



Prof. Dr. Antony Unwin, Dr. Ali Ünlü
Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse
Institut für Mathematik
Universität Augsburg
<http://stats.math.uni-augsburg.de/>



Stochastik IV – Multivariate statistische Verfahren

Übungsblatt 8

Bearbeitung: Dienstag 19. Juni 2007, 12.15 - 13.45 Uhr
Raum 2001 T

1. Leite für $g = 2$ Gruppen und $p \geq 2$ Variablen die Entscheidungsregel der quadratischen Diskriminanzanalyse (unter Einbeziehung von Kosten für Fehlklassifikationen und beliebigen a-priori Wahrscheinlichkeiten) auf der Folie Seite 200 ab. Gehe hierbei idealerweise, Schritt für Schritt, nach der Musterlösung zu Aufgabe 1 von Übungsblatt 7 vor. Verstehe, übertrage und verallgemeinere Letztere an entsprechenden Stellen.
2. Betrachtet werden wie in Aufgabe 4 von Übungsblatt 7 die beiden Variablen *stearic* und *oleic* des **Olive Oils** Datensatzes.
 - (a) Parametrisiere ein hinreichend feines Raster auf der Scatterplot-Ebene der Wertepaare der beiden Variablen.
 - (b) Erstelle mit der Diskriminanzfunktion in Aufgabe 4 von Übungsblatt 7 (der linearen Diskriminanzanalyse) für dieses Raster die Gruppenzugehörigkeitsvorhersagen. Es soll dabei das Streudiagramm (mit den farblich markierten Regionen und 95%-Konfidenzellipsen) um die entsprechend eingefärbten Entscheidungsbereiche erweitert werden (vgl. Folie Seite 193).
3.
 - (a) Führe für die Variablen und Daten in Aufgabe 4 von Übungsblatt 7 eine quadratische Diskriminanzanalyse (mittels der Funktion `qda` im Paket MASS in R) durch, zum einen ohne 'Training' und 'Test' (im Allgemeinen nicht empfehlenswert) und zum anderen über 'leave-one-out' Kreuzvalidierung.
 - (b) Erstelle die resultierenden Konfusionsmatrizen der Gruppenzugehörigkeitsvorhersagen. Berechne die Prozentsätze korrekter Klassifikationen und vergleiche diese zusammenfassenden 'Gütemaße' mit dem Prozentsatz korrekter Klassifikationen, wenn jeder Fall derjenigen Gruppe mit größtem Umfang in der Gesamtstichprobe zugeordnet wird.
 - (c) Visualisiere die Entscheidungsbereiche für die quadratische Diskriminanzanalyse wie in Aufgabe 2 von Übungsblatt 8.
4. Betrachtet wird der **Crabs** Datensatz.
 - (a) Berechne aus den 5 morphologischen Messvariablen 4 Cluster mittels `kmeans` in R. Erstelle eine Scatterplot-Matrix der 5 Variablen, in der die erhaltenen Cluster farblich hervorgehoben sind; kennzeichne hierbei jede der vier Gruppen ('Blue Male', 'Blue Female', 'Orange Male' und 'Orange Female') mit unterschiedlichen Symbolen.
 - (b) Nehme als Grundlage nicht die Daten selbst, sondern die zweite und dritte Hauptkomponente der 5 Messvariablen, und führe erneut eine 'k-means' Cluster Analyse mit 4 Clustern durch. Erstelle eine Scatterplot-Matrix der 5 Variablen, in der die nun erhaltenen Cluster farblich hervorgehoben sind; kennzeichne hierbei jede der vier Gruppen ('Blue Male', 'Blue Female', 'Orange Male' und 'Orange Female') mit unterschiedlichen Symbolen.

5. Betrachtet wird das in R als dist-Objekt verfügbare Beispiel **Euro Distances** (vgl. Aufgabe 5 von Übungsblatt 6). Korrigiere in diesem Beispiel die Entfernung zwischen Rom und Athen auf 1420 km.
- (a) Führe hierarchische Clusterungen mittels 'Average Linkage' und 'Ward's Method' durch, und plote diese in jeweiligen Dendrogrammen.
 - (b) Plote die Punkte (inklusive der Städtenamen) einer klassischen multidimensionalen Skalierung in $k = 2$ Dimensionen, und färbe hierbei die Punkte nach der 4-Cluster Lösung des 'Average Linkage' ein.