

Statistisches Data Mining (StDM)

Woche 1

Aufgabe 1 PCA in R

Arbeiten Sie das Kapitel 10.4 Lab 1 Principal Components Analysis in ISLR durch.

Aufgabe 2 Abstimmungsverhalten der Kantone

Besorgen Sie sich den Datensatz der Volksabstimmungen in der Schweiz von der Webseite: <http://www.bfs.admin.ch/bfs/portal/de/index/themen/17/03/blank/data/01.html> dort unter **Online-Datenrecherche Volksabstimmungen (Ergebnisse Ebene Kanton seit 1866)**. Wählen Sie alle Kantone und alle Vorlagen ab 1990 aus. Wählen Sie als Ergebniss das Feld **Ja in \%/**. Alternativ können Sie auch den Datensatz 'px-x-1703030000_100.csv' verwenden, der mit den oben beschriebenen Methoden erzeugt worden ist.

- a) Laden Sie den Datensatz und bringen ihn mit der Funktion **reshape** in eine Matrixform, bei der die Kantone die Zeilen und die Abstimmungsergebnisse die Spalten sind.
- b) Verschaffen Sie sich einen Überblick über die Daten in dem Sie eine PCA durchführen, plotten Sie auch die Kantonsnamen.
- c) Ist eine zweidimensionale Darstellung angebracht? Ploten Sie dazu die **explained variance** gegen die Anzahl der Hauptkomponenten.

Aufgabe 3 Wordembedding

Visualization of a word-embedding: The data-set (`samplewordembedding.csv`) has been produced by the word2vec algorithm (see [here](#) for details.) Each of the 871 word in the corpus is assigned to a 300 dimensional vector.

- a) Load the data set and perform a PCA.
- b) Evaluate the quality of the 2 dimensional PCA plot by showing the explained variance.