Assignment 1

April 4, 2021

Question 1: 5 points

Extend `l1_n1.ipynb` to try at least 3 more models that you can find in scikit-learn.

Question 2: 20 points

- Extend `l1_n3.ipynb` to try a Support Vector Machine regressor (sklearn.svm.SVR) with various hyperparameters, such as
  - kernel="linear" with various values for the C hyperparameter
  - kernel="rbf" with various values for the C and gamma hyperparameters.
- Dont worry about what these hyperparameters mean for now.
- Use GridSearchCV and RandomizedSearchCV.
- How does the best SVR predictor perform?

Question 3: 10 points

Add a transformer in the preparation pipeline to select only the most important attributes.

Hint:

- One way to decide which features are important, for the purposes of this assignment, is to use the correlation matrix as in cell 30 of `l1_n3.ipynb` and drop features whose `|correlation|` with the label is less than some threshold.
- This task belongs to `fit()` function of the transformer.
- The `fit()` function, as an argument, takes features and labels and therefore has complete information to perform such a calculation and decide which features can be dropped and save this information in the object for `transform()` function to use.

- The value of the threshold should be given as an argument to the constructor of the transformer.
- The `transform()` function would use the unimportant features computed in `fit()` to drop them from the frame and return the corresponding numpy array.
- Use cell 62 as a template for writing transformers.
- Cell 64 shows how to insert a transformer into the preparation pipeline

Question 4: not graded but highly recommended

- Study my Linux video lessons if you have not yet done so. Let me know if you did not get an invitation to the course. Knowledge of Linux will help you in many MSCA courses and at work.
- Follow `midway.pdf` to make sure you can run all the notebooks on midway2. For some assignments your laptop might not be powerfull enough.
- Note:
  - in the notebooks, you might have to split the part that downloads data from the rest and download on a login node but do computations on a compute node
  - the syntax of pandas, sckit-learn and other packages keeps changing and some things might require a small modification to run on RCC or your laptop.

Requirements

These are the general rules for submitting the homework not only for this assignment.

1. Submit jupyter notebook(s) in ipynb and html formats: html format allows the grader to easily read it in canvas while ipynb allows to execute it if something is not clear from html.
2. The notebook should be well formatted:
   - Use markdown to break it into sections, bullet points, etc.
     - See `l1_n3.ipynb` for example how to use it.
     - If you double click on markdown cell, you can see the code.
     - To show the formatted text, execute the corresponding markdown cell.
     - Start the notebook with a markdown title cell of the form: '# Assignment X, Jane Doe, date'.

   - When answering a particular question from the assignment, start the answer with markdown cell
     '## Question Y'
     that indicates what question you are answering.

3. Name your notebook (and the corresponding html file) according to the following template: `Assignment_X_Jane_Doe.ipynb`

4. If you are submitting several notebooks, append to the name `'_partZ'` and explain in canvas what each notebook is about.

5. There should not be any failed cells in the notebook. Every cell should work and have the results of the execution.

6. The notebook should contain only what is needed to answer a particular question and nothing else. More is as bad as less.

   - For example, in Question 2, do not submit the whole `l1_n3.ipynb` plus your code. Select from the `l1_n3.ipynb` only what is necessary before adding your code.

7. All the plots, if needed for the assignment, should be of production quality with readable labels, titles, etc. Also, try to find the most approriate type of plot to clearly demonstrate your statement.

8. Due date: 04/09/2021