WMuEMaL

Friedrich-Schiller-Universität Jena

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

WERKZEUGE MUSTERERKENNUNG & MASCHINELLES LERNEN

Aufgabenblatt 5

(Ausgabe am Fr 18.5.2018 — Abgabe bis So 27.5.2018)

Aufgabe 1 — 12 P

Es geht darum, mit der Momentenmethode den Schwerpunkt und den Neigungswinkel von zehn abgebildeten Großraumdrohnen zu bestimmen. Die benötigten Grauwertbilder finden Sie in der Datei `moment.rda`.

(a) *Schreiben Sie eine Funktion* `moment(x,plot=TRUE)`, *die zum Grauwertbild* `x` *mit Hilfe der zentralen Momente (ME-Skript IV.4, Blatt 6) die Koordinaten des Objektschwerpunkts und den Neigungswinkel des Objekts (in Grad) berechnet; Rückgabe als benamter Vektor.*
TIPP: Vektorisieren Sie beispielsweise mit Hilfe von `col()` *und* `row()`*!*

(b) *Bei Aufruf mit* `plot=TRUE` *soll das Bild* `x` *gezeichnet werden und der berechnete Schwerpunkt mit einem farbigen Fadenkreuz (z.B.* `?abline`*) markiert werden; beachten Sie den Unterschied zwischen Bildmatrix- und Grafikkoordinaten.*

(c) *Schreiben Sie den berechneten Neigungswinkel (in Grad; Nord/West/Ost entspricht* $0°/90°/-90°$*) unter das Bild (Funktion* `mtext()`*). Denken Sie unbedingt wieder an* `atan2()`*! Fügen Sie einen weiteren* `abline()`*-Aufruf hinzu, der eine Gerade durch den Schwerpunkt mit dem berechneten Neigungswinkel einzeichnet.*

(d) *Erstellen Sie nun je Beispielbild eine* (2×2)*-Leinwand mit je einem* `moment()`*-Aufruf für das Original* `x`*, für sein Negativ* `1-x` *sowie für die Binärversionen (Funktion* `binarise()` *zur Binarisierung mit* 2-means*-Schwelle finden Sie in* `moment.rda`*) von Original und Negativ.*

(e) *Für welche der vier Verarbeitungsvarianten in (d) werden Schwerpunkt und Neigungswinkel der ersten acht Objekte verläßlich berechnet? Welche Bildeigenschaft führt zum Versagen des Verfahrens bei den beiden Raketen? Welche Vorfilterung wäre geeignet, damit* `moment()` *auch hier den korrekten Neigungswinkel entdeckt?*

Abzugeben sind die Programmdatei `moment.R` *sowie Ihre drei schriftlichen Antworten zu (e).*

Aufgabe 2 — 8 P

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 aufsteigend 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 sowie die Grafikausgabe* `zip.pdf`*.*

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

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