Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering
IDP Institut für Datenanalyse
und Prozessdesign

**HS 2016**

**Statistisches Data Mining (StDM)**

# Woche 5

## Aufgabe 1 Lab

Read and do the excersises of chapter 10.5.2 in ILSR

## Aufgabe 2 Clustering und MDS

Die Unahnlichkeitsmatrix CD.dis im File CountriesDis.RDA (mit load() laden) enthält Unähnlichkeiten zwischen Ländern. Die Unähnlichkeiten stammen aus einer Studie, in der Studierende aufgefordert waren, paarweise die Unaehnlichkeit zwischen den 12 Ländern Belgien (BEL), Brasilien (BRA), Chile (CHI), Kuba (CUB), Ägypten (EGY), Frankreich (FRA), Indien (IND), Israel (ISR), Vereinigte Staaten (USA), Sowjetunion (USS), Jugoslawien (YUG) und Zaire (ZAI) anzugeben. In der Unähnlichkeitsmatrix sind die durchschnittlichen Unähnlichkeitsbewertung der Studierenden festgehalten.

a) Führen Sie mit dieser Unähnlichkeitsmatrix hierarchische Cluster-Analysen durch, verwenden Sie verschiedene Linkage-Methoden. Vergleichen Sie dabei die Resultate der Cluster-Methoden single, complete, average und “ward” bezüglich der Gruppenbildung

b) Führen Sie eine ordinale multidimensionale Skalierung mit zwei Komponenten durch und tragen Sie die Länder in eine 2D-Darstellung ein. Beschreiben Sie die wesentlichen Eigenheiten der Daten, wie sie in dieser Darstellung ersichtlich sind. Wie weit finden Sie darin die Resultate der hierarchischen Cluster-Analyse wieder? Scheiden Sie dazu den Baum aus a) in eine geeignete Anzahl von Gruppen und färben das MDS Ergebniss ein.

## Aufgabe 3 Clustering und MDS

In dieser Aufgabe untersuchen wir nochmals die Daten zu den letzten 20 eidgenössischen Volksabstimmungen von 1998 und 1999.

a) Führen Sie eine hierarchische Cluster-Analyse durch unter Verwendung von euklidischen Distanzen zwischen den Kantonen. Vergleichen Sie die Resultate der Cluster-Methoden Single Linkage, Complete Linkage und Average Linkage bezüglich der Gruppenbildung. Was ist jeweils eine optimale Anzahl Cluster? Wie verhält sich in den Resultaten jeweils die Schweiz (CH)? Tipps: Die Daten können Sie wie folgt einlesen:

```
X.t <- read.table(file.path(baseDir, "../PCA/px-x-1703030000_100.csv"),
                  header=T, sep=";", skip=1, stringsAsFactors = FALSE)
X <- reshape(X.t, idvar = "Kanton", timevar = "Datum.und.Vorlage", direction = "wide")
```

Zur besseren Visualisierung können Sie den Parameter `cex=0.3` setzten.

b) Führen Sie eine Cluster-Analyse mit K-means (K=3) durch.

```r
abst.km0 <- kmeans(abst.dist, centers = 3)
abst.km0$cluster # Clu
```

c) Wählen Sie die optimale Anzahl K-means-Cluster. Um zu sehen, wie stark das Ergebniss vom Startwert abhängt führen sie 5 Verschiedene Durchläufe aus. Verändern Sie auch den Parameter `nstart`, was fällt auf.

```r
getRes = function(){
  x.res <- rep(NA,16)
  # nun wird f?r jede Abstimmung die Varianz der Ergebnisse in den
  # verschiedene Kantonen berechnet - statt einer for-Schleife wird
  # hier mit apply gearbeitet
  for(i in 2:16){
    abst.km <- kmeans(abst.dist, centers = i, nstart=10)
    x.res[i] <- sum(abst.km$withinss)  # bestimme die summe der Varianzen ?ber alle Cluster
  }
  return (x.res)
}
plot(2:16, getRes()[2:16], type="b")
for (i in 1:10){
  lines(2:16, getRes()[2:16], col='gray')
}
# Schwer zu sehen, gehen wir mal von K=4 aus.
# Wählt man nstart gross, so gleichen sich die Kurven an.
```

d) [Optional] Wählen Sie ein geeignetes K, führen Sie ein K-Mean Clustering durch und färben Sie bei einem hierachischen gemäss der Gruppenzugehörigkeit ein. Tipp: http://stackoverflow.com/questions/18802519/label-and-color-leaf-dendrogram-in-r

```r
abst.km <- kmeans(abst.dist, centers = 4, nstart=100)
abst.W <- hclust(abst.dist, method="ward.D2")
plot(abst.W, labels=names, main="Ward", cex=cex.val, col=abst.km$cluster)
# TODO make coloring of nodes
```