

Vorlesungsplan

- 17.10. Einleitung
- 24.10. Ein- und Ausgabe
- 31.10. Reformationstag, Einfache Regeln
- 7.11. Naïve Bayes, Entscheidungsbäume
- 14.11. Entscheidungsregeln, Assoziationsregeln
- 21.11. Lineare Modelle, Instanzbasiertes Lernen
- 28.11. Clustering
- 5.12. Evaluation
- 12.12. Evaluation
- 19.12. Lineare Algebra für Data Mining
- 9.1. Statistik für Data Mining
- 16.1. Lineare Modelle, Support Vector Machines (SVM)
- 19.1. Vorlesung statt Übung: Bayes-Netze
- 23.1. **Clustering**
- 26.1. Vorlesung statt Übung: Clustering II
- 30.1. Finden von häufigen Teilstrukturen
- 2. 2. Klausur

Probabilistische Clusteranalyse

- Probleme von heuristischen Ansätzen:
 - Aufteilung in k Cluster?
 - Reihenfolgeabhängigkeit von Algorithmen?
 - Sind die Zuordnungsoperationen sinnvoll?
 - Was ist das Optimierungskriterium und wird zumindest ein *lokales* Minimum erreicht?
- Probabilistische Perspektive \Rightarrow suche die *wahrscheinlichsten* Cluster bei gegebenen Daten
- Instanzen gehören zu jedem Cluster mit einer gewissen Wahrscheinlichkeit

Mixturen von Verteilungen

- Modelliere die Data mittels einer Mixtur von einer endlichen Anzahl von Verteilungen
- Ein Cluster entspricht einer Verteilung
 - bestimmt Whr. von Attributwerten in diesem Cluster
- *Endliche Mixturen*: endliche Anzahl von Clustern
- Cluster werden meist durch Normalverteilungen beschrieben
- Die Cluster-Verteilungen werden durch Gewichte kombiniert

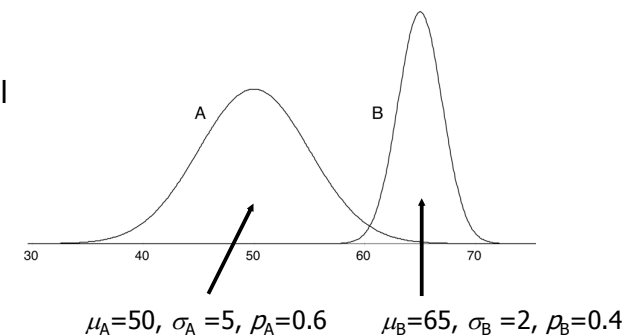
Zwei Klassen Mixtur-Modell

Daten
(eindimensional)

A	51	B	62	B	64	A	48	A	39	A	51
A	43	A	47	A	51	B	64	B	62	A	48
B	62	A	52	A	52	A	51	B	64	B	64
B	64	B	64	B	62	B	63	A	52	A	42
A	45	A	51	A	49	A	43	B	63	A	48
A	42	B	65	A	48	B	65	B	64	A	41
A	46	A	48	B	62	B	66	A	48		
A	45	A	49	A	43	B	65	B	64		
A	45	A	46	A	40	A	46	A	48		

Modell

Parameter



Anwendung des Mixtur-Modelles

- Whr., daß Instanz x zu Cluster A gehört:

$$\Pr[A|x] = \frac{\Pr[x|A]\Pr[A]}{\Pr[x]} = \frac{f(x; \mu_A, \sigma_A)p_A}{\Pr[x]}$$

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

- *Likelihood* einer Instanz bei gegeben Clustern:

$$\Pr[x | \text{die Verteilungen}] = \sum_i \Pr[x | \text{Cluster}_i] \Pr[\text{Cluster}_i]$$

Allgemein: Mixtur-Dichten

$$p(\mathbf{x}) = \sum_{i=1}^k p(\mathbf{x} | G_i) P(G_i)$$

- G_i sind Komponenten/Gruppen/Cluster, $P(G_i)$ Mixtur-Verhältnisse (priors), $p(\mathbf{x} | G_i)$ Komponenten-Dichten
- Gauß-Mixtur mit: $p(\mathbf{x}|G_i) \sim N(\boldsymbol{\mu}_i, \Sigma_i)$
- Parameter: $\Phi = \{P(G_i), \boldsymbol{\mu}_i, \Sigma_i\}_{i=1}^k$
- Trainingsdaten ohne Klassen: $X = \{\mathbf{x}^t\}_t$ (unüberwachtes Lernen)

Klassen vs. Cluster

- Überwacht: $X = \{\mathbf{x}^t, \mathbf{r}^t\}_t$
- Klassen $C_i, i=1, \dots, K$
- $p(\mathbf{x}) = \sum_{i=1}^K p(\mathbf{x} | C_i) P(C_i)$
- mit $p(\mathbf{x} | C_i) \sim N(\boldsymbol{\mu}_i, \Sigma_i)$
- $\Phi = \{P(C_i), \boldsymbol{\mu}_i, \Sigma_i\}_{i=1}^K$
- $\hat{P}(C_i) = \frac{\sum_t r_i^t}{N}$ $\mathbf{m}_i = \frac{\sum_t r_i^t \mathbf{x}^t}{\sum_t r_i^t}$
- $\mathbf{s}_i = \frac{\sum_t r_i^t (\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x}^t - \mathbf{m}_i)^T}{\sum_t r_i^t}$
- Unüberwacht: $X = \{\mathbf{x}^t\}_t$
- Cluster $G_i, i=1, \dots, k$
- $p(\mathbf{x}) = \sum_{i=1}^k p(\mathbf{x} | G_i) P(G_i)$
- mit $p(\mathbf{x} | G_i) \sim N(\boldsymbol{\mu}_i, \Sigma_i)$
- $\Phi = \{P(G_i), \boldsymbol{\mu}_i, \Sigma_i\}_{i=1}^k$

Zuordnungen, \mathbf{r}^t ?

Lernen der Cluster

- Annahme:
 - es gibt k Cluster in den Trainingsdaten
- Lerne die Cluster \Rightarrow
 - bestimme ihre Parameter
 - d.h. z.B. Durchschnitt und Standardabweichung
- Gütekriterium:
 - *Likelihood der Trainingsdaten bei gegeben Clustern*
- EM Algorithmus
 - finde ein lokales Maximum der Likelihood-Fkt.

Expectation-Maximization (EM)

- Log-Likelihood eines Mixtur-Modelles

$$L(\Phi | X) = \log \prod_i p(\mathbf{x}^i | \Phi)$$

$$= \sum_i \log \sum_{j=1}^k p(\mathbf{x}^i | G_j) P(G_j)$$

- Kann nicht analytisch optimiert werden
- Annahme
 - Versteckte Variablen z , welche die Opt.vereinfachen, wenn sie bekannt wären, (z.B. Zuordnungen zu Clustern)
 - Vollständige Likelihood, $L_c(\Phi | X, Z)$, bezüglich \mathbf{x} und z
 - Unvollständige Likelihood, $L(\Phi | X)$, bezüglich \mathbf{x}

EM mit Gauß-Mixturen

- $z_i^t = 1$, falls \mathbf{x}^t zu G_i gehört, 0 sonst (entspricht Klassen r^t_i beim überwachten Lernen); Annahme $p(\mathbf{x} | G_i) \sim N(\boldsymbol{\mu}_i, \Sigma_i)$

- E-step: $E[z_i^t | X, \Phi^l] = \frac{p(\mathbf{x}^t | G_i, \Phi^l) P(G_i)}{\sum_j p(\mathbf{x}^t | G_j, \Phi^l) P(G_j)}$
$$= P(G_i | \mathbf{x}^t, \Phi^l) \equiv h_i^t$$

- M-step: $P(G_i) = \frac{\sum_t h_i^t}{N}$ $\mathbf{m}_i^{l+1} = \frac{\sum_t h_i^t \mathbf{x}^t}{\sum_t h_i^t}$

Nutze geschätzte Label anstelle der unbekanntem Label

$$\mathbf{S}_i^{l+1} = \frac{\sum_t h_i^t (\mathbf{x}^t - \mathbf{m}_i^{l+1})(\mathbf{x}^t - \mathbf{m}_i^{l+1})^T}{\sum_t h_i^t}$$

E- und M-Schritte

- Iteriere die zwei Schritte
- 1. E-Schritt: schätze z gegeben X und akt. Φ
- 2. M-Schritt: Finde neues Φ' gegeben z , X , und altes Φ .

$$\text{E-step: } Q(\Phi | \Phi^l) = E[L_c(\Phi | X, Z) | X, \Phi^l]$$

$$\text{M-step: } \Phi^{l+1} = \arg \max_{\Phi} Q(\Phi | \Phi^l)$$

Wenn Q erhöht wird, steigt auch die unvollständige Likelihood

$$L(\Phi^{l+1} | X) \geq L(\Phi^l | X)$$

