Homework 1
EECS 498/598: Applied Machine Learning for Affective Computing

Due: January 26th, 2022

1 Introduction

This homework contains two parts, Part 1 on Linear Regression and Part 2 on Logistic Regression. Both parts use real world data and will introduce you to techniques used in the workforce today!

This pdf will contain broad instruction as well as analysis questions to do after you have finished the coding for each part of the homework. hw1.ipynb contains more detailed instructions for coding the homework and serves as a "driver" for the code you will write. answers.py is where you will fill in your answers to the coding questions as directed by hw1.ipynb. **answers.py should be the ONLY file you actually edit!** Submit answers.py to gradescope to receive feedback on your answers. The scores shown on gradescope are final, functions that produce graphs will not be autograded.

1.1 Grading and Due Date

This assignment is due on January 26th, 2022. Overall, each part is worth 50 points: 35 points for the coding and 15 points for the analysis questions. Point values are attached to each function as well as each analysis question. The coding output is graded on correctness and **not** style, the analysis questions will be graded for correctness.

1.2 Submission

Submit your answers.py to the autograder on gradescope, most of your functions will automatically be graded here and you have unlimited submissions. On canvas, submit your hw1.ipynb as well as a pdf containing the answers to the analysis questions **and any graphs** you produced during the coding portion.

In summary the three files that need to be submitted are your answers.py file to gradescope and hw1.ipynb and pdf with graphs and analysis questions to canvas.

2 Installing/Learning Python

If you've never used python before you'll want to run through this google class to nail down the basics: https://developers.google.com/edu/python

If you need to install python on your computer we recommend using Anaconda, which can be found here https://www.anaconda.com/products/individual

We also have created our own cheat sheets just for you all! You can find them in the canvas Files/Cheatsheets. These cheatsheets will get you caught up on how to use Jupyter Notebook, matplotlib, pandas, sklearn, and other basics. If you're completely new to python let me clue you in on a little secret, these libraries are SUPER IMPORTANT. Especially for this class! You'll be glad you learned them, and by the end of this homework you will know how to use all of them to some extent.

3 Part 1: Linear Regression

Part 1 of the homework will focus applying Linear Regression to real world data to try and predict the dollar price of a Big Mac given world economic information. Because this is real world data, you will have to clean some invalid values as well as verify the assumptions of Linear Regression. In this section you will learn how to manipulate pandas DataFrames to clean data, use statsmodels Linear Regression, and graph to verify the statistics of your Linear Regression fit.

After you have completed the code for Part 1 by following along with hw1.ipynb and filling in the answers in answers.py, come back and answer these reflection questions about different parts of the Linear Regression Analysis.

3.1 Analysis Questions

1. (2 pts.) Looking at the results from localPriceOutlier and dollarExOutliers functions, should these two rows be removed? Does the local price and dollar exchange rate in this country invalidate this dollar price data?
2. (2 pts.) Looking at the results from the graphIndepVsDep function, which variables should be removed? Why?
3. (2 pts.) Referring to the correlationHeatmap function, which variables are highly correlated to each other?

4. (2 pts.) Looking at the results from the correlationHeatmap function, which variables should be removed? How many variables need to be removed?

5. (2 pts.) Do the graphs produced by graphIndepVsResidual satisfy our assumptions? Why or why not.

6. (1 pts.) Does the histogram of residuals generated by histOfResiduals and the qq-plot generated by graphQQPlot fit well? What, if anything, is unusual?

7. (1 pts.) Consider the plot for the residuals generated by graphFittedVsResidual, which of the three conditions (normal distribution, independence, homoscedasticity) are met?

8. (2 pts.) Is our Linear Regression model good? This is a slightly ambiguous question on purpose, use some of the data given by model.summary() to justify your answer.

9. (1 pts.) What is one challenge associated with replacing all values with the mean in replaceNaNWithMean?

4 Part 2: Logistic Regression

In this section we will be predicting whether an IMDB movie review is positive or negative using logistic regression! The IMDB movie review dataset is a popular sentiment analysis dataset that is used in many parts of machine learning. To keep this assignment manageable, we have provided a subset of 6,000 movie reviews as well as their associated labels positive (1) and negative (0). These labels were made and validated by humans so we accept them as a ground truth.

After you have completed the code for Part 2 by following along with hw1.ipynb and filling in the answers in answers.py, come back and answer these reflection questions about different parts of the Logistic Regression Analysis.

4.1 Analysis Questions

1. (2 pts.) Does removing stop words increase the accuracy of our model? Why?

2. (2 pts.) Does removing words with non-alphabetic characters in them increase the accuracy of our model? Why?

3. (2 pts.) Does using TfIdf increase the accuracy of our model? Why?

4. (3 pts.) Using the scoreText function make up a sentence that is in your view scored incorrectly. Give the sentence and explain why our model falls short in this case.

5. (3 pts.) If only 1/3 of our labels were positive would accuracy be a good metric? Why?

6. (3 pts.) Can two identical Bag of Words vectors have different meanings in text form? Why? What about Tf-Idf?