Denotationen

- $P(h_i)$ sei die Anfangswahrscheinlichkeit dafür, dass die Hypothese h_i erfüllt ist (**A-Priori Probability** von h_i).
- $P(D)$ sei die Anfangswahrscheinlichkeit dafür, dass die Daten D beobachtet werden.
- $P(D|h_i)$ sei die Wahrscheinlichkeit dafür, dass man D beobachtet, sofern die Hypothese h_i erfüllt ist.
- $P(h_i|D)$ sei die Wahrscheinlichkeit dafür, dass h_i bei den Daten D gilt (**A-Posteriori Probability** von h_i).

- Die **A-Priori Wahrscheinlichkeit** für ein Ereignis *A* beschreibt die Wahrscheinlichkeit für *A*, wenn keine Informationen über andere Ereignisse vorliegen.

- In der Bayesschen Statistik modellieren A-Priori Wahrscheinlichkeiten unser **Vorwissen** über die Häufigkeit einer Klasse und das Auftreten von Merkmalen.
  - $P(Klasse=apfel) = 0.25$ (d.h. 25% der Objekte sind Äpfel)
  - $P(Klasse=birne) = 0.40$ (d.h. 40% der Objekte sind Birnen)
  - $P(Klasse=orange) = 0.35$ (d.h. 35% der Objekte sind Orangen)

Die Bayessche Statistik
Die A-Posteriori Probability

- Die **A-Posteriori-Wahrscheinlichkeit** für ein Ereignis *A* beschreibt die Wahrscheinlichkeit für *A*, wenn wir auch zusätzlich wissen, dass ein bestimmtes Ereignis eintritt.

- In der Bayesschen Statistik modellieren damit dann die **Zusammenhänge** zwischen Klassen und Merkmalen:
  - *P (Form=rund | Klasse=orange) = 1.0*

    *(*d.h. wenn ich weiß, dass ich eine Orange betrachte, dann ist die Wahrscheinlichkeit dafür, dass sie rund ist 100%)

- Man betrachtet eine Menge *H* von Hypothesen und ist daran interessiert, die wahrscheinlichste Hypothese *h* aus *H* für die gegebenen Daten *D* zu finden.
- Die **Maximum A-Posteriori Hypothese** ist

$$h_{MAP} = \underset{h \in H}{\arg\max}\, P(h \mid D) =$$

$$\underset{h \in H}{\arg\max} \frac{P(D \mid h) P(h)}{P(D)} =$$ *(Mit dem Satz von Bayes)*

$$\frac{1}{P(D)} \underset{h \in H}{\arg\max}\, P(D \mid h) P(h)$$ *(Herausziehen der Konstante 1/P(D), die von h unabhängig ist)*

- Wenn kein Vorwissen über die Wahrscheinlichkeiten der einzelnen Hypothesen besteht, macht man die spezielle Annahme, dass jede Hypothese aus *H* die gleiche Anfangswahrscheinlichkeit hat, d.h.

$$P(h_i) = P(h_j) \quad \forall h_i, h_j \in H$$

- Eine Hypothese, die $P(D|h)$ maximiert, wird dann die **Maximum Likelihood Hypothesis** h_{ML} genannt.

$$h_{ML} = \underset{h \in H}{\arg\max}\, P(D \mid h)$$

- Bei der **Naiven Bayes Klassifikation** fragen wir nach den Wahrscheinlichkeiten der Hypothesen, wenn wir wissen, dass ein bestimmtes Ereignis eingetroffen ist.

- **Gegeben**:
  - Eine endliche Menge von **Klassen** $V = \{v_1, v_2, ..., v_m\}$
  - Eine Menge von klassifizierten **Trainingsbeispielen**
  - Eine **Instanz** als ein Tupel $(a_1, a_2, ..., a_n)$ von **Attributen**
- **Gesucht**:
  - Die wahrscheinlichste Klasse für die neue Instanz

- Die wahrscheinlichste Klasse v_{MAP} wird berechnet durch:

$$v_{MAP} = \underset{v_j \in V}{\arg\max}\, P(v_j \mid a_1, a_2 \ldots a_n)$$

- Anwendung des Satz von Bayes ergibt diesen Ausdruck:

$$v_{MAP} = \underset{v_j \in V}{\arg\max}\, \frac{P(a_1, a_2 \ldots a_n \mid v_j) P(v_j)}{P(a_1, a_2 \ldots a_n)}$$

- Herausziehen der Normalisierungskonstante ergibt:

$$v_{MAP} = \underset{v_j \in V}{\arg\max}\, P(a_1, a_2 \ldots a_n \mid v_j) P(v_j)$$

- Wir schätzen die Terme $P(a_1, a_2 \ldots a_n \mid v_j)$ und $P(v_j)$ auf Basis der bekannten Trainingsdaten durch **Abzählen**.

- Abschätzung von $P(v_j)$ durch Bestimmung der relativen **Häufigkeit**, der Klasse v_j innerhalb der Trainingsdaten.

- Abschätzung von $P(a_1, a_2 \ldots a_n \mid v_j)$ ist **problematisch**
  - Bei wenig Trainingsdaten ist es nicht ungewöhnlich, dass wir keine solche Kombinationen haben, d.h. wir würden für diesen Ausdruck dann lediglich *0* erhalten.
  - Gute Abschätzung nur bei **riesiger Trainingsmenge**.

- Zur Vereinfachung nimmt man an, dass alle $a_1, a_2 \ldots a_n$ **bedingt unabhängig** voneinander sind (**naiver Ansatz**)

$$P(a_1, a_2 \ldots a_n \mid v_j) = \prod_{i=1}^{n} P(a_i \mid v_j)$$

- Einsetzen ergibt dann den **Naiven Bayes Klassifikator**

$$v_{NB} = \underset{v_j \in V}{\arg\max}\, P(v_j) \prod_{i=1}^{n} P(a_i \mid v_j)$$

- Die **A-Posteriori-Wahrscheinlichkeit** für die Klasse v_{MAP} mit der höchsten Wahrscheinlichkeit wird gegeben durch

$$P(v_{MAP} \mid a_1, a_2 \ldots a_n) = \frac{P(a_1, a_2 \ldots a_n \mid v_{MAP}) P(v_{MAP})}{\sum_{i=1}^{m} P(a_1, a_2 \ldots a_n \mid v_i) P(v_i)}$$

- Die **wichtigsten Formeln** in der Bayeschen Statistik:

  - **Naive Bayes Klassifikator:**

$$v_{NB} = \underset{v_j \in V}{\arg\max}\, P(v_j) \prod_{i=1}^{n} P(a_i \mid v_j)$$

  - **A-Posteriori- Wahrscheinlichkeit:**

$$P(v_{MAP} \mid a_1, a_2 \ldots a_n) = \frac{P(a_1, a_2 \ldots a_n \mid v_{MAP}) P\left(v_{MAP}\right)}{\sum_{i=1}^{m} P(a_1, a_2 \ldots a_n \mid v_i) P(v_i)}$$

- Das Schätzen der Wahrscheinlichkeit $P(v_j)$ geschieht durch Abzählen der relativen Häufigkeiten der Klassen innerhalb der vorhandenen Menge von Trainingsdaten.

- **Beispiel**:
  - *P(Klasse=A)=4/5*
  - *P(Klasse=B)=1/5*

ID	***Form***	***Farbe***	***Klasse***
1	*rund*	*blau*	*A*
2	*rund*	*grün*	*A*
3	*rund*	*gelb*	*A*
4	*eckig*	*grün*	*A*
5	*oval*	*weiß*	*B*

- Die bedingten Wahrscheinlichkeiten $P(a_i \mid v_j)$ werden als relative Häufigkeiten durch folgende Formel abgeschätzt

$$P(a_i \mid v_j) = \frac{n_{i,j}}{n_j}$$

- Wobei n_j die Anzahl der Trainingsbeispiele für v_j ist und $n_{i,j}$ die Anzahl der Instanzen mit Attribut a_i und Klasse v_j.

- **Beispiel**:
  - *P(Form=rund | Klasse=A)=3/4*
  - *P(Farbe=grün | Klasse=A)=2/4*
  - *P(Form=oval | Klasse=A)=0/4*

ID	*Form*	*Farbe*	*Klasse*
1	*rund*	*blau*	*A*
2	*rund*	*grün*	*A*
3	*rund*	*gelb*	*A*
4	*eckig*	*grün*	*A*
5	*oval*	*weiß*	*B*

Naive Bayes Klassifikation
Wahrscheinlichkeit Abschätzen

- Bei kontinuierlichen metrischen Attributen:
  - Abschätzung durch eine ***Diskrete Approximation*** durch diskrete Intervalle:
    - *P(9.0 < Durchmesser ≤ 9.5 | Orange)=10%*
    - *P(9.5 < Durchmesser ≤ 10.0 | Orange)=30%*
    - *P(10.0 < Durchmesser ≤ 10.5 | Orange)=30%*
    - *P(10.5 < Durchmesser ≤ 11.0 | Orange)=10%*
    - *P(11.0 < Durchmesser | Orange)=5%*
  - Abschätzung durch **Wahrscheinlichkeitsdichtefunktionen** (meist Berechnung nach Normalverteilung):

$$P(x \mid C) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad \text{wobei} \quad \mu = \frac{\sum_{x \in TR} x}{|TR|} \quad \text{und} \quad \sigma = \sqrt{\frac{\sum_{x \in TR} (x-\mu)^2}{|TR|}}$$

Naive Bayes Klassifikation
Beispiel des Tennisspielers

- Die Instanzen, welche betrachtet werden sind Tage.
- Jeder Tag besteht aus einem Tupel der Attribute
  - *Outlook*
  - *Temperature*
  - *Humidity*
  - *Wind*

- Ein neuer Tag soll dahingehend klassifiziert werden, ob Luigi an dem neuen Tag Tennis spielen wird oder nicht.
- Die Trainingsdatenmenge ist in einer Tabelle gegeben.

Naive Bayes Klassifikation
Beispiel des Tennisspielers

Day	Outlook	Temperature	Humidity	Wind	Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Naive Bayes Klassifikation
Beispiel des Tennisspielers

Universität Augsburg University

- Klassifiziert werden soll nun die folgende neue Instanz:

$$(Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong)$$

- Wir wenden den Naive-Bayes Klassifikator an um die wahrscheinlichste Klasse (*yes* oder *no*) zu berechnen:

$$v_{NB} = \underset{v_j \in V}{\arg\max}\, P(v_j)\prod_i P(a_i \mid v_j)$$

$$= \underset{v_j \in \{yes, no\}}{\arg\max}\, P(v_j)\prod_i P(a_i \mid v_j)$$

$$= \underset{v_j \in \{yes, no\}}{\arg\max}\, P(v_j)\begin{pmatrix} P(Outlook = sunny \mid v_j) \\ P(Temperature = cool \mid v_j) \\ P(Humidity = high \mid v_j) \\ P(Wind = strong \mid v_j) \end{pmatrix}$$

Naive Bayes Klassifikation
Beispiel des Tennisspielers

Universität Augsburg University

- Abschätzen der Wahrscheinlichkeiten für *yes* und *no*

$$P(Tennis = yes) = 9/14$$

$$P(Tennis = no) = 5/14$$

- Bedingte Wahrscheinlichkeiten:

$$P(Outlook = sunny \mid Tennis = yes) = 2/9$$

$$P(Temperature = cool \mid Tennis = yes) = 3/9$$

$$P(Humidity = high \mid Tennis = yes) = 3/9$$

$$P(Wind = strong \mid Tennis = yes) = 3/9$$

$$P(Outlook = sunny \mid Tennis = no) = 3/5$$

$$P(Temperature = cool \mid Tennis = no) = 1/5$$

$$P(Humidity = high \mid Tennis = no) = 4/5$$

$$P(Wind = strong \mid Tennis = no) = 3/5$$

Day	Outlook	Tem	Hum	Wind	Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

- Die Vorhersage des Zielwerts für die neue Instanz ergibt:

$$P(yes)\begin{pmatrix} P(sunny \mid yes) \\ P(cool \mid yes) \\ P(high \mid yes) \\ P(strong \mid yes) \end{pmatrix} \approx .0053, \qquad P(no)\begin{pmatrix} P(sunny \mid no) \\ P(cool \mid no) \\ P(high \mid no) \\ P(strong \mid no) \end{pmatrix} \approx .0206$$

- Die Naive Bayes Klassifizierung wählt also *no* als das wahrscheinlichste Ergebnis für die neue Instanz.
- Die A-Posteriori Wahrscheinlichkeit für *no* kann man dann mittels der totalen Wahrscheinlichkeit berechnen.

$$P(yes \mid \ldots) \approx \frac{0.0053}{0.0053 + 0.0206} \approx \frac{0.0053}{0.0259} \approx 0.205$$

$$P(no \mid \ldots) \approx \frac{0.0206}{0.0053 + 0.0206} \approx \frac{0.0206}{0.0259} \approx 0.795$$

- **Problem**:
  - Die bisherige Methode zur Abschätzung führt zu einer Unterschätzung bei kleinem $n_{i,j}$ oder wenn $n_{i,j} \rightarrow 0$

- **Beispiel**:
  - Wenn wir keine Trainingsbeispiele haben, bei denen starker Wind vorlag und nicht Tennis gespielt wurde:

$$P(Wind = strong \mid PlayTennis = no) = 0$$

  - Dieser Wert dominiert bei einer neuen Instanz mit *Wind=strong* alle anderen Beobachtungen und führt zur Unterschätzung der Gesamtwahrscheinlichkeit.

- **M-Estimate of Probability**:

$$\frac{n_{i,j} + m \cdot p}{n_j + m}$$

  - *m* Anzahl der neu hinzugefügten virtuellen Trainingsbeispiele
  - n_j Anzahl der Trainigsbeispiele aus der Klasse v_j
  - $n_{i,j}$ Anzahl der Beispiele mit Attribut a_i und Klasse v_j
  - *p* A-Priori Wahrscheinlichkeit für Beispiele mit dem Attribut a_i und Klasse v_j. Um p zu wählen, nimmt man eine **Gleichverteilung** an, d.h falls das Attribut a_i *k* mögliche Werte hat, dann wählt man *1/k*

$$\frac{n_{i,j} + m \cdot p}{n_j + m}$$

- **Beispiel**:
  - Anzahl der Trainingsbeispiele mit *overcast* bei *no:*

    $$n_{overcast,no} = 0$$

  - Anzahl der Trainingsbeispiele für die Klasse *no:*

    $$n_{no} = 5$$

  - Annahme der Gleichverteilung für Attribut *outlook:*

    $$p = 1/3$$

  - M-Estimate für $P(overcast \mid no)$ bei verschiedenen *m*:

$$m = 0 \rightarrow P(overcast \mid no) = (0 + 1/3 \cdot 0)/(5 + 0) = 0$$

$$m = 1 \rightarrow P(overcast \mid no) = (0 + 1/3 \cdot 1)/(5 + 1) = 0.055$$

$$m = 100 \rightarrow P(overcast \mid no) = (0 + 1/3 \cdot 100)/(5 + 100) = 0.318$$

Multimodale Analyse

Naive Bayes - Bonusaufgabe

Elisabeth André
Stephan Hammer
Chi-Tai Dang

Human Centered Multimedia
Institute of Computer Science
Augsburg University
Universitätsstraße. 6a
86159 Augsburg, Germany

Naive Bayes Klassifikation Bonusaufgabe

Universität Augsburg University

- EINZELABGABEN!!
- Sie nehmen jetzt an einem kurzen Experiment teil, das mit dem Naive-Bayes Klassifikator zu tun hat. Nächste Vorlesung erfahren Sie was es genau damit auf sich hat.
- Natürlich werden Ihre Daten anonym behandelt und es werden keine personenbezogenen Daten gespeichert.
- Bitte beantworten Sie zunächst die folgenden Fragen.
- Nummerieren Sie bitte die Fragen für die Auswertung.

Naive Bayes Klassifikation
Bonusaufgabe

1. Welches Bild gefällt Ihnen besser?

a)

b)

Naive Bayes Klassifikation
Bonusaufgabe

Universität Augsburg University

2. Welches Paar Schuhe halten sie für bequemer?

a)

b)

c)

Naive Bayes Klassifikation
Bonusaufgabe

3. Welche Art von Komödie gefällt Ihnen besser?

a)

b)

4. Wenn Sie ein Auto wären, welches Auto käme ihrer Vorstellung am nächsten?

a)

b)

5. Sehen Sie sich gerne im Spiegel an?

a) Nein, ich mag das gar nicht.

b) Es ist ok, mich im Spiegel zu sehen.

c) Ja, ich mag meinen Anblick im Spiegel.

6. Was trifft eher zu?

a) Nutzen Sie eher Android?

b) Oder Nutzen Sie eher Apple?

c) Oder Nutzen Sie eher Windows?

Multimodale Analyse

Naive Bayes - Textklassifikation

Elisabeth André
Stephan Hammer
Chi-Tai Dang

HCM
Human Centered Multimedia
Institute of Computer Science
Augsburg University
Universitätsstraße. 6a
86159 Augsburg, Germany

Naiver Bayes Klassifikator
Textklassifikation

- NB Klassifikatoren in der **Textanalyse** weit verbreitet

Vorteile:

- Hohe Trainings- und Klassifizierungsgeschwindigkeit
- NB verbessert sich mit jeder neu klassifizierten Instanz
- Einsatz zur Kategorisierung von Dokumenten
  - Kategorien: z.B. Finanzen, Sport
  - Genres: z.B. Nachrichten, Kritiken
  - Meinungen: z.B. Positive, Negative
  - Labels: z.B. Spam und NoSpam

Die guten Klassifizierungseigenschaften von NB-Klassifikatoren machen sie insbesondere beim Einsatz in **Spam-Filtern** sehr beliebt

- **Gegeben:**
  - Trainingsmenge von Emails, von denen bekannt ist, ob sie in die Klasse *Spam* oder in die Klasse *No-Spam* klassifiziert wurden.
  - Neue Email, die in *Spam* oder *No-Spam* klassifiziert werden soll
- **Gesucht:**
  - Klassifiziere die neue Email in die Klasse *Spam* oder *No-Spam*
- **Attribute:**
  - Häufigkeit bestimmter Worte aus einem Vokabular

- Vereinfachende Annahme:
  - Klassifikation ist unabhängig von der Aufeinanderfolge von Worten
  - Quasi eine Vernachlässigung der grammatikalischen Regeln
- Naive Bayes Klassifikation

positions: alle Textpositionen im zu klassifizierenden Dokument
w_i: *Wort an der Position i*

$$v_{NB} = \underset{v_j \in V}{\arg\max}\, P(v_j) \prod_{i \in positions} P(w_i \mid v_j)$$

- Abschätzen von $P(v_j)$ wie gewöhnlich durch Abzählen
- Abschätzen von $P(w_k|v_j)$ mit m-estimate

Naiver Bayes Klassifikator
Textklassifikation

- Gleichverteilung der Wahrscheinlichkeiten aller Wörter $w_1,...,w_l$ im Vokabular, da jedes Wort w_k einmal im Vokabular vorkommt
- Anzahl der neuen virtuellen Beispiele (m) ist Größe des Vokabulars

$$P(w_k \mid v_j) \approx \frac{n_k + 1}{n + |V|}$$

- n ist die Anzahl der Wortpositionen in allen Trainingsbeispielen der Klasse v_j
- n_k ist die Häufigkeit der Auftreten des Wortes w_k unter diesen n Wortpositionen.
- |V| ist die Anzahl unterschiedlicher Worte in den Trainingsdaten

1. ***LearnNaiveBayesText*** **(** *E*, *V* **)**
   - *E* Menge aller Beispieldokumente und *V* Menge aller Klassen
   - Extrahiere ein Vokabular *Voc* aus der Trainingsbeispiel-Menge
   - Berechne die Abschätzungen für $P(v_j)$ und $P(w_k|v_j)$ wie folgt

1. Für jede der Klassen v_j :
   - $D_j \leftarrow$ Menge der Dokumente aus der Klasse v_j
   - $P(v_j) \leftarrow |D_j| / |E|$
   - $T_j \leftarrow$ Ein einziges Dokument bestehend aus allen D_j
   - $n_j \leftarrow$ Anzahl der Wörter im Dokument T_j
   - Für jedes Wort w_k aus dem Vokabular
     - $n_{k,j} \leftarrow$ Häufigkeit von Wort w_k im Dokument T_j
     - $P(w_k|v_j) \leftarrow (n_{k,j}+1) / (n_j + |Voc|)$

2. ***ClassifyNaiveBayesText(** D **)***

- *Sei D* das zu klassifizierende Dokument und *positions* die Menge der Wortpositionen in *D* besetzt mit einem Wort aus dem Vokabular.
- Sei w_i das Wort an Position *i* im Dokument *D* dann berechne

$$v_{NB} = \underset{v_j \in V}{\arg\max}\, P(v_j) \prod_{i \in positions} P(w_i \mid v_j)$$

Naiver Bayes Klassifikator Textklassifikation

	Doc	Words	Class
Training	1	free all free	spam
	2	free lottery free	spam
	3	free trip	spam
	4	free erasmus place	no_spam
Test	5	free free free erasmus place	?

$P(w_k|v_j) \leftarrow (n_{k,j}+1) / (n_j + |V|)$
$n_{k,j}$ Häufigkeit von Wort w_k in Dokument D_j
n_j Anzahl der Wörter in Dokument D_j

P(spam)=3/4
P(no_spam)=1/4
P(free|spam) = (5+1)/(8+6)=3/7
P(place|spam) = (0+1)/(8+6)=1/14
P(erasmus|spam)=(0+1)/(8+6)=1/14
P(free|no_spam)=(1+1)/(3+6)=2/9
P(place|no_spam)=(1+1)/(3+6)=2/9
P(erasmus|no_spam)=(1+1)/(3+6)=2/9

$$v_{NB} = \underset{v_j \in V}{\arg\max} P(v_j) \prod_{i \in positions} P(w_i | v_j)$$

P(spam|doc5) ~ 3/4 * $(3/7)^3$*1/14*1/14 =0,0098
P(no_spam|doc5) ~1/4 * $(2/9)^3$*2/9* 2/9=0,0023

Naiver Bayes Klassifikator
Performanz des Verfahrens

- **Experiment von Joachims, 1996:**
  - Sammlung von je 1.000 Artikeln aus 20 Newsgroups
  - Davon 2/3 Trainingsbeispiele und 1/3 Testbeispiele
- **Vokabularbildung:**
  - Entferne die 100 häufigsten Worte („the“, „of“, …)
  - Entferne Worte, die weniger als 3 mal auftauchen
- **Erkennungsrate:**
  - Zufälliges Raten: 5%
  - Mit Algorithmus: 89%

Naiver Bayes Klassifikator Zusammenfassung

- Die vorgestellten Bayes-Methoden ermitteln A-Posteriori Wahrscheinlichkeiten für Hypothesen basierend auf den angenommenen oder irgendwie abgeschätzten A-Priori Wahrscheinlichkeiten und den neu beobachteten Daten.
- Mit jeder neu klassifizierten Instanz verbessert sich der Naive Bayes Klassifikator und bietet eine gute Performanz
- NB gibt sogar eine Wahrscheinlichkeitsverteilung an.
- Der Naive Bayes-Klassifikator basiert auf der Annahme, die Attribute seien bedingt unabhängig voneinander.
- Obwohl diese Annahme in der Praxis häufig verletzt wird liefert der Ansatz doch oft gute Klassifikationsresultate.

Literaturhinweise zu Bayes Klassifikation

- T. Mitchell: Machine Learning
- Russel, Norvig: Artificial Intelligence: A Modern Approach.
- C.M. Bishop: Pattern Recognition and Machine Learning.