CS412 Machine Learning

Assignment 3: Naive Bayes and Logistic Regression

Due date: Sunday, May 2, 2021, 23:55
Late submission: till Tuesday, May 4, 2021, 23:55
(-10pts penalty for **each** late submission day)

In this assignment, you will try to identify the topic of given documents according to given features using Machine Learning approaches. You are asked to implement a naive bayes and a logistic regression classifier 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

You will use 20 Newsgroups dataset on this assignment. This dataset can be downloaded via scikit-learn. There is already a cell in the colab notebook that will download the dataset for you.

Implementation:

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

`https://colab.research.google.com/drive/1D5h_5ogTrCfYcO_Iw1vQjL8cHD3TKyKp?usp=sharing`

To start working on your homework, take a copy of this notebook to your own google drive. Each step of this assignment is commented on in the shared colab. Please follow the instructions.

You will do your implementations on the file you copied into your own drive 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, explain your reasoning)
- Talk about train/test sets, size and how 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 feature importances of models
- Comment on anything 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:

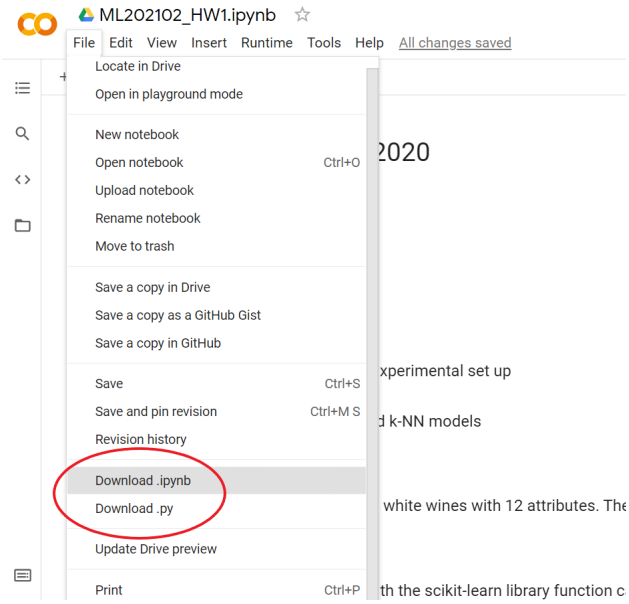
There will be dense office hours between 17:00 and 20:00 (three hours) every day between 26-30 April. You can join the office hours using the link below: https://sabanciuniv.zoom.us/j/8153671435

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 **shared with anyone in edit mode.**

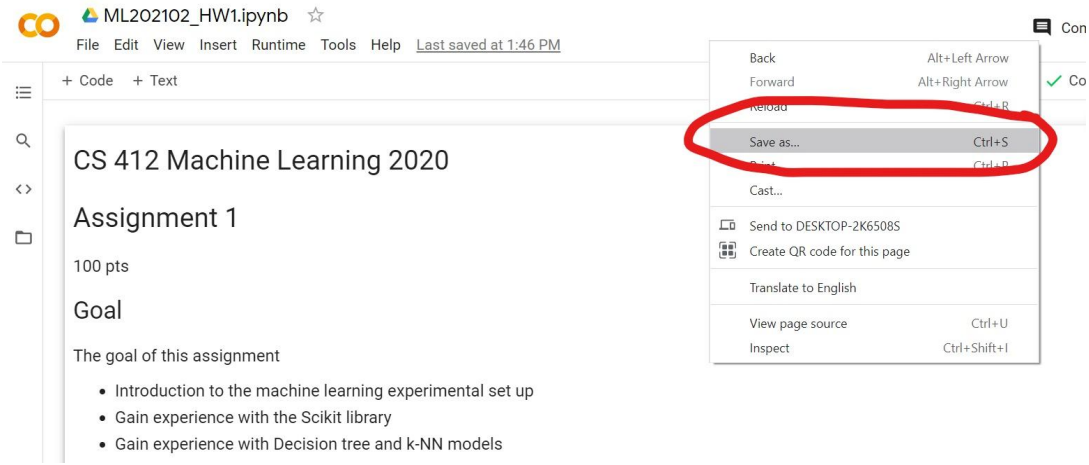
- Download the **.ipynb and the .html** file and upload both of them to Sucourse.
- Please do your assignment individually, do not copy from a friend or the Internet. Plagiarized assignments will receive -100.

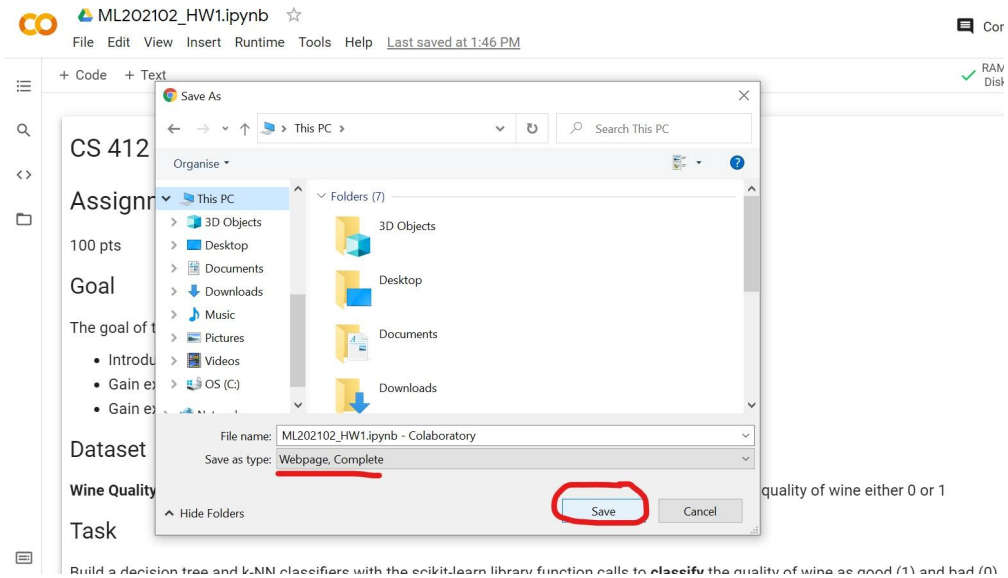
For .ipynb file:



For html file:

Right click on the page




ML202102_HW1.ipynb
File Edit View Insert Runtime Tools Help Last saved at 1:46 PM
+ Code + Text
CS 412
Assignn
100 pts
Goal
The goal of t
Introdu
Gain ex
Gain ex
Dataset
Wine Quality
quality of wine either 0 or 1
Task
Build a decision tree and k-NN classifiers with the scikit-learn library function calls to classify the quality of wine as good (1) and bad (0)
Save As
This PC
Search This PC
Organise
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Videos
OS (C:)
Folders (7)
3D Objects
Desktop
Documents
Downloads
File name: ML202102_HW1.ipynb - Colaboratory
Save as type: Webpage, Complete
Hide Folders
Save
Cancel