Vertiefende statistische Verfahren

5. Übungsblatt SS 2024

Allgemeine Information

Alle Aufgaben sind mit R zu lössen, wenn nicht explizit anders angegeben. Die Berechnungen sollen nachvollziehbar und dokumentiert sein. Um die vollständige Punktezahl zu erreichen, müssen alle Ergebnisse und Fragen entsprechend interpretiert bzw. beantwortet werden. Code alleine ist nicht ausreichend! Die Abgabe erfolgt über Moodle entsprechend der Abgaberichtlinien als pdf und Rmd File. Bitte inkludieren Sie namentlich alle beteiligten Gruppenmitglieder sowohl im Bericht als auch im Source Code. Die jeweiligen Datensätze die für diese Übung relevant sind finden Sie ebenfalls in Moodle.

1 Clustering - Diabetes[3P]

Verwenden Sie den Datensatz `diabetes_RM.csv`. Der Datensatz enthält fünf Messungen, die an 145 nicht adipösen erwachsenen Patienten durchgeführt wurden (Beschreibung siehe UE4). Reaven und Miller [ref] wendeten in Anlehnung an Friedman und Rubin (1967) eine Clusteranalyse auf die drei primären Variablen (`insulin`,`glucose` und `sspg`) an und identifizierten drei Cluster: "normal", "chemical" und "overt" diabetische Probanden. Die Variable `group` enthält die Klassifizierungen der Probanden in diese drei Gruppen und dient hier als Ground Truth.

- Führen Sie eine Clusteranalyse durch. Verwenden Sie eine Clusteranzahl von 3 und vergleichen Sie die Genauigkeit (gegenüber Ground Truth) folgender Cluster-Algorithmen:
  - k-means, k-medoids
  - hierarchisches Clustering
  - hierarchischer k-means
  - Modell-basiertes Clustering
- Stellen Sie die Ergebnisse grafisch dar (Scatter Plot).
- Welches Verfahren ist am besten geeignet?
- Finden Sie für k-means die optimale Anzahl an Cluster, und beurteilen Sie ob sich die Ground Truth Clusterstruktur reproduzieren lässt.

2 Clustering - Breast Cancer [4P]

Brustkrebs ist weltweit die häufigste bösartige Erkrankung bei Frauen und eine der Hauptursachen für krebsbedingte Todesfälle sowohl in Entwicklungs- als auch in Industrieländern. Verwenden Sie den Datensatz `breast_cancer.csv`. Die Merkmale werden aus einem digitalisierten Bild eines Feinnadelaspirats einer Brustmasse berechnet. Sie beschreiben Merkmale der im Bild vorhandenen Zellkerne. Das Zielmerkmal erfasst die Prognose gutartig (B) oder bösartig (M) und dient hier als Ground Truth. Achten Sie auf uninformative Features (z.B. ID) und fehlende Daten (Missing Values).

- Führen Sie eine Clusteranalyse durch, um eine etwaige Clusterstruktur zwischen gutartigen und bösartigen Zellen zu identifizieren.
- Vergleichen Sie die Genauigkeit folgender Cluster-Algorithmen:
  - k-means
  - hierarchisches Clustering
  - Modell-basiertes Clustering

  - DBSCAN
- Welches Verfahren ist am besten geeignet?
- Stellen Sie die Ergebnisse grafisch dar (Scatter Plot z.B. radius_mean vs. texture_mean).
- Lässt sich das Ergebnis verbessern, wenn vor dem Clustering der Merkmalsraum mittels PCA reduziert wird? Vergleichen Sie die Ergebnisse mit den vorherigen Resultaten.

3 Clustering - Heart Disease Patients [3P]

Verwenden Sie den Datensatz `heart_disease_patients.csv`. Der Datensatz enthält anonymisierte Daten von Patienten, bei denen eine Herzerkrankung diagnostiziert wurde. Patienten mit ähnlichen Merkmalen könnten auf die gleichen Behandlungen ansprechen, und Ärzte könnten davon profitieren, etwas über die Behandlungsergebnisse von Patienten zu erfahren, die denen ähneln, die sie behandeln. Zu diesem Zweck führen Sie bitte eine Clusteranalyse durch. Vergleichen Sie unterschiedliche Algorithmen und versuchen Sie ein bestmögliches Ergebnis zu erreichen. Begründen Sie ihre Entscheidungen. Verwenden Sie nur numerische Merkmale und achten Sie auf uninformative Features und fehlende Daten. Versuchen Sie die resultierenden Cluster zu interpretieren.