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

Mischmodelle für dyadische Daten

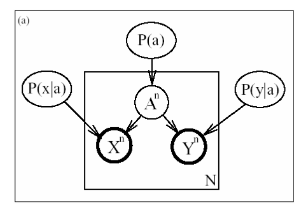
- Dyadische Daten: Paare $(x, y) \in X \times Y$
  $X = \{x_1, \ldots, x_I\}, Y = \{y_1, \ldots, y_J\}$
- Wenn $|X \times Y|$ sehr groß ist:
  - viele Paare (x,y) haben kleine Whr.
  - Häufigkeit ist oft Null
  - anfällig für kleine Störungen
- Anwendungen
  - Textanalyse, Suche:
    - X = Dokumente, Y = Wörter,
    - ein Text besteht aus vielen Paaren (Dok., Wort)
  - Bild-Segementierung
    - X=Bildpositionen, Y=Textureigenschaften
  - Empfehlungssystem für Filme
    - X=Anwender, Y=Filme, Paar (x,y) beschreibt Wertung

Formale Beschreibung der Daten

- Trainingsdaten $S = \{(x^n, y^n)\}_{n=1,\ldots,N}$ sind eine Realisierung von N Paaren von Zufallsvariablen $(X^n, Y^n)_{n=1,\ldots,N}$
- Modell für versteckte Variablen
  - Beobachtungen (x^n, y^n) werden mit versteckten Zufallsvariablen A^n verknüpft
  - A^n kann Werte aus $A = \{a_1, \ldots, a_K\}$ annehmen
  - Eine Realisierung $\vec{a} = (a^n)_{n=1,\ldots,N}$ partitioniert S in K Gruppen

Aspekt-Modell

- Paare $(x^n, y^n) \in S$ sind iid.
- Zufallsvariablen X^n und Y^n sind unabhängig bei gegebenem Aspekt A^n



- Generatives Modell
  - 1. wähle einen Aspekt $a \in A$ mit $Pr[a]$
  - 2. wähle $x \in X$ mit $Pr[x|a]$
  - 3. wähle $y \in Y$ mit $Pr[y|a]$

Aspekt-Modell (2)

- Verbundwahrscheinlichkeit der Trainingsdaten und einer hypothetischen Zuordnung von Aspekten

$$Pr[S,\vec{a}] = \prod_{n=1}^{N} Pr[x^n, y^n, a^n] \text{ mit}$$

$$Pr[x^n, y^n, a^n] = Pr[a^n] \cdot Pr[x^n|a^n] \cdot Pr[y^n|a^n]$$

- Durch Summieren über alle möglichen Realisierungen von $\vec{a} \in A^N = A \times A \times \cdots \times A$ mit $|A^N| = K^N$

$$Pr[S] = \prod_{x \in X} \prod_{y \in Y} Pr[x,y]^{n(x,y)} \text{ mit}$$

$$Pr[x,y] = \sum_{a \in A} Pr[a] Pr[x|a] Pr[y|a]$$

$$n(x,y) = |\{(x^n, y^n) \colon X^n = x \wedge y^n = y\}|$$

EM Algorithmus für Aspekt-Modell

- Parameter $\Phi = \{Pr[a], Pr[x|a], Pr[y|a]\}$
- E-Schritt

$$h_i^n := Pr[a_i|x^n, y^n, \Phi^l] = \frac{Pr[x^n|a_i] Pr[y^n|a_i] Pr[a_i]}{\sum_{a' \in A} Pr[x^n|a'] Pr[y^n|a'] Pr[a']}$$

- M-Schritt

$$Pr[a_i] = \frac{1}{N} \sum_{n=1}^{N} h_i^n \qquad Pr[x|a_i] = \frac{\sum_{n=1 \ldots N : x^n = x}^{N} h_i^n}{\sum_{n=1}^{N} h_i^n}$$

$$Pr[y|a_i] = \frac{\sum_{n=1 \ldots N : y^n = y}^{N} h_i^n}{\sum_{n=1}^{N} h_i^n}$$

Probabilistische Faktoranalyse für diskrete Daten

- Die Matrix der Zähler $C = (n(x_i, y_j))_{i,j}$ kann durch SVD in Faktoren zerlegt werden

$$C = U \Sigma V^T$$

- Das Aspekt-Modell kann als Matrixzerlegung der Verbundwhr. $P = (Pr[x_i, y_j])_{i,j}$ verstanden werden, mit

$$\hat{\Sigma} = diag(Pr[a_k])_k$$

$$\hat{U} = (Pr[x_i|a_k])_{i,k}$$

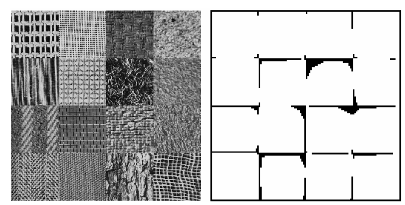
$$\hat{V} = (Pr[y_j|a_k])_{j,k}$$

$$P = \hat{U} \hat{\Sigma} \hat{V}^T$$

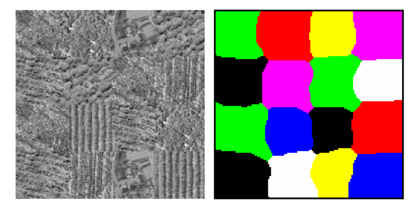
Hier sind $\hat{U}$ und $\hat{V}$ nicht orthogonal!

Anwendung Bild-Segementierung

- X=Bildpositionen, Y=Textureigenschaften



16 verschiedene Texturen



7 verschiedene Texturen aus Luftbildern auf 16 Felder verteilt

- Weitere Anwendungen (Siehe Arbeiten von Thomas Hofmann)
  - Textanalyse, Suche:
    - X = Dokumente, Y = Wörter,
    - ein Text besteht aus vielen Paaren (Dok., Wort)
  - Empfehlungssysteme für Filme
    - X=Anwender, Y=Filme, Paar (x,y) beschreibt Wertung

Spektrale Clusteranalyse

- Algorithmus von A.Y. Ng, M.I.Jordan und Y.Weiss

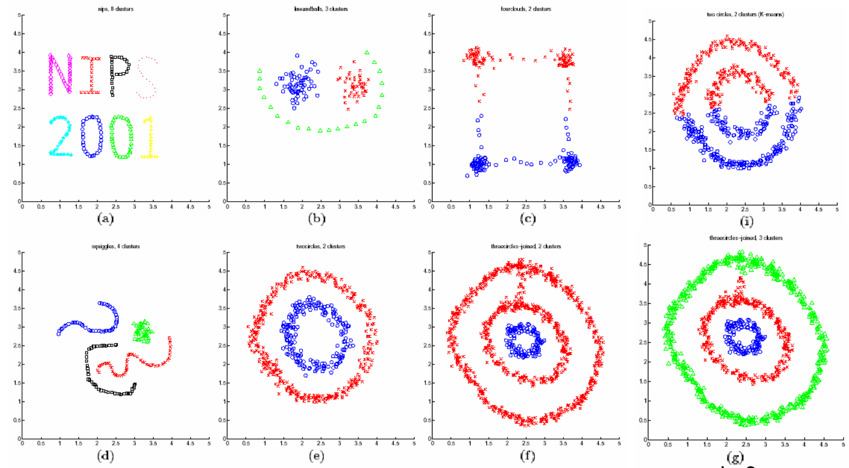
Gegeben: eine Distanzmatrix P von n Objekten

Gesucht: k Cluster

1. Berechne Affinitätsmatrix $A \in \mathbb{R}^{n \times n}$ $A_{ij} = \begin{cases} e^{-P_{ij}/2\sigma^2} & , i \neq j \\ 0 & , sonst \end{cases}$
2. $D = diag(\sum_j A_{1j}, \ldots, \sum_j A_{nj})$ $L = D^{-1/2} A D^{-1/2}$
3. Finde die Eigenvek. von L mit k größten Eigenwerten (orthogonal im Fall von wiederholten Eigenwerten) $X = [x_1, \ldots, x_k]$
4. Normalisiere die Zeilen von X, $Y_{ij} = X_{ij} / (\sum_{j'} X_{ij'}^2)^{-1/2}$
5. Clustere Zeilen von Y als Punkte im $\mathbb{R}^k$ mit k-Means
6. Weise i-tes Obj. dem Cluster j zu, wenn Zeile i von Y in dem k-Means Cluster j ist

Beispiele

Normal k-Means



k=2

k=3