CA3 - Work for a secret customer

Lever oppgave

Leveres innen Mandag innen 13:00 **Poeng** 0
Må leveres en filopplasting **Filtyper** pdf, py, og ipynb

Background

You are a data scientist employed by the new startup company "Kristian & Oliver AI". A new customer wants our startup company to analyse some highly confidential data set. Since the data are so confidential the customer provides no additional information. All you know is that the customer wants you to train simple classifiers that give the best possible predictions for the test data.

Kaggle competition

You will participate in an In-Class Kaggle competition (only students enrolled in DAT200 course are allowed to participate), where you will compete with other fellow students on how well your model predicts the test data.

Link to our In-class competition on Kaggle:

DAT200-CA3-2020

https://www.kaggle.com/c/dat200-ca3-2020/
(https://www.kaggle.com/c/dat200-ca3-2020/)

Here is the link to **enter** the competition (please don't forward this link to others outside our DAT200-course, since we want to keep the competition In-class) https://www.kaggle.com/t/986f2002a23c43309ad956d684ddaec0 (https://www.kaggle.com/t/986f2002a23c43309ad956d684ddaec0)

On the data website (https://www.kaggle.com/c/dat200-ca3-2020/data) of the In-Class competition you will find three files:

- training data
- test data
- sample submission

Rules and context of the competition

1. Please find the suggested workflow in file 'DAT200_CA03_workflow.pdf'. Use the workflow as a guide for how to train your classifiers, make predictions and upload the predictions to Kaggle.
2. You learned only a handful of classifiers in DAT200 so far. Therefore, you have only these available for prediction. You can use the following classifiers implemented in scikit-learn:

- Perceptron
- Logistic regression
- Support vector classifier (SVC kernel='linear')
- Support vector classifier (SVC kernel='rbf')
- Decision trees
- Random forests
- K-Nearest Neighbours (K-NN)

3. You have not yet learned all details of cross validation and how to apply it in scikit-learn. Therefore, we will continue our practice of using many train_test_splits (of the training data) while searching of the best parameters of each classifier as we have done so far in the lectures.
4. You may try out training classifiers with different number of features in X,

that is by removing features (feature selection) or making new features (feature engineering). You can also remove potential outliers if you identify some and if that improves your predictions.

5. You may collaborate with other students on the compulsory assignment, but you need to make your own submission to Kaggle and Canvas.

Deliverables / submissions

To have the compulsory assignment approved you need to deliver the following:

1. Your name must appear on the leaderboard of our own Kaggle competition, which means that you must submit at least one prediction (link to leaderboard: https://www.kaggle.com/c/dat200-ca3-2020/leaderboard (https://www.kaggle.com/c/dat200-ca3-2020/leaderboard))
2. You need to beat the benchmark classification accuracy score of **0.65256**. If you don't get above this benchmark, you will fail the compulsory assignment.
3. Submit (I) a Jupyter notebook; (II) PDF of Jupyter notebook; (III) Python code of your Jupyter notebook on Canvas with the code for training of your best classifier (please don't submit code for all seven classifiers) and the computation of the prediction. Please make short comments throughout your notebook/code on what you are doing and how you choose the parameters of your final best classifier. Leave out everything that is not necessary, keep only what is needed (we will reject PDF's that are full of unnecessary stuff)!
4. If you use an alias in the Kaggle leaderboard, you must provide your Kaggle alias AND your real name at the beginning of your Jupyter notebook
5. Remember that you can get your compulsory assignment approved during the exercises, where you provide an oral discussion of your work.