WMuEMaL

Friedrich-Schiller-Universität Jena

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

WERKZEUGE MUSTERERKENNUNG & MASCHINELLES LERNEN

Aufgabenblatt 8

(Ausgabe am Fr 8.6.2018 — Abgabe bis So 17.6.2018)

Aufgabe 1 — 8 P

In dieser Aufgabe geht es um die Anpassung von Ausgleichspolynomen (in einer Veränderlichen) mit der 'R'-Funktion `lm()` *durch lineare Regression (ME-Skript VI.5).*

(a) Laden Sie den Datensatz `xydata` *(#1=Quellvariable und #2=Zielvariable) aus der Datei* `limo.rda` *und lesen Sie die Dokumentation zu* `lm()` *und* `formula`*-Objekten.*

(b) Schreiben Sie eine Funktion `polyfun(x,a)`*, die für alle Einträge des Vektors* `x` *den Funktionswert des Polynoms mit Koeffizienten* `a` *(aufsteigend als* $a_0, \ldots, a_n$ *gespeichert) berechnet und als Vektor zurückgibt.*

(c) Schreiben Sie eine Funktion `polyfit(xy,n)`*, die für den Datensatz* `xy` *ein Ausgleichspolynom* `n`*-ten Grades zur Vorhersage des zweiten aus dem ersten Attribut berechnet und als Regressionsobjekt (Klasse* `lm`*) zurückliefert. Rufen Sie dazu* `lm()` *mit einer geeigneten Modellformel auf.*

(d) Schreiben Sie eine Funktion `polyfits(xy,order,plot=FALSE)`*, die eine Liste der Ausgleichspolynomobjekte für* `xy` *zu allen Polynomgraden in* `order` *abliefert. Im Fall* `plot=TRUE` *extrahiert sie das Akaike- und das Bayes-Informationskriterium aller Modelle (Funktionen* `AIC()` *und* `BIC()`*) und trägt die Werte in einer gemeinsamen Grafik über den Polynomgraden auf. Führen Sie nun* `polyfits(xydata,0:11,plot=TRUE)` *aus; welcher Grad ist der Gewinner?*

(e) Zeichnen Sie nun auf drei (2×2)*-Leinwänden für jedes Polynom ihrer Liste eine Grafik mit (1) der Punktwolke aus dem Datensatz* `xydata`*, (2) dem Funktionsverlauf des gefitteten Ausgleichspolynoms und (3) dem BIC-Wert und (4) den Polynomkoeffizienten (sinnvoll verknappt) als Texteintrag.*

(f) Wiederholen Sie die Grafikaufrufe aus Teil (d) und (e) für den klassischen Datensatz `cars` *(Wissenswertes zu dieser Erhebung mit* `?cars`*).*

(g) Wiederholen Sie diese Grafikaufrufe für den fünfspaltigen `LifeCycleSavings`*-Datensatz, und zwar für alle sechs Kombinationen der drei Attribute* `pop15`*,* `pop75` *und* `dpi`*.*

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

Aufgabe 2 — 12 P

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, für deren etwas hausbackene Schnittstelle wir im Folgenden einige einfache Hüllfunktionen schreiben werden.*

(a) *Schreiben Sie eine 'R'-Konstruktorfunktion* `kNN(x,neighbors=1)` *für einen k-NN-Regel-Klassifikator mit Lerndaten* `x` *und* `neighbors` *nächsten Nachbarn. Rückgabe ist ein Listenobjekt der Klasse* `kNN` *mit den benötigten Daten und Parametern.*

(b) *Schreiben Sie eine 'R'-Prädiktorfunktion* `predict.kNN(o,newdata)`*, welche die Zeilenvektoren der Datenmatrix* `newdata` *(Matrix oder Dataframe; ohne Faktor!) mit der k-NN-Regel* `o` *klassifiziert. Rückgabe ist der Klassenfaktor.*

(c) *Erweitern Sie* `predict.kNN()`*, so dass bei Aufruf mit* `newdata=NULL` *die Leave-One-Out-Klassifikation der Lerndaten des* `o`*-Objekts berechnet wird. Konsultieren Sie* `?knn.cv`*.*

(d) *Reanimieren Sie die Auswertefunktion* `heldout(x,newdata=x,method,...)` *vom letzten Aufgabenblatt und modifizieren Sie ihren 'R'-Kode, so dass bei Aufruf mit* `newdata=NULL` *die Leave-One-Out-Fehlerrate des* `method`*-Klassifikators für die* `x`*-Daten berechnet wird.*

(e) *Programmieren Sie einen Testlauf* `run.1st(x,y,choice=1+2*0:9)`*, der eine dreizeilige Matrix von Fehlerraten erzeugt. In Spalte* `j` *wird die k-NN-Regel mit* `choice[j]` *Nachbarn getestet. In Zeile 1 wird* `x` *zum Lernen und* `y` *zum Testen genutzt. In Zeile 2 werden die Rollen von* `x` *und* `y` *getauscht. In Zeile 3 wird die Leave-One-Out-Fehlerrate für die Vereinigungsmenge von* `x` *und* `y` *ermittelt.*

(f) *Programmieren Sie einen Testlauf* `run.2nd(x,y,choice=2^(0:13))`*, der einen Vektor von Fehlerraten erzeugt. In Komponente* `j` *stehe die Fehlerrate der* 1*-NN-Regel mit Testdaten* `y` *und den ersten* `choice[j]` *Mustern von* `x` *zum Lernen.*

(g) *Programmieren Sie einen Testlauf* `run.3rd(x,choice=2:ncol(x)-1)`*, der einen Vektor von Fehlerraten erzeugt. In Komponente* `j` *stehe die Leave-One-Out-Fehlerrate der* 1*-NN-Regel für die Daten* `x`*, wobei alle Attribute außer einem — dem* „Knock-out“*-Attribut* `choice[j]` *— als Merkmalsatz zur Klassifikation genutzt wurden.*

(h) *Laden Sie jetzt die drei Datensätze* `diabetes`*,* `letter` *und* `germany` *aus den* `*.rda`*-Dateien und führen Sie damit (in obiger Reihenfolge zugeordnet) die drei Testreihen durch. Für die* `germany`*-Studie werden Lern- und Testdatenteil vereinigt und an* `x` *übergeben. Speichern Sie die drei Fehlertabellen mit* `save(pe.1,pe.2,pe.3,file='kNN.rda')` *ab.*

(i) *Erzeugen Sie abschließend vier* `barplot`*-Grafiken, zwei für die* `diabetes`*-Fehlermatrix und je eine für die beiden Fehlervektoren zu* `letter` *und* `germany`*. Gestaltungsvorschläge siehe Ausgabebeispiel* `kNN-bsp.djvu`*. Für eine ansprechende Darstellung ist darauf zu achten, dass die Funktionen aus (e,f,g) informative Beschriftungen in* `colnames` *und* `rownames` *ablegen.*

Abzugeben sind der R-Code `kNN.R` *und die Fehlertabellen in* `kNN.rda`*.*

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