MA124 Maths by Computer: Assignment 4

Machine Learning Applied to Bike Sharing Demand Data (20 Marks)

In a recent research article published in the journal Computer Communications, authors Sathishkumar V E, Jangwoo Park, and Yongyun Cho sought to predict the "bike count required at each hour for the stable supply of rental bikes"[1]. They employed several regression models, including linear regression. The dataset used in the original study is available here.

Assignment: Apply machine learning to a modified version of the original dataset and report the results.

[1] Sathishkumar V E, Jangwoo Park, and Yongyun Cho. 'Using data mining techniques for bike sharing demand prediction in metropolitan city.' Computer Communications, Vol.153, pp.353-366, March, 2020. web link.

The original research article and a modified dataset are posted on the Moodle page. You will need to refer to the article for some of the tasks below. You will need to download SeoulBikeData_mod.csv and put it into the folder with your assignment notebook. You do not need to submit SeoulBikeData_mod.csv with your assignment (see below).

SeoulBikeData_mod.csv has been modified from the original dataset to remove the categorical variables, and to convert dates to months. Months have been coded by number, e.g. 1 = January, etc. Only half the months are included in the modified dataset.

While the number of tasks is large, this is in part because the instructions are rather specific. Many of them follow directly from the Week 7 and 8 notebooks.

Computational tasks:

1. Import needed libraries. (You will need pandas, seaborn, as well as things from sklearn, and of course numpy and matplotlib.)

2. Using pandas, read SeoulBikeData_mod.csv into a Dataframe.

3. `describe` the Dataframe.

4. Plot a histogram of `Rented Bike Count`. Do not plot this as a density, but as a count. See Fig. 3 of the article. The vertical axis in the article is labelled "frequency", but is the same as the count.

   Produce a box plot similar to that in Fig. 3 of the article.

   Try to generate both the histogram and box plot to look approximately as they do in the article.

5. Produce two violin plots: one showing `Rented Bike Count` for different values of the `Month` and the other showing `Rented Bike Count` for different values of the `Hour`.

6. From the full Dataframe, create a new Dataframe `X` containing all the columns except `Rented Bike Count` and a Series `y` containing only the `Rented Bike Count` column. These are your design matrix and target respectively.

7. Perform a test-train split to create `X_train`, `X_test`, `y_train` and `y_test`. You **must** use the same percentage of data for testing and training as was used in the article and you **must** state what they are. You can find these in the article.

8. Create and train a linear regression model.

9. Use the trained model to obtain `y_pred`, the prediction on the test data `X_test`. Form the residual `resid = y_test - y_pred`.

10. Compute and report: Rsquared (R2), the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), and the Coefficient of Variation (CV). Compare these results to those on the top, right of Table 4 of the article. (Note, the modified dataset we are studying is different from that used in the article. Hence the results will not be identical. However, the procedure is very close to that used in the article.)

11. Produce and comment on the following plots.

- Histograms of `y_test` and of `y_pred` (on the same plot). These should be reported as counts rather than densities.

- A scatter plot of `resid` as a function of `y_test` corresponding to Fig. 9 of the article. (Recall what `y_test` represents and label the plot appropriately.) Unlike Fig. 9 of the paper, you should use a colormap to plot the different `Hours` in different colours.

- A scatter plot of `resid` as a function of `X_test['Month']`. Use a colormap to indicate the absolute value of `resid`.

- A scatter plot of `resid` as a function of `X_test['Rainfall(mm)']`. Use a colormap to indicate the absolute value of `resid`.

  (For all of the scatter plots, feel free to also vary the point size and colours to make attractive and informative plots. Choose a colormap that looks good to you.)

1. (Challenge material, 4 of the 20 marks) You will see in the article that most of the results involve "Trees". There are several types of trees used in machine learning. Sklearn provides a DecisionTreeRegressor.

- Create and train a DecisionTreeRegressor with `max_depth=8`. Here is an example that you will want to use.
- The train model to obtain `y_pred` in this case.
- Compute and report: Rsquared (R2), the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), and the Coefficient of Variation (CV). Compare with what is what

was obtained from the linear regression model above.
- Plot histograms of of `y_test` and `y_pred` (on the same plot).

Further 5 marks

A further 5 marks will be awarded for this assignment based on overall quality and clarity of the submitted notebook. Clearly this assignment lends itself to producing a nice document. Such a document might be useful to you in the future, for example in applying for internships.

Submission

You should not submit the SeoulBikeData_mod.csv file. You will submit **one Jupyter notebook**.

- The last thing you should do before submitting the notebooks is to Restart Kernel and Run All Cells. You should then save the notebooks and submit the .ipynb files. **You will lose one mark if you submit notebooks that have not been run.**
- No template will be provided, but you should be able to create your own notebook based on this assignment sheet and past submissions.
- If the notebook is run and all code cells are collapsed, the notebook should be readable as a well-formatted report, primarily consisting of:
    - A short introduction making reference to the original research article. (Approximately 50-100 words might be appropriate for this assignment. Restate from the Abstract or Introduction what the article is about. Obviously you won't understand all the details, but in a few words you should be able to summarise the motivation and goals.)
    - Computational tasks. Descriptions of these can be brief (from one to a few sentences, enough for the reader to follow without looking at the code.)
    - Properly labelled figures. There should be a short description of each (this is very important).
    - Ending summary of the results making connection back to the original research article, for example by comparing your results to those in the article.
    - Somewhere in the report a full citation to the research article should appear. Use this assignment sheet as a model.

Use the example notebooks as a guide for Python style. One assumes the reader understands Python. Add comments to set off blocks of code or to note anything tricky. In most cases Python code explains itself.

In [ ]: