# 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

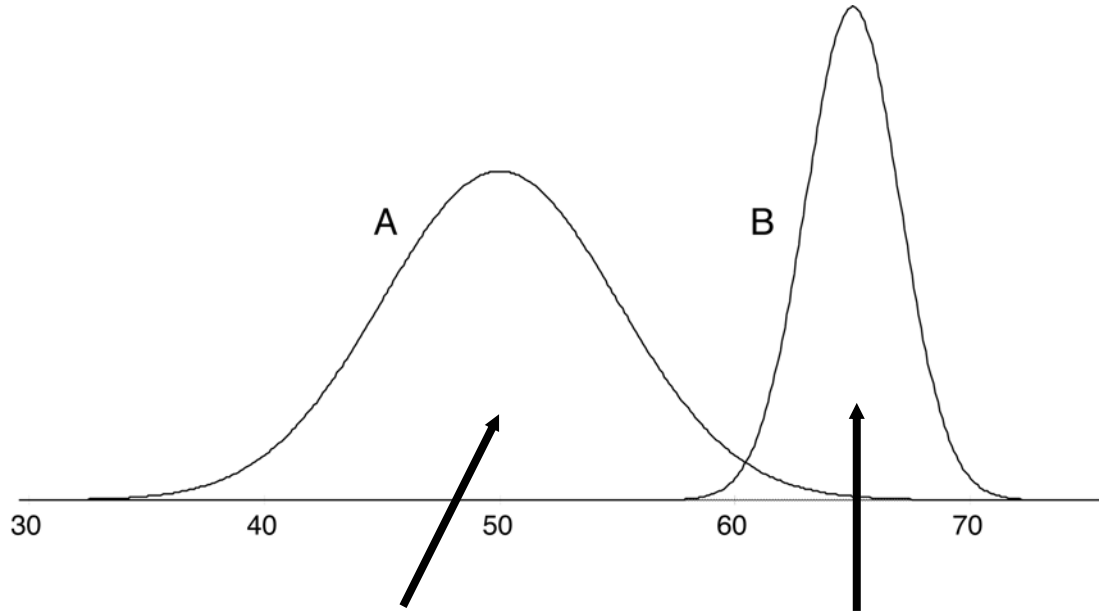
# Zwei Klassen Mixtur-Modell

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




$\mu_A$=50, $\sigma_A$ =5, $p_A$=0.6 $\mu_B$=65, $\sigma_B$ =2, $p_B$=0.4

# Anwendung des Mixtur-Modelles

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

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

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:

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

# Allgemein: Mixtur-Dichten

$$p(\boldsymbol{x}) = \sum_{i=1}^{k} p(\boldsymbol{x} \mid \mathsf{G}_i) P(\mathsf{G}_i)$$

- $\mathsf{G}_i$ sind Komponenten/Gruppen/Cluster,
  $P(\mathsf{G}_i)$ Mixtur-Verhältnisse (priors),
  $p(\boldsymbol{x} \mid \mathsf{G}_i)$ Komponenten-Dichten

- Gauß-Mixtur mit: $p(\boldsymbol{x}|\mathsf{G}_i) \sim \mathsf{N}(\boldsymbol{\mu}_i, \sum_i)$
- Parameter: $\Phi = \{P(\mathsf{G}_i), \boldsymbol{\mu}_i, \sum_i\}^k_{i=1}$
- Traingingsdaten ohne Klassen: $X=\{\boldsymbol{x}^t\}_t$ (unüberwachtes Lernen)

# Klassen vs. Cluster

- Überwacht: X = $\{ \boldsymbol{x}^t, \boldsymbol{r}^t \}_t$
- Klassen $C_i$ $i$=1,...,$K$

$$p(\mathbf{x}) = \sum_{i=1}^{K} p(\mathbf{x} \mid C_i) P(C_i)$$

mit $p$ ( $\boldsymbol{x}$ | $C_i$) ~ N ( $\boldsymbol{\mu}_i$ , $\sum_i$ )

- $\Phi = \{P(C_i), \boldsymbol{\mu}_i, \sum_i\}^K_{i=1}$

$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} \quad \mathbf{m}_i = \frac{\sum_t r_i^t \mathbf{x}^t}{\sum_t r_i^t}$$

$$\mathbf{S}_i = \frac{\sum_t r_i^t (\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x}^t - \mathbf{m}_i)^T}{\sum_t r_i^t}$$

- Unüberwacht: X = $\{ \boldsymbol{x}^t \}_t$
- Cluster $G_i$ $i$=1,...,$k$

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

mit $p$ ( $\boldsymbol{x}$ | $G_i$) ~ N ( $\boldsymbol{\mu}_i$ , $\sum_i$ )

- $\Phi = \{P(G_i), \boldsymbol{\mu}_i, \sum_i\}^k_{i=1}$

Zuordnungen, $\boldsymbol{r}^t_i$ ?

# 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

$$\mathrm{L}(\Phi \mid \mathrm{X}) = \log \prod_t p\left(\mathbf{x}^t \mid \Phi\right)$$

$$= \sum_t \log \sum_{i=1}^{k} p\left(\mathbf{x}^t \mid \mathrm{G}_i\right) P\left(\mathrm{G}_i\right)$$

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

# 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.

$$\mathrm{E}\text{-step}: \mathrm{Q}\left(\Phi \mid \Phi^{l}\right) = E\left[\mathrm{L}_{C}\left(\Phi \mid \mathrm{X},\mathrm{Z}\right) \mid \mathrm{X},\Phi^{l}\right]$$

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

Wenn Q erhöht wird, steigt auch die unvollständige Likelihood

$$\mathrm{L}\left(\Phi^{l+1} \mid \mathrm{X}\right) \geq \mathrm{L}\left(\Phi^{l} \mid \mathrm{X}\right)$$

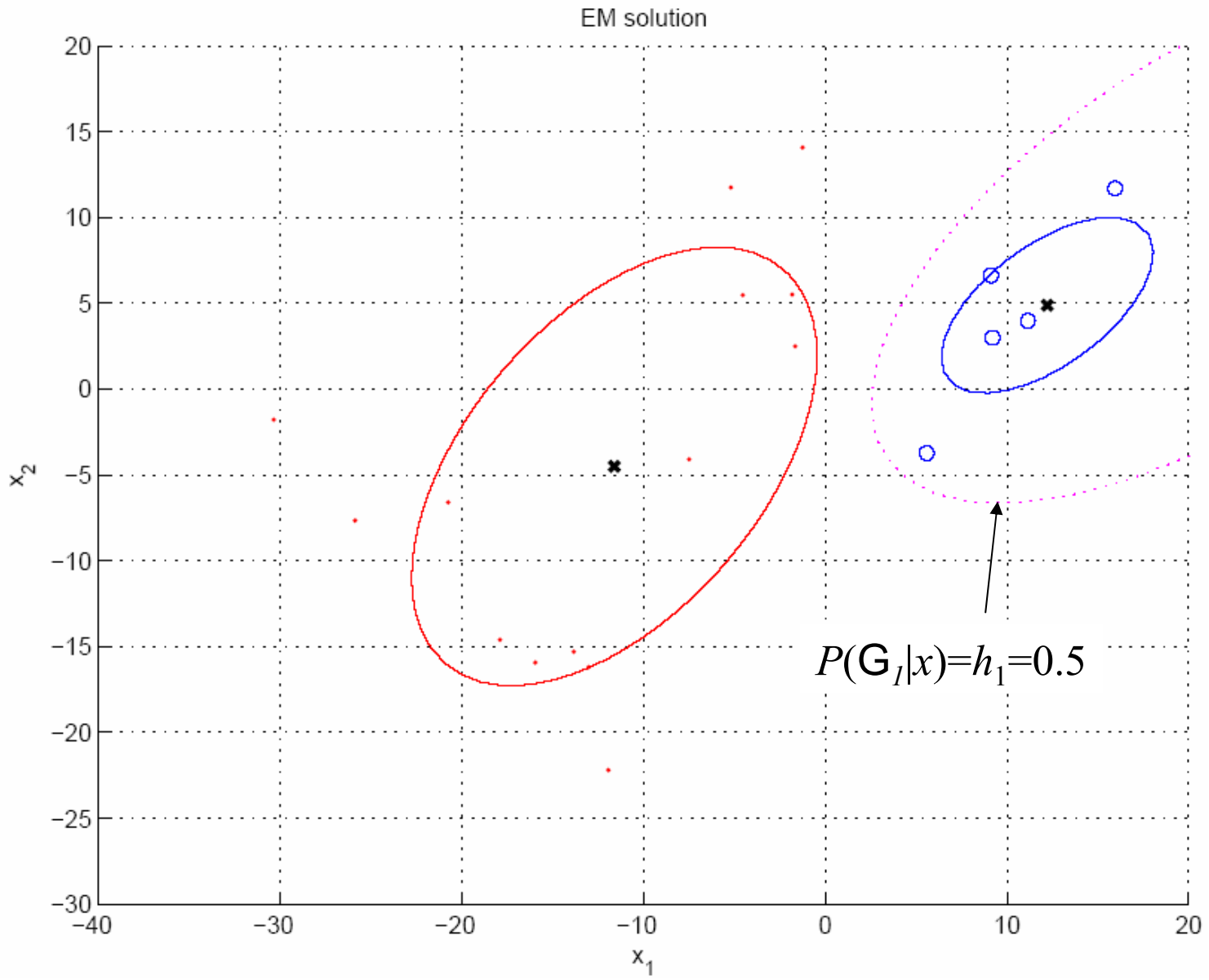
# EM mit Gauß-Mixturen

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

- E-step:
$$E\left[z_i^t \middle| \mathrm{X}, \Phi^l\right] = \frac{p\left(\mathbf{x}^t \mid \mathrm{G}_i, \Phi^l\right) P\left(\mathrm{G}_i\right)}{\sum_j p\left(\mathbf{x}^t \mid \mathrm{G}_j, \Phi^l\right) P\left(\mathrm{G}_j\right)}$$
$$= P\left(\mathrm{G}_i \mid \mathbf{x}^t, \Phi^l\right) \equiv h_i^t$$

- M-step:
$$P\left(\mathrm{G}_i\right) = \frac{\sum_t h_i^t}{N} \qquad \mathbf{m}_i^{l+1} = \frac{\sum_t h_i^t \mathbf{x}^t}{\sum_t h_i^t}$$
$$\mathbf{S}_i^{l+1} = \frac{\sum_t h_i^t \left(\mathbf{x}^t - \mathbf{m}_i^{l+1}\right)\left(\mathbf{x}^t - \mathbf{m}_i^{l+1}\right)^T}{\sum_t h_i^t}$$

*Nutze geschätzte Label anstelle der unbekannten Label*


EM solution
20
15
10
5
0
−5
−10
−15
−20
−25
−30
−40
−30
−20
−10
0
10
20
x2
x1
P(G1|x)=h1=0.5