Klassifikation

Überblick

- Grundkonzepte
- Entscheidungsbäume
- Evaluierung von Klassifikatoren
- Lernen von Regeln
- Klassifikation mittels Assoziationsregeln
- **Naïver Bayescher Klassifikator**
- Naïve Bayes für Text Klassifikation
- Support Vektor Maschinen
- Ensemble-Methoden: Bagging und Boosting
- Zusammenfassung

Bayesche Klassifikation

- Probabilistische Sicht: Überwachtes Lernen kann auf elegante Weise probabilistisch formuliert werden.
- Seien A_1 bis A_k diskrete Attribute. Das Klassenattribut sei C.
- Gegeben sei ein Testbeispiel *d* mit den beobachteten Attributwerten a_1 bis a_k.
- Die Klassifikation besteht darin die folgende Aposteriori-Wahrscheinlichkeit zu berechnen. Die Vorhersage ist die Klasse c_j s.d.

$$\Pr(C = c_j \mid A_1 = a_1, \ldots, A_{|A|} = a_{|A|})$$

maximal wird

Anwendung von Bayes' Regel

$$\Pr(C = c_j \mid A_1 = a_1, \ldots, A_{|A|} = a_{|A|})$$

$$= \frac{\Pr(A_1 = a_1, \ldots, A_{|A|} = a_{|A|} \mid C = c_j) \Pr(C = c_j)}{\Pr(A_1 = a_1, \ldots, A_{|A|} = a_{|A|})}$$

$$= \frac{\Pr(A_1 = a_1, \ldots, A_{|A|} = a_{|A|} \mid C = c_j) \Pr(C = c_j)}{\sum_{r=1}^{|C|} \Pr(A_1 = a_1, \ldots, A_{|A|} = a_{|A|} \mid C = c_r) \Pr(C = c_r)}$$

- $\Pr(C=c_j)$ ist die Klasse *Prior*-Wahrscheinlichkeit: leicht aus den Trainingsdaten zu schätzen.

Berechnung der Wahrscheinlichkeiten

- Der Nenner ist irrelevant für Entscheidung , da er für jede Klasse gleich ist.
- Nur $P(A_1=a_1,...,A_k=a_k \mid C=c_i)$ wird gebraucht, was umgeformt werden kann als

  $Pr(A_1=a_1|A_2=a_2,...,A_k=a_k, C=c_j)* Pr(A_2=a_2,...,A_k=a_k |C=c_j)$
- Der zweite Faktor kann rekursiv auf die gleiche Weise weiter zerlegt werden.
- Jetzt kommt noch eine Annahme.

Bedingte Unabhängigkeitsannahme

- Alle Attribute sind bedingt unabhängig bei gegebener Klasse $C = c_j$.
- Formal wird angenommen,

  $\Pr(A_1=a_1 \mid A_2=a_2, ..., A_{|A|}=a_{|A|}, C=c_j) = \Pr(A_1=a_1 \mid C=c_j)$

  und so weiter für A_2 bis $A_{|A|}$. d.h.,

$$\Pr(A_1 = a_1,...,A_{|A|} = a_{|A|} \mid C = c_i) = \prod_{i=1}^{|A|} \Pr(A_i = a_i \mid C = c_j)$$

Naïver Bayescher Klassifikator

$$\Pr(C = c_j \mid A_1 = a_1, \ldots, A_{|A|} = a_{|A|})$$

$$= \frac{\Pr(C = c_j) \prod_{i=1}^{|A|} \Pr(A_i = a_i \mid C = c_j)}{\sum_{r=1}^{|C|} \Pr(C = c_r) \prod_{i=1}^{|A|} \Pr(A_i = a_i \mid C = c_r)}$$

- Wie wird P($A_i = a_i | C=c_j$) geschätzt?

Klassifikation einer Testinstanz

- Wenn die wahrscheinlichsten Klasse vorhergesagt wird, braucht nur der Zähler berechnet werden, da der Nenner für alle Klassen gleich ist.
- Für ein gegebenes Testbeispiel, wird das folgende berechnet

$$c = \arg\max_{c_j} \Pr(c_j) \prod_{i=1}^{|A|} \Pr(A_i = a_i \mid C = c_j)$$

Beispiel

- Berechne alle benötigten Wahrscheinlichkeiten für die Klassifikation

A	B	C
m	b	t
m	s	t
g	q	t
h	s	t
g	q	t
g	q	f
g	s	f
h	b	f
h	q	f
m	b	f

$Pr(C = t) = 1/2$, $Pr(C = f) = 1/2$

$Pr(A=m \mid C=t) = 2/5$ $Pr(A=g \mid C=t) = 2/5$ $Pr(A=h \mid C=t) = 1/5$

$Pr(A=m \mid C=f) = 1/5$ $Pr(A=g \mid C=f) = 2/5$ $Pr(A=h \mid C=n) = 2/5$

$Pr(B=b \mid C=t) = 1/5$ $Pr(B=s \mid C=t) = 2/5$ $Pr(B=q \mid C=t) = 2/5$

$Pr(B=b \mid C=f) = 2/5$ $Pr(B=s \mid C=f) = 1/5$ $Pr(B=q \mid C=f) = 2/5$

Now we have a test example:

$A = m \quad B = q \quad C = ?$

Beispiel, Fortsetzung

- Für C = t, ergibt sich

$$\Pr(C=t)\prod_{j=1}^{2}\Pr(A_j=a_j\,|\,C=t)=\frac{1}{2}\times\frac{2}{5}\times\frac{2}{5}=\frac{2}{25}$$

- Für Klasse C = f, ergibt sich

$$\Pr(C=f)\prod_{j=1}^{2}\Pr(A_j=a_j\,|\,C=f)=\frac{1}{2}\times\frac{1}{5}\times\frac{2}{5}=\frac{1}{25}$$

- C = t ist wahrscheinlicher. t ist deshalb die Vorhersage.

Probleme

- **Numerische Attribute**: Naïves Bayesches Lernen nimmt an, dass alle Attribute kategorisch sind. Numerische Attribute müssen diskretisiert werden.
- **Anzahl=Null**: Ein bestimmter Attributwert taucht unter Umständen nicht mit einer Klasse gemeinsam auf. Abhilfe durch Glättung.

$$\Pr(A_i = a_i \mid C = c_j) = \frac{n_{ij} + \lambda}{n_j + \lambda n_i}$$

- **Fehlende Werte**: werden ignoriert

Diskussion zum naïven Bayesches Klassifikator

- Vorteile:
  - Leicht zu implementieren
  - Sehr effizient
  - Liefert gute Ergebnisse bei vielen Anwendungen
- Nachteile
  - Annahme: Klassen sind bedingt unabhängig, deshalb geht Vorhersagegenauigkeit verloren, wenn diese Annahme stark verletzt wird. (z.B. bei stark korrellierten Daten)

Überblick

- Grundkonzepte
- Entscheidungsbäume
- Evaluierung von Klassifikatoren
- Lernen von Regeln
- Klassifikation mittels Assoziationsregeln
- Naïver Bayescher Klassifikator
- **Naïve Bayes für Text Klassifikation**
- Support Vektor Maschinen
- Ensemble-Methoden: Bagging und Boosting
- Zusammenfassung

Text Klassifikation/Kategorisierung

- Weil die Anzahl elektronischer Dokumente stark steigt, wird automatische Dokumentklassifikation immer wichtiger.
- Die bisher vorgestellten Techniken können zwar angewendet werden, sind aber nicht so effektiv wie die nachfolgenden Methoden.
- Heute wird eine naïve Bayesche Methode diskutiert, die speziell für Texte zugeschnitten ist, und Text-spezifische Eigenschaften nutzt.
- Die Ideen sind ähnlich zu naïve Bayes.

Probabilistischer Rahmen

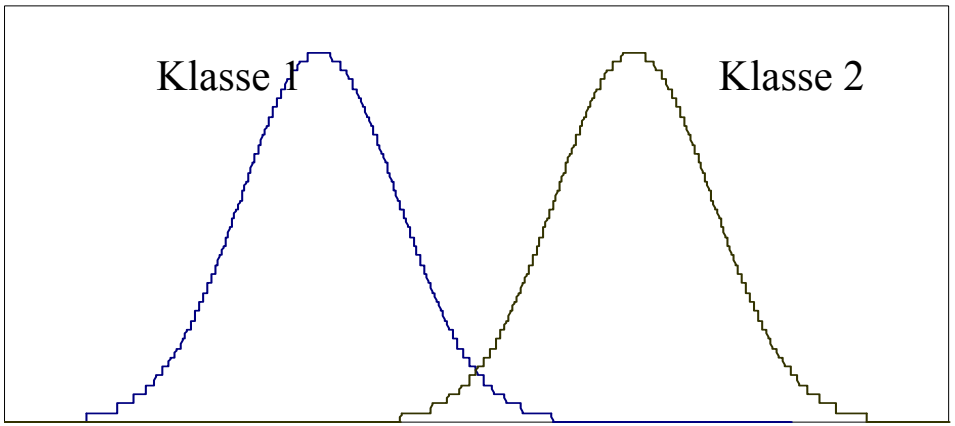
- Generatives Modell: Jedes Dokument wird durch eine parametrisierte Verteilung erzeugt, die von versteckten Parametern beeinflußt ist.
- Das generative Modell macht zwei Annahmen
  - Die Daten (oder Textdokumente) werden von einem Mischmodell erzeugt,
  - Zwischen den Komponenten der Mischverteilund und den Dokumentklassen gibt es eine eins-zu-eins Beziehung.

Mischmodell

- A **Mischmodell** beschreibt die Daten durch mehrere statistische Verteilungen.
  - Jede Verteilung korrespondiert zu einem Daten-Cluster und die Parameter der Verteilung sind eine Beschreibung des Clusters.
  - Jede Verteilung im Mischmodell wird auch **Mischkomponente** genannt.
- Eine Verteilung/Komponente kann von beliebiger Art sein

Beispiel

- Die Abbildung zeigt eine Wahrscheinlichkeits-dichtefunktion einer 1-dimensionale Datenmenge (mit zwei Klassen) erzeugt durch
  - eine Mischung von zwei Gauss-Verteilungen,
  - eine pro Klasse, deren Parameter (beschrieben durch θ_i) der Durchschnitt (μ_i) und die Standardabweichung (σ_i), d.h., $\theta_i = (\mu_i, \sigma_i)$.



Mischmodell (Fortsetzung …)

- Sei *K* die Anzahl der Mischkomponenten (oder Verteilungen) in einem Mischmodell.
- Die *j*te Verteilung hat die Parameter θ_j.
- Sei Θ die Menge der Parameters alle Komponenten, $\Theta = \{\varphi_1, \varphi_2, \ldots, \varphi_K, \theta_1, \theta_2, \ldots, \theta_K\}$, wobei φ_j das *Mischgewicht* (oder *Prior Wahrscheinlichkeit*) eine Mischekomponente *j* sei und θ_j die Parameter der Komponente *j*.
- Wie erzeugt das Modell die Dokumente?

Dokumenterzeugung

- Wegen der eins-zu-eins Beziehung zwischen Klassen und Mischkomponenten sind die Mischgewichte die *Klassen-Prior-Wahrscheinlichkeiten*, d.h., $\varphi_j = \Pr(c_j|\Theta)$.
- Das Mischmodell erzeugt ein Dokument d_i durch:
  - Auswahl der Mischkomponente (oder Klasse) bezüglich der Klassen-Prior.-Wahrscheinlichkeiten (d.h., der Mischgewichte), $\varphi_j = \Pr(c_j|\Theta)$.
  - Nachdem eine Komponente (c_j) gewählt ist, wird ein Dokument d_i bezüglich der Parameter mit der Verteilung $\Pr(d_i|c_j; \Theta)$ erzeugt, oder genauer $\Pr(d_i|c_j; \theta_j)$.

$$\Pr(d_i \mid \Theta) = \sum_{j=1}^{|C|} \Pr(c_j \mid \Theta) \Pr(d_i \mid c_j; \Theta) \qquad (23)$$

Modellierung von Textdokumenten

- Der naïve Bayes-Klassifikator behandelt jedes Dokument als “bag of words”.
- Das erzeugende Modell macht weitere Annahmen:
  - Wörter werden bei gegebener Klasse unabhängig von einander erzeugt (wie beim naïven Bayes).
  - Die Wahrscheinlichkeit eines Wortes ist unabhänigi von seiner Position im Dokument. Die Dokumentlänge wird unabhängig von der Klasse gewählt.

Multinomiale Verteilung

- Wegen der bisherigen Annahmen, kann eine Dokument durch eine Multinomial-Verteilung erzeugt werden.
- D.h. jedes Dokument wird aus einer Multinomial-Verteilung von Worten gezogen, die Anzahl der Versuche entspricht der Dokumentlänge.
- Die Worte sind aus einem gegeben Vokabular $V = \{w_1, w_2, \ldots, w_{|V|}\}$.

Verteilungsfunktion einer Multinomial-Verteilung

$$\Pr(d_i \mid c_j;\Theta) = \Pr(\mid d_i \mid) \mid d_i \mid ! \prod_{t=1}^{\mid V \mid} \frac{\Pr(w_t \mid c_j;\Theta)^{N_{ti}}}{N_{ti}!} \qquad (24)$$

wobei N_{ti} ist die Anzahl des Auftretens von Wort w_t in Dokument d_i und

$$\sum_{t=1}^{\mid V \mid} N_{it} = \mid d_i \mid \qquad \sum_{t=1}^{\mid V \mid} \Pr(w_t \mid c_j;\Theta) = 1. \qquad (25)$$

Parameter-Schätzung

- Parameter werden durch empirische Anzahlen geschätzt.

$$\Pr(w_t \mid c_j; \hat{\Theta}) = \frac{\sum_{i=1}^{|D|} N_{ti} \Pr(c_j \mid d_i)}{\sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N_{si} \Pr(c_j \mid d_i)}. \qquad (26)$$

- Um 0 Anzahlen für seltene Worte, die nicht in der Trainingsmenge aber in der Testmenge auftauchen, wird die Wahrscheinlichkeits-schätzung geglättet. *Lidstone* Glättung, $0 \leq \lambda \leq 1$

$$\Pr(w_t \mid c_j; \hat{\Theta}) = \frac{\lambda + \sum_{i=1}^{|D|} N_{ti} \Pr(c_j \mid d_i)}{\lambda |V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N_{si} \Pr(c_j \mid d_i)}. \qquad (27)$$

Parameter-Schätzung (Fortsetzung…)

- Klassen-Prior-Wahrscheinlichkeiten, welche die Mischgewichte φ_j sind können leicht aus den Trainingsdaten geschätzt werden

$$\Pr(c_j \mid \hat{\Theta}) = \frac{\sum_{i=1}^{|D|} \Pr(c_j \mid d_i)}{|D|} \qquad (28)$$

Klassifikation

- Gegeben eine Test-Dokument d_i, mit den Gleichungen (23) (27) und (28)

$$\Pr(c_j \mid d_i; \hat{\Theta}) = \frac{\Pr(c_j \mid \hat{\Theta}) \Pr(d_i \mid c_j; \hat{\Theta})}{\Pr(d_i \mid \hat{\Theta})}$$

$$= \frac{\Pr(c_j \mid \hat{\Theta}) \prod_{k=1}^{|d_i|} \Pr(w_{d_i,k} \mid c_j; \hat{\Theta})}{\sum_{r=1}^{|C|} \Pr(c_r \mid \hat{\Theta}) \prod_{k=1}^{|d_i|} \Pr(w_{d_i,k} \mid c_r; \hat{\Theta})}$$

where $w_{d_i,k}$ is the word in position k of document d_i. If the final classifier is to classify each document into a single class, then the class with the highest posterior probability is selected:

$$\arg\max_{c_j \in C} \Pr(c_j \mid d_i; \hat{\Theta}) \qquad (30)$$

Diskussion

- Die meisten Annahmen beim naïven Bayes-Klassifikator werden zu einem gewissen Grad in der Praxis verletzt.
- Trotzdessen, ergibt der naïve Bayes Klassifikator brauchbare Modelle.
  - Die Hauptannahme ist die der Mischverteilung. Wenn diese Annahme stark verletzt wird, kann die Klassifikationsgenauigkeit rapide sinken.
- Der naïve Bayes-Klassifikator ist sehr effizient.