Home Exercise 2

Statistical Learning (AMI22T)

[This exercise consists of 20 points. In order to get full points, you have to answer all the questions, below. Your answer must be clearly written, well structured, and easy to follow. Your methods and models should be appropriately motivated. A cover page to your solution should make clear of your identifier, e.g. name, e-mail etc. The length of your solution should not exceed 12 pages, Times New Roman fonts with font size 12 and single line space. Upload your solution on Learn by June 1, 2020; 09:00 CET. For any further questions, do not hesitate to contact Ilias Thomas, e-mail: ith@du.se]

The Data Cortex Nuclear data set (on Learn) concerns the expression levels of 77 proteins, detectable in the nuclear fraction of cortex of mice. There are 38 control mice and 34 mice with Down syndrome in the dataset, which are further divided into 4 categories each (8 categories total, 4 for the control mice and 4 for the Down syndrome mice). A detailed data description is available at https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression#. There are numerical and categorical variables on the dataset. After handling the missing values appropriately, answer the following questions.

1. a) Use the 77 proteins as predictors for decision trees and support vector machines models to make binary and multiple class classification.
   b) Perform principal component analysis on the 77 numerical features. Use an appropriate number of principal components as predictors and perform the same classification task.
   c) Using bagging, random forest, and boosting perform the same classification task. Compare the results of the three methods.

2. Use the dataset to perform clustering. You should try both k-means clustering and hierarchical clustering. In every case, find a number of clusters that make sense and try to explain what each cluster describes.