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 31.5.2019 — Abgabe bis So 9.6.2019)

Aufgabe 1 — 10 P

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.ldf` *(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* `fisher (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* `fisher()` *um ein Argument* `method=c('FDA','PCA','BSA','orig')` *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* `'orig'` *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) die beiden ersten FDA-Merkmale, (3) PCA-Merkmale, (4) BSA-Merkmale.

(g) *Datensatz* `mafia` *enthält Personen zweier Musterklassen (±* `Ndrangheta`*) mit ihren Erwerbshäufigkeiten einschlägiger Konsumartikel als Merkmale. Nutzen Sie einen geschickten* `fisher()`*-Aufruf um herauszubekommen, welche fünf der zweiundzwanzig gelisteten Produkte die verläßlichsten Indikatoren für die Mitgliedschaft im organisierten Verbrechen sind.*

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

Aufgabe 2 — 10 P

Wir implementieren Lern- und Testphase eines einfachen statistischen Klassifikators — der naiven Bayesregel mit klassenweise normalverteilten Merkmalen (ME-Skript VI.4 und VII.2).

(a) ***Lernphase:*** *Die Konstruktorfunktion* `naivegauss(x)` *erwartet einen Lerndatensatz* `x` *(Klasse* `data.frame`*) mit der Etikettierung (Klasse* `factor`*) in letzter Position. Sie erzeugt ein Listenobjekt der Klasse* `naivegauss`*, das alle nötigen Informationen zur Klassifikation enthält, also z.B. die Klassenwahrscheinlichkeiten und die gelernten Normalverteilungsparameter.*

(b) ***Abrufphase:*** *Die Funktion* `predict.naivegauss(o,newdata)` *erwartet ein Listenobjekt* `o` *der Klasse* `naivegauss` *sowie einen Testdatensatz* `newdata` ***ohne*** *Etikettierung. Sie retourniert einen Faktorvektor, der zu jedem Eingabemuster (Zeilenvektoren von* `newdata`*) die geratene Klasse enthält.*

HINWEIS: Stellen Sie sicher, dass `predict` *auch unter Extrembedingungen (Datensätze mit einem Merkmal und/oder einem Muster) funktioniert!*

(c) ***Fehlertest:*** *Die Funktion* `heldout (x, newdata=x, method, ...)` *erwartet je einen etikettierten Lern- und Testdatensatz. Sie lernt aus* `x` *und klassifiziert damit* `newdata`*. Dabei verwendet sie das Klassifikationsverfahren, das in der 'R'-Klasse* `method` *(mit gleichnamigem Konstruktor, dem wir auch* `...` *weiterleiten) implementiert ist. Nach Vergleich mit den wahren Klassenzugehörigkeiten der Testmuster liefert sie die (geschätzte) Fehlerwahrscheinlichkeit als Rückgabewert. Diesem* `numeric[1]`*-Objekt sei als Attribut (Name:* `confused`*) die Matrix der absoluten Klassenverwechslungshäufigkeiten beigefügt.*

(d) *Laden Sie die Iris-Daten und starten Sie* `heldout (iris, iris, naivegauss)`*. Die Reklassifikationsfehlerrate sollte* 4 *Prozent (*6/150*) betragen.*

(e) *Lesen Sie die Datensätze* `diabetes.lern` *und* `diabetes.test` *ein. Starten Sie alle vier möglichen Aufrufkombinationen (Lern/Test) von* `heldout()` *für diese Daten. Erklären Sie, inwiefern die Größenrelationen zwischen den Fehlerraten der vier* `diabetes`*-Läufe exakt Ihren Erwartungen entsprechen (ME-Skript VI.7).*

(f) ***Kreuzvalidierung:*** *Schreiben Sie eine Funktion* `leave1out (x, method, ...)`*, welche die „*leave-one-out*“-Fehlerrate eines Datensatzes* `x` *berechnet. Wie* `heldout` *soll auch* `leave1out` *für jeden syntaktisch wie* `naivegauss` *ausgelegten Klassifikatortyp* `method` *anwendbar sein. Wie groß ist der* L^1O*-Fehler für die Iris-Daten? (Tipp:* 7/150*) Und für die (Gesamtheit der)* `diabetes`*-Daten?*

Abzugeben sind die Datei `naivegauss.R` *mit dem Programmcode sowie schriftlich je* 4 *Fehlerraten zu (e) und (f) und der Kommentar zu (e).*

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)
- 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