EEP 596: Adv Intro ML || Programming Assignment 1

Dr. Karthik Mohan

Univ. of Washington, Seattle

Jan 6, 2022

Submission Guidelines

- Submit a Jupyter/IPython notebook file as your assignment solution to canvas
- Follow any setup instructions mentioned in the assignment - A good setup enables you to do programming seamlessly for future assignments
- You may discuss/brainstorm ideas to solve the assignment with peers - However, your submission should be your own and show your code implementation.

#1. Set Up

Colab environment

- **Notebook setup:** Access `https://colab.research.google.com/` - Create a notebook, pick any publicly available ML data set (e.g. Spam Filtering Dataset) of your choice and load it using the pandas library. Print the first few lines of this data in the notebook.
- Pick any categorical variable in the attributes of the data set and use an appropriate pandas library to create dummy attributes that capture the values of the categorical variable.
- Pick any numeric attribute from the list of attributes and plot a histogram of the values of the attribute in the data set. Use Matplotlib library to do so.

#1.Set up

Environment

- **Environment setup:** Connect a code-server or VS code to the colab server and post a screen shot. This will enable you to write code on VS code/code-server while doing your computations on google colab server. (Especially useful for mini-projects and more involved assignments).

#2. Housing Prices Data Set

Data Set

- You are given two data sets `train.csv` and `test.csv`
- Inspect the data set using Pandas and print first 10 lines
- How many attributes does the data set have and how many data points ?

#2. Housing Prices Data set

Linear Regression Model

- You want to fit a linear regression model. Before that, you want to get the data in shape. What pre-processing would you implement to do just that?
- Are there any attributes with missing data? If so, what do you do with these attributes when it comes to training?
- Fit a linear regression model by training on the pre-processed 'train' data.
- Evaluate your result on the 'test' data. What's your evaluation metric?
- Do you see a difference in the value of the evaluation metric between train and test data sets?
- Would you suspect over-fitting and why?

#2. Housing Prices Data set

Linear Regression Model

- How would you fix over-fitting in this case? Is over-fitting now reduced with your fix ? Definitely try ℓ_1 regularization as one of your fix strategies.
- For the ℓ_1 regularization strategy, plot your validation error as a function of the regularization hyper-parameter. Which hyper-parameter values gives you the lowest error?
- Based on the linear model you just learned, print the 10 most important attributes/features that seem to have a big impact on the prediction of the sale price? Do the important attributes make sense and match your intuition/knowledge on housing prices?
- What's the R^2 coefficient for your linear model?
- All your steps towards the solution to this problem should be captured in your python notebook for full credit.