WMuEMaL

Friedrich-Schiller-Universität Jena

Fakultät für Mathematik und Informatik
INSTITUT FÜR INFORMATIK
Prof. E.G. Schukat-Talamazzini

WERKZEUGE MUSTERERKENNUNG & MASCHINELLES LERNEN

Aufgabenblatt 7

(Ausgabe am Fr 26.5.2017 — Abgabe bis So 11.6.2017)

Aufgabe 1 — 10P

In dieser Aufgabe geht es um die Fisher-Diskriminanten (Lineare Diskriminanzanalyse nach Kitano, ME-Skript V.6).

(a) Laden Sie wieder den Irisdatensatz, lesen Sie die sechs Datensätze aus `fda.rda` *(⇝ Aufgabenwebseite) ein und verwenden Sie Ihre alte (korrigierte) Funktion* `plot_lfd` *(Übung 6, Aufgabe 2).*

(b) Schreiben Sie eine Funktion `class_scatter(x,f)`*, die für den mit Faktor* `f` *etikettierten Datensatz* `x` *den Mittelwertvektor $\boldsymbol{\mu}$ und die drei Streuungsmatrizen $\boldsymbol{S}$, $\boldsymbol{S}_W$, $\boldsymbol{S}_B$ (total, Inner- und Außerklassen) berechnet und in einer Liste mit den Einträgen* `mean`*,* `total`*,* `within`*,* `between` *zurückliefert.*

(c) Erweitern Sie `class_scatter()` *um einen Test ('R'-Funktion* `stopifnot`*) auf die Gültigkeit der Zerlegung $\boldsymbol{S} = \boldsymbol{S}_W + \boldsymbol{S}_B$ und korrigieren Sie nötigenfalls die Kovarianzberechnung; lesen Sie dazu bitte* `?cov` *durch.*

(d) Schreiben Sie nun eine Funktion `FDA (x, train=x, n=)` *zur FDA-Transformation des Datensatzes* `x`*. Es sind die* `n` *ersten Diskriminanten zu berechnen. Wählen Sie Kitanos Kernmatrix $\boldsymbol{S}_W^{-1}\boldsymbol{S}_B$ und verwenden Sie die Funktionen* `class_scatter()` *und* `eigen()`*; alles weitere wie bei* `PCA()`*.*

(e) Erweitern Sie `FDA()` *um ein Argument* `method=c('FDA','PCA','BSA','ID')` *für die Alternativen* `PCA` *(gewöhnliche PCA) und* `BSA` *('between-scatter' Analyse), welche als Kernmatrix der Transformation $\boldsymbol{Q} = \boldsymbol{S}$ bzw. $\boldsymbol{Q} = \boldsymbol{S}_B$ statt $\boldsymbol{Q} = \boldsymbol{S}_W^{-1}\boldsymbol{S}_B$ (im Fall* `FDA`*) zu Grunde legen. (Bei* `'ID'` *entfällt das Transformieren.)*

(f) Starten Sie nun für jeden der sieben Datensätze eine (2×2)-Leinwand und zeichnen Sie den Scatterplot für die jeweils beiden ersten

(1) Originalmerkmale, (2) PCA-Merkmale, (3) BSA-Merkmale, (4) FDA-Merkmale.

(g) Datensatz `ALDI` *enthält Personen zweier Musterklassen (± `jena`) mit ihren Erwerbshäufigkeiten einschlägiger Konsumartikel als Merkmale. Nutzen Sie einen geschickten* `FDA()`*-Aufruf um herauszubekommen, welche fünf der zwanzig gelisteten Produkte die verläßlichsten Indikatoren für den Wohnsitz Jena sind.*

Abzugeben ist die Datei `fda.R` *mit Ihrem Programmcode und Ihre schriftliche Antwort zu (g).*

Aufgabe 2 — 10P

Laden Sie das 'R'-Paket `class` *mit dem Kommando* `library(class)` *und lesen Sie sich die Beschreibung zu den Methoden* `knn` *und* `knn.cv` *des Nächste-Nachbarin-Klassifikators (ME-Skript VI.6, Blatt 14,15) durch.*

(a) Schreiben Sie eine 'R'-Funktion `knn.heldout (train, test, k=1)`*, die einen Klassifikatortest mit den angegebenen Lern- und Testdaten durchführt und dabei die k-Nächster-Nachbar-Regel der Funktion* `knn` *verwendet; der Ausgabewert sei die Fehlerrate.*

(b) Schreiben Sie eine 'R'-Funktion `knn.leave1out (data, k=1)`*, welche einen Kreuzvalidierungstest („leave-one-out") durchführt; der Ausgabewert sei wiederum die Fehlerrate.*

(c) Laden Sie die `diabetes.rda`*-Daten und führen Sie die k-NN-Klassifikation durch (Lerndaten ≠ Testdaten). Verwenden Sie die Befundgrößen $k \in \{1, 2, 3, 5, 7, 10, 14, 19, 25\}$ und schreiben Sie alle Fehlerraten (in Prozent!) in eine erste Tabellenzeile.*

(d) Vertauschen Sie nun die Rolle von Lern- und Testdatensatz und füllen Sie die nächste Tabellenzeile. Wie deuten Sie die Ergebnisse?

(e) Gehen Sie jetzt zur Kreuzvalidierung über (dritte Zeile); verwenden Sie dazu die Vereinigungsmenge von Lern- und Testdatensatz. Was gibt es Auffälliges zu berichten?

(f) Wiederholen Sie die Kreuzvalidierung zweimal (vierte und fünfte Zeile). Wo und warum unterscheiden sich die Resultate?

(g) Laden Sie jetzt Lern- und Testdatensatz aus `letter.rda` *und evaluieren Sie die 1-NN-Regel, wobei Sie stets den Testdatensatz zur Fehlerwertung verwenden, aber aus den Lerndaten Anfangspartien wechselnder Längen zwischen 1 und 8192 Mustern (T ∈* `2^(0:13)`*) zum Anlernen selektieren.*

(h) Analysieren Sie jetzt den Gesamtdatensatz (Lern+Test) aus `germany.rda`*. Ermitteln Sie bitte für jedes $x_1, \ldots, x_{24}$ die Fehlerrate (1-NN-Regel/Kreuzvalidierung), die sich mit den restlichen 23 Merkmalen — nach Entfernung des aktuellen x_n — ergibt (Knockoutrate).*

(i) Welches Merkmal ist am unverzichtbarsten? Welches Merkmal ist am überflüssigsten? Sind die Unterschiede dieser minimalen/maximalen Fehlerrate zu derjenigen des ***vollständigen*** *Merkmalinventars eigentlich statistisch signifikant?*

Abzugeben sind der R-Code `knnrule.R`*, eine Datei* `errors.rda` *mit den drei Fehlerratentabellen zu (c,d,e,f), (g) und (h) als gesavete Matrixobjekte (mit informativen Beschriftungen in* `colnames` *und* `rownames`*) und Ihre Kommentare zu (d,e,f,i) als* ***schriftlicher*** *Lösungsteil.*

Hinweise zum Übungsablauf

- Die wöchentliche WMM-Vorlesung findet am Mittwoch um 12:15 Uhr statt.
  Das Aufgabenblatt gibt es immer am Freitag (PDF im Netz).
  Der späteste Abgabetermin ist Sonntag 23:59 Uhr.
- Die Übungsaufgaben dürfen natürlich (und sollten sogar) in Gruppenarbeit (2 Mitglieder) gelöst werden.
- Schriftliche Lösungen (*„Textantworten“*) sind als PDF beizufügen oder direkt im e-Mail-Textkörper unterzubringen.
- Alle anderen Lösungen (Programmieraufgaben, Daten und Grafiken) sind als elektronischer Anhang der Lösungs-e-Mail abzuliefern.
- Programmcode (Dateien `*.R`) muss auch wirklich in 'R' ausführbar sein.
  (Kommando `Rscript «name.R»` auf einem der Rechner des FRZ-Pools)
- Ganz wichtig:
  Schriftliche Antworten werden von mir gedruckt, gelesen, kommentiert und korrigiert. Deshalb diese Textteile bitte **niemals** im abgegebenen Programmcode verstecken!
- Je Gruppe und je Aufgabenblatt ist **genau eine** e-Mail zu senden:
  - Vermerk »**WMM**/n« und Gruppenname im `subject`-Feld
    ($n \in \mathbb{N}$ ist die laufende Nummer des Übungsblattes)
  - die Namen der beteiligten Gruppenmitglieder im Textrumpf
  - Tabellen, Bilder, Programmcode, Sensordaten als Attachments
    (elektronische Anlagen)
  - etwaige schriftliche Antworten im Textrumpf der Post oder als Attachment
    (Text/PDF)
- *Pfingstfrieden:* Am Freitag 2.6. gibt es kein Übungsblatt. Die Lösungen für das Übungsblatt vom Freitag 26.5. müssen erst am Sonntag 11.6. abgeliefert werden.
- Einige Aufgabentexte verweisen Sie zum Nachschlagen von Details auf das Folienskript zur Vorlesung Mustererkennung; Sie finden es unter der URL
  `http://www.minet.uni-jena.de/fakultaet/schukat/ME/Scriptum/`.
  Die Angabe *ME-Skript II.6* bedeutet: Kapitel II, Abschnitt 6