MAD 2020-21, Assignment 5

Bulat Ibragimov, Kim Steenstrup Pedersen

hand in until: 05.01.2021 at 23:59

General comments: The assignments in MAD must be completed and written individually. You are allowed (and encouraged) to discuss the exercises in small groups. If you do so, you are required to list your group partners in the submission. The report must be written completely by yourself. In order to pass the assignment, you will need to get at least 40% of the available points. The data needed for the assignment can be found in the assignment folder that you download from Absalon.

Submission instructions: Submit your report as a PDF, not zipped up with the rest. Please add your source code to the submission, both as executable files and as part of your report. To include it in your report, you can use the `lstlisting` environment in LaTeX, or you can include a "print to pdf" output in your pdf report.

Exercise 1 (16 points. Classification). *Please check the corresponding jupyter notebook file with the assignment code and examples.*

An iris flower database is given in assignment files. There are 3 iris type, and each iris flower is characterized with 4 attributes. Please read the database using functions from the jupyter notebook and perform the following tasks:

a) *Use PCA code from previous assignment to convert data from 4D to 2D by preserving only 2 most representative eigenvector. If you did not manage to implement PCA, you can simply use first 2 dimensions in the database. Note that using PCA will give you 4 points, whereas taking 2 random dimensions will give you only 1 point.*
b) *Implement kNN classification algorithm and apply it to classify iris database (2 points).*
c) *Try different values of k = 1, 2, 3, 4, 5 and print prediction accuracy for validation set. Please select optimal value of k and justify your selection (3 points).*
d) *Implement random forest classification and apply it to classify iris database. Use any reasonable parameters you want (3 points).*
e) *Visualize results of kNN and random forests. Visualization should be similar to Figure 8 in https://towardsdatascience.co a-simple-classification-problem-with-python-fruits-lovers-edition-d20ab6b071d2. Please take care that three things are correctly visualized in the plot, a) coordinates of the datapoints, b) reference labels of datapoints shown as the color of visualization points, and c) classification decision areas shown as different colors. Make colors of reference labels to be similar to the decision areas, but note that reference labels do not necessary always agree with decision boundary colors (4 points).*