WMuEMaL

Friedrich-Schiller-Universität Jena

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

WERKZEUGE MUSTERERKENNUNG & MASCHINELLES LERNEN

Aufgabenblatt 6

(Ausgabe am Fr 19.5.2017 — Abgabe bis So 28.5.2017)

Aufgabe 1 — 10P

Diese Aufgabe behandelt die maschinelle Gruppierung landessprachlicher Texte nach einem informationstheoretischen Distanzkriterium (Skript[1] „Stochastische Grammatikmodelle“ *VIII.6, S. 13–15).*

(a) Laden Sie die Liste (`zip.rda`) der Zeichenkettenvektoren von 43 Übersetzungen des UDHR-Dokuments (Menschenrechtedeklaration der UN).

(b) Schreiben Sie eine Funktion `bits(x,compress=TRUE)`, die einen Textvektor `x` mit dem GZIP*-Verfahren ('R'-Funktion `memCompress()`) komprimiert und als Ergebnis die Anzahl der erzeugten Bits (Zählen mit `nchar()`) abliefert. Für `compress=FALSE` geben Sie die Bitzahl des Originals zurück.*

(c) Erzeugen Sie eine Cleveland-Grafik (`?dotchart`) mit den absteigend sortierten Kompressionsfaktoren für alle Landessprachen.

(d) Nach Shannon benötigt ein Komprimierer $\mathcal{H}(p)$ Bits/Zeichen (Entropie), um einen p-verteilten Text x_p zu kodieren, wenn er die Verteilung p zum Verschlüsseln verwendet. Verschlüsselt er mit abweichender Verteilung q, so werden es $\mathcal{H}(p\|q)$ Bits/Zeichen (Kreuzentropie). Schreiben Sie eine Funktion `entropy(xp,xq)`, welche näherungsweise die Kreuzentropie $\mathcal{H}(p\|q)$ für die Verteilungen p und q der Texte `xp` und `xq` berechnet. Die Bitzahl einer q-Verschlüsselung von `xp` sollten Sie durch Aufrufe `bits(c(xq,xp))` und `bits(xq)` ermitteln können.

(e) Schreiben Sie den Einzeiler `divergence(xp,xq)` zur Berechnung der Kullback-Leibler-Divergenz $\mathcal{D}(p\|q) = \mathcal{H}(p\|q) - \mathcal{H}(p\|p)$ sowie die Funktion `distance(X)`, die für die Textliste `X` eine Distanzmatrix (Klasse `dist`) mit allen wechselseitigen Textdistanzen $d_{ij} = \mathcal{D}(p_i\|p_j) + \mathcal{D}(p_j\|p_i)$ (symmetrische Divergenz) erzeugt. Vergessen Sie bitte nicht die Mitnahme der Textprobenamen aus `X`.

(f) Und nun clustern Sie die Textproben, indem Sie ihre Distanzmatrix den Methoden `agnes` bzw. `diana` ('R'-Paket `cluster`) zur agglomerativen/divisiven Gruppierung übergeben und die Dendrogrammgrafiken ausgeben.

Abzugeben ist die Datei `zip.R` mit Ihrem Programmcode.

[1] URL: `http://www.minet.uni-jena.de/fakultaet/schukat/SGM/Scriptum/lect08-NLP.pdf`

Aufgabe 2 — 10P

*In dieser Aufgabe geht es um **etikettierte** Merkmaldaten, ihre graphische Darstellung und ihre Transformation nach Karhunen-Loéve (PCA, ME-Skript V.5).*

Wir stellen Merkmaldaten in 'R' als `data.frame` mit $N+1$ Spalten dar; jede Zeile entspricht einem Muster; die Spalten $1, 2, \ldots, N$ enthalten die Merkmalwerte (Typ `numeric`) und die letzte Spalte zeigt die wahre Klassenzugehörigkeit (Typ `factor`) an.

(a) Laden Sie den Irisdatensatz mit dem Kommando `data(iris)` und lesen Sie die sieben Datensätze aus `load('pca.rda')` (⇝ Aufgabenwebseite) ein.

(b) Schreiben Sie eine Grafikausgabefunktion `plot_lfd (x, subset=?, ...)` zur Scatterplotdarstellung (siehe `?plot.data.frame`) der multivariaten Datensätze. Die Punkte der Zeichnung sind nach Klassenzugehörigkeit einzufärben. Es bezeichne `x` den Datensatz (mit Klassenfaktor in der letzten Spalte) und `subset` ist ein `integer`-Vektor mit den Indizes der zu berücksichtigenden Merkmale. Ergänzen Sie einen sinnvollen Defaultwert.

(c) Testen Sie `plot_lfd()` mit den `iris`-Daten für die verschiedenen `subset`-Argumente `1:4`, `2:4`, `3:4` und `4:4`.

(d) Schreiben Sie nun eine Funktion `PCA (x, train=x, n, center=TRUE, scale=TRUE)` zur PCA-Transformation des Datensatzes `x`. Es sind die `n` ersten Hauptachsen zu verwenden; die logischen Parameter `center` und `scale` geben an, ob auch zentriert bzw. skaliert werden soll. Als Grundlage der Eigenwertberechnung diene die Kovarianzmatrix des — i.a. von `x` verschiedenen — Lerndatensatzes `train`. Rückgabeobjekt ist ein `data.frame` mit den (nach wie vor etikettierten!) transformierten Merkmalen. Brauchbare 'R'-Funktionen für die Matrizenmanipulation sind z.B. `apply`, `cov`, `eigen`, `sweep`.

(e) Starten Sie eine (2×2)-Leinwand und rufen Sie `plot_lfd (PCA (iris, n=2))` mit den vier Kombinationen für `center` und `scale` auf.

(f) Starten Sie eine (2×2)-Leinwand und rufen Sie wieder `plot_lfd (PCA (x, train, n=2))` auf. Für `x` und `train` setzen Sie wahlweise `iris` ein und den Teildatensatz `iris.part`, der lediglich die 50 `setosa`-Muster enthält.

(g) Starten Sie nun eine Schleife über die acht Datensätze (inklusive `iris`) mit je einer (2×2)-Leinwand und den vier Scatterplots für

> *(1) die beiden ersten Originalmerkmale, (2) die beiden letzten Originalmerkmale, (3) die beiden ersten Hauptkomponentenmerkmale, (4) die beiden letzten Hauptkomponentenmerkmale.*

Für Teil (f) und (g) verwenden Sie bitte stets `center=TRUE` und `scale=TRUE`.

Abzugeben ist die Datei `pca.R` mit Ihrem Programmcode.

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