CS412 Machine Learning

Assignment 1: Decision Tree and K-NN

Due date: Sunday, March 28, 2021, 23:55
Late submission: till Tuesday, March 30, 2021, 23:55
(-10pts penalty for **each** late submission day)

In this assignment, you will classify wines as good or bad quality according to given features using Machine Learning approaches. You are asked to implement a decision tree and a k nearest neighbor classifiers to perform this task using the scikit-learn library. You will write your findings, results, and interpretations into a report and submit that as well.

Dataset Information

Wine quality data from UCI Machine Learning Repository:

https://archive.ics.uci.edu/ml/datasets/wine+quality

The dataset contains 4898 instances with 12 numeric attributes prepared by Paulo Cortez. Each row represents a unique wine with its related information. Refer to the dataset website for the details of the attributes.

The train and test datasets that you will use in this assignment can be found under this folder:

https://drive.google.com/drive/folders/1PC6M332CTdW-OOrgJ-1GU1F3UaRupka8?usp=sharing

You have 12 numeric attributes. You must split training data into 70% training and 30% validation sets. And use this validation set to choose your hyper-parameters.
You should use the test set provided in the folder to evaluate the performance of your models.

Implementation:

In this assignment, you are expected to use Google Colab.

https://colab.research.google.com/drive/1sWv6wwEGexq40SydkgSAu63VpBeYnYYt?usp=sharing

To start working on your homework, take a copy of this folder to your own google drive. You will do your implementations on this file and submit it with the expected outputs. We may just look at your notebook results; so make sure each cell is run and outputs are there.

Report:

Write an at most 1/2 page summary of your approach to this problem at the end of your notebook; this should be like an abstract of a paper or the executive summary. You should write the report under the cell in the Colab notebook.

Your report must include statements such as:

- Include the problem definition: 1-2 lines
- Talk about any preprocessing you did, how you handle missing values, and explain your reasoning
- Talk about train/val/test sets, size, and how to split
- State what your test results are with the chosen method, parameters: e.g. "We have obtained the best results with the ….. classifier (parameters=....), giving classification accuracy of …% on test data…."
- Comment on the speed of the algorithms and anything else that you deem important/interesting

You will get full points from here as long as you have a good (enough) summary of your work, regardless of your best performance or what you have decided to talk about in the last few lines.

You are expected to write this report on your own.

Office Hours:

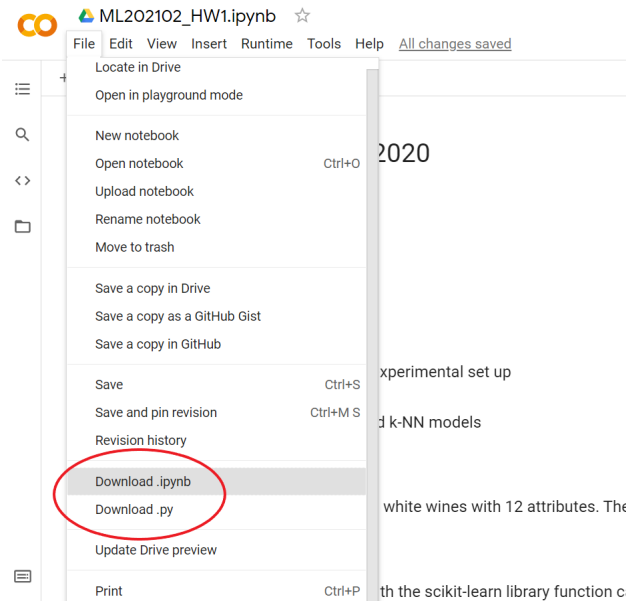
There will be dense office hours between 17:00 and 20:00 (three hours) every day between 22-27 March. You can join the office hours using the link below: https://sabanciuniv.zoom.us/j/98841829652?pwd=eXdmMGFXNldJemlib3lDbVFWdTAzZz09

Meeting ID: 988 4182 9652
Passcode: cs412

Submission Instructions

- You will submit this homework via SUCourse.
- Please read this document again before submitting it.
- Please submit your **"share link" INLINE in Sucourse submissions.** That is we should be able to click on the link and go there and run (and possibly also modify) your code. For us to be able to modify, in case of errors, etc, you should get your "share link" as **share with anyone in edit mode.**
- Download the **.ipynb and the .html** file and upload both of them to Sucourse.
- Please do your assignment individually, do not copy from a friend or the Internet. Plagiarized assignments will receive -100.

For .ipynb file:



For html file:

Right click on the page