Exercise Lesson01: PCA

Machine Learning using R

Exercise 1: Wisconsin breast cancer data

The Wisconsin data set describes features computed from digitized images of fine needle asprirates (FNA) of breast mass. In addition to the extracted features, the data set includes an ID number of the image and the diagnosis of the sample (M = malignant, B = benign). For more information see https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.

a) Load the data "breastCancer_Wisconsin.csv" using the `read.csv('breastCancer_Wisconsin.csv', stringsAsFactors=TRUE)` command. The data file can be found in the folder of this exercise. Is there a variable which is inappriopriately coded? Recode the variable into the correct format.

b) How many variables are there in total in the data? How many observations?

c) Perform a Principal Component Analysis (with scaling) on the data. Do not include the `id` factor in the PCA. Also exclude the `diagnosis` factor from the PCA. Why can't we include these factors in the PCA? (**Hint**: `prcomp(..., scale.=TRUE)`)

d) Check the results of the PCA. What proportion of the variance is explained by the first PC? What proportion of the variance is explained by the first four PCs together?

e) Which of the variables contributes mostly to the first PC (in absolute terms)? How large is this loading? (**Hint**: `which.max()`, `abs()`, `pca_object$rotation`)

f) What are the coordinates of the first picture (first row) expressed in PC-coordinates? Only give the coordinates of the first three PCs. (**Hint**: `pca_obj$x`)

g) Now perform a dimensionality reduction by looking only at the first two PCs. Plot the values of the two first PCs in a scatterplot and make it so the points' colour represents the diagnosis of the pictures. What can we observe in the plot?

h) We now want to see how the 2D representation changes when no scaling is used. Perform the PCA again but without scaling and create the 2D plot of the first two PCs. What can you observe?

i) Using the PCs generated with scaling, we wish to find out how many PCs are needed to explain the data well. Create a scree-plot and decide how many PCs should be chosen.

j) In order to make the situation a bit easier to oversee, we perform the PCA again but only with the first five variables (again excluding `id` and `diagnosis`, use scaling). Create a biplot of this PCA, showing the projections of the original variables. Which of the variables has the strongest contribution to PC2?