Series 4

The first four tasks are meant as repetition. The last two tasks cover the new material.

First download the file `u7.rda` into your working directory. It contains all the datasets you will need for the first four tasks. Then download the file `dataU6.rda`. It contains all the datasets for the last two tasks. Load the files: `load("u7.rda")` and `load("dataU6.rda")`. You will require the packages `MASS`, `mice`, `mvoutlier` and `cluster`.

1. The data needed in this excercise is `dat1`. Look at the covariance matrix, correlation matrix and column means of `dat1`.

a) Which variable has the largest variance and how large is it?

b) Which pair of variables has the largest correlation?

c) Compute the Mahalanobis distance for the first data point in the data set (use the function "mahalanobis")

d) What does the Mahalanobis distance of a point indicate?

1. The distance from the center in terms of standard deviations in the direction of the point.
2. The variance in the direction of the point.
3. The distance of the point from the center in the units given in the original data set.
4. The percentage of variance captured by the point.

2. The data set `dat2` contains demographic information in 8 variables on 50 US states. Visualize this data using a stars plot (include a legend; don't worry if the legend is not perfectly readable; use the default scaling). See ?state.x77 if you want more information on this data set.

a) In Mississippi, there are two variables that are much more prominent than the other variables. What are these variables?

b) Find one other state that is in this respect like Mississipi.

c) Compare Alaska and California. Name three prominent differences between the two states.

3. We now look at the dataset `dat3`.

a) In the data set `dat3` is one severe outlier. In which row is the outlier?

b) Assume that we have data from a multivariate normal distribution. What is the distribution of the squared Mahalanobis distance of the data points from the center?

1. Normal distribution
2. Chi distribution
3. Chi-Squared distribution
4. Uniform distribution

4. The data set `dat6` contains 10 variables (columns) measuring different chemical properties of different substances (rows). Each variable uses different units.

a) Compute the principal components (use cor = TRUE). How many PCs are necessary to retain at least 80% of the total variance?

b) Make a scree plot (explained variance vs. nmb. of PCs). Between which two Components is the largest drop?

c) Make a biplot of the data. Is there any pair of variable that show a strong negative correlation?

5. Principal components find the (orthogonal) directions with the largest spread, while linear discriminants, also called canonical variables, find the (orthogonal) directions in which the groups are separated best. The goal of this excercise is to show to you, that linear discriminants are indeed much better suited for visualizing group differences in low dimensions. You will need the `MASS` package.

a) The datafiles for this excercise are `dat` (data) and `cl` (true classes).
Since the data is two-dimensional, we can easily plot it and inspect it visually: Make a plot of X1 vs. X2. What do you guess: In which direction will the first PC and the first LD be?

b) Project the points onto the first PC and investigate how well the groups are separated. In which direction is the first PC? Make a plot which is suitable for showing the quality of separation of the two groups along the first PC. [Do not scale the input data. Confirm that the data is already approximately centered around the origin (i.e., you do not need to center).]

c) Project the points onto the first LD and investigate how well the groups are separated. In which direction is the first LD? Make a plot which is suitable for showing the quality of separation of the two groups along the first LD.

d) What is the (leave-one-out) cross validation error for LDA on this data set? Do you think that LDA is useful for classification in this problem?

6. A retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. There are roughly two controls per case of CHD. The goal of this excercise is to classify with LDA ten new patients into CHD/no-CHD using only variables, that can be measured without highly trained people or a laboratory. Using CV, we will try to assess whether we should expect LDA to perform well.

- `datSA` contains samples; `clSA` contain the corresponding labels for chd (2=CHD present, 1=not present)
- `datSATest` contains info on 50 new patients which we have to predict; afterwards, they all get tested more thoroughly, so in the end we know the true CHD status, given in `clSATest`

You will need the package `MASS`.

a) Have a look at the meaning of the variables involved. SAheart is the origin of datSA, but we will later on work with datSA, cl, datSATest and clSATest.
R-hint: `library(ElemStatLearn)`, `?SAheart`

b) We exclude ldl (needs laboritory), famhist (LDA can only deal with numerical predictors), typea (not enough background information to interpret this variable).
Look at a scatterplot matrix and the summary of datSA

c) Compute LDA and project datSA onto the first (and in this case only) LD. Visualize the result. Can you detect any separation of the two groups?

d) Now we want to assess how confident we can be in predictions of LDA on new patients. For this, we use (leave-one-out) CV. What error rate would you expect for new samples from the same population?

e) Make a prediction for the 50 new patients (datSATest). You can check the predictions using the true labels (clSATest). What error rate do you observe? Does it agree with the CV estimate?

Preliminary discussion: 22.04.13.

Deadline: No hand-in.