Prof. Dr. E. Spodarev
Judith Olszewski

SoSe 2015
11.06.2015

Statistik-Praktikum/WiMa-Praktikum II - Übungsblatt 7

Vorstellung der Ergebnisse in der Übung am 18.06.2015

Aufgabe 1

a) Betrachte die Daten in der SAS-Datei 'zns.sas7bdat'. Die Variable IND enthält Informationen darüber, ob eine untersuchte Person gesund (G) oder krank (K) war. Bei der Untersuchung wurden zudem die Zellarten Nervenzelle (N), Astrozyt (A), Oligodendrozyt (O), Mikroglia (M) und Glia (G) ausgezählt und die Quotienten AN=A/N, ON=O/N, M/N etc. in der Datei gespeichert. Kann man bereits anhand der Mittelwerte der drei gemessenen Quotienten A/N, O/N und M/N (getrennt für Gesunde und Kranke) eine Zuordnung in gesunde bzw. kranke Personen erkennen?

b) Erzeuge eine graphische Veranschaulichung der Merkmale A/N, O/N, M/N für die 98 Datensätze von Gesunden und Kranken. Verwende die Prozedur G3D um einen Scatterplot in 3 Dimensionen zu erstellen. Wähle für Gesunde und Kranke verschiedene Symbole und Farben. Lässt sich bereits eine Trennungseigenschaft feststellen?

c) Anhand dieser Lernstichprobe will man auch in Zukunft Patienten aufgrund des Merkmalsvektors (A/N, O/N, M/N) als krank bzw. gesund diagnostizieren. Führe deshalb eine lineare Diskriminanzanalyse (DA) mit Bayes-Entscheidungsregel zu diesen ZNS-Daten durch. Gehe von normalverteilten Merkmalen aus und benutze die gepoolte Kovarianzmatrix und geschätzte a priori Wahrscheinlichkeiten (Prozedur DISCRIM). Interpretiere die Ergebnisse.

d) Eine andere Möglichkeit der DA bieten sogenannte nichtparametrische Verfahren. Die Datei 'kristall.sas7bdat' ist eine erweiterte Fassung der bekannten Kristalldaten. Erzeuge einen 3-dim. Scatterplot der Größen pH-Wert, Calcium-Konzentration und spezifisches Gewicht (G) mit verschiedenen Symbolen und Farben für kristallbildend bzw. nicht kristallbildend. Führe danach eine nichtparametrische, auf Dichteschätzern basierende DA unter Verwendung des Epanechnikov-Kerns mit Bandbreite $r = 1,2$ durch. Interpretiere auch hier deinen Output.

e) Versuche das eher schlechte Ergebnis in Aufgabe 1c) (19 fehlklassifizierte Gesunde) zu interpretieren und durch einen eigenen Versuch (nichtparametrische DA mit Epanechnikov-Kern und Bandbreite $r = 1,2$) zu verbessern.

f) Analysiere die Daten der Testdatei 'zns2.sas7bdat' sowohl mittels einer parametrischen als auch einer nichtparametrischen DA. Verwende jeweils die Datei 'zns.sas7bdat' als Lernstichprobe. (Tipp: Option TESTDATA)

Bitte zweite Seite beachten

Aufgabe 2

Die Datei 'zufall.sas7bdat' enthält 140 Realisierungen von unabhängigen Zufallsvektoren, die von verschiedenen zweidimensionalen Normalverteilungen erzeugt wurden.

a) Erzeuge einen Plot dieser zweidimensionalen Daten.

b) Führe eine Clusteranalyse zur Bildung von 5 Clustern mittels der Prozedur CLUSTER mit einem Density-Linkage Verfahren (Nearest-Neighbor mit Parameter $K = 8$) für diese Daten durch. Benutze das unten beschriebene Vorgehen um das Ergebnis zu visualisieren. Zu welchem Schluss kommst du?

Hilfestellung:

- Beispiel zur Prozedur G3D

  Die folgenden Anweisungen erzeugen einen Scatterplot, bei dem entsprechend dem Wert der Variablen var ein anderes Symbol für den Scatterplot gewählt wird. Versuche selbstständig auch eine Farbunterscheidung (die globale Option SYMBOL funktioniert in diesem Zusammenhang nicht).

```
DATA...;
    SET ...;
    IF var='Wert' THEN DO
    shapev='PYRAMID';
    END;
    ELSE DO
    shapev='STAR';
    END;
RUN;

PROC G3D DATA=...;
    SCATTER an*on=mn /SHAPE=shapev;
RUN;
```

- Beispiele zur Prozedur DISCRIM

  Parametrische DA mit Bayes-Regel wird durchgeführt.

```
PROC DISCRIM DATA=datei1 METHOD=NORMAL POOL=YES;
      CLASS...
      VAR...;
      PRIORS PROP;
RUN;
```

Lernstichprobe: datei1 zu analysierende Daten: datei2

```
PROC DISCRIM DATA=datei1 TESTDATA=datei2 METHOD=...;
      CLASS...
      VAR...;
RUN;
```

Nichtparametrische DA unter Verwendung des Epanechnikov-Kerns mit Bandbreite 1,2.

```
PROC DISCRIM DATA=datei1 METHOD=NPAR KERNEL=EPA R=1,2;
      CLASS...
      VAR...;
      PRIORS PROP;
RUN;
```

- Die Prozedur CLUSTER

  Das Ergebnis der Prozedur CLUSTER kann unter Benutzung der Prozedur TREE wie folgt veranschaulicht werden:

```
PROC CLUSTER DATA=... METHOD=.... OUTTREE=baum;
    VAR ....;
RUN;
```

  Plotten eines Dendrogramms und Berechnung des Outputs der Clusteranalyse bei k Clustern

```
PROC TREE DATA=... NCL=k  OUT=out;
    COPY x1 x2;
RUN;
```

  Der Datensatz out kann dann mit den Prozeduren SORT, PRINT und GPLOT nach Clustern geordnet, ausgegeben und visualisiert werden, z.B

```
PROC GPLOT DATA=out;
    PLOT x2*x1=cluster;
RUN;
```