Statistik IV für Nebenfachstudierende

6.2 Nicht-hierarchische Clusterverfahren

Prof. Dr. Andreas Mayr

Institut für Statistik, LMU München

Sommersemester 2017

Optimale Partitionierung (nicht-hierarchisch)

- Grundidee: Berechne alle möglichen Partitionierungen und wähle die beste.
- Im Allgemeinen kein hierarchisches Vorgehen.
- Formal:
  - $H(\mathcal{E})$ ist Heterogenität der Partition.
  - Gesucht wird optimale Partitionierung $\mathcal{E}_{\text{opt}}$.

$$\mathcal{E}_{\text{opt}} = \min_{\mathcal{E}} H(\mathcal{E})$$

- Wie findet man ε_{opt}?
- Ansatz: *Try all, keep best*
- Hauptproblem: Dimensionalität
  - Schon bei $n = 10$ und 3 Clustern gibt es 9330 Möglichkeiten.
  - Bei $n = 100$ und 3 Clustern $\approx 6 \cdot 10^{29}$
- Meist verwendet man einen Austauschalgorithmus:
  - Gestartet wird mit einer zufälligen Partitionieren *mit fixer Cluster-Anzahl*
  - Es werden dann iterativ so lange die Cluster-Zentren verschoben um $H(\varepsilon)$ zu verringern bis es keine Änderung der Zugehörigkeiten mehr gibt.

Unterschied von partitionierenden Clustern zum hierarchischen:

- Die Anzahl der Cluster muss von Anfang an feststehen.
- Die Beobachtungen ändern typischerweise im Laufe des Verfahrens ihre Clusterzugehörigkeit (Austauschverfahren).
- Dies ist bei hierarchischen Clustern nicht möglich: Einmal gebildete Cluster bleiben, sie können nur verfeinert (*divisiv*) oder verallgemeinert (*agglomerativ*) werden.

Optimierungskriterien

Basieren auf Streuungszerlegung

$$\mathbf{T} = \underline{\underline{\mathbf{W}}}(\mathcal{E}) + \underline{\underline{B}}(\mathcal{E})$$

- Varianzkriterium (auch Spur-Kriterium)
  - Für jedes Cluster wird quadratische Abweichung zum Centroid bestimmt und über alle Cluster aufsummiert.

$$H(\mathcal{E}) = \sum_{k=1}^{g} \sum_{\underline{x}_i \in C_k} \| \underline{x}_i - \bar{\underline{x}}_k \|^2$$

- Determinantenkriterium
  - Berücksichtigt zusätzlich Korrelation zwischen den Merkmalen.

$$H(\mathcal{E}) = |\underline{\mathbf{W}}(\mathcal{E})|$$

- Verallgemeinertes Determinantenkriterium
  - Noch flexibler

$$H(\mathcal{E}) = \sum_{k=1}^{g} n_k \ln\left(|\frac{1}{n_k}\mathbf{W}(C_k)|\right)$$

Mischverteilungsansätze

Annahme

- Grundgesamtheit zerfällt in unterschiedliche Cluster.
- Beobachtungen $\underline{x}_1, \ldots, \underline{x}_n$ kommen aus Mischverteilung:

$$f(\underline{x}) = \sum_{i=1}^{k} p(i) \cdot f(\underline{x}, i)$$

- Ergebnis: Führt über Satz von Bayes zu Posteriori-Wahrscheinlichkeiten:

$$\hat{\mathbb{P}}(k \mid \underline{x}_i)$$

Stochastische Partitionierung

Ähnlich wie optimale Partitionierung, nur jetzt basierend auf Verteilungen

$\underline{x}_i$ im Cluster k:

- $\underline{x}_i \sim N(\mu_k, \sigma^2 \mathbf{I})$
- $\underline{x}_i \sim N(\mu_l, \underline{\underline{\Sigma}})$
- $\underline{x}_i \sim N(\mu_l, \underline{\underline{\Sigma}}_k)$

Übersicht unterschiedliche Verfahren