Homework 8 - Support Vector Machines

Washington DC Bikeshare Data (155 points)

Instruction:

- This is a group-work assignment!
- You are expected to submit the **.ipynb** file and the exported **.html**.
- Only one member in each group needs to submit the assignment. It will be automatically submitted for the rest of group members.
- This is a long assignment, start early!

Question 1 EDA (30 points)

In this exercise I want you to do a quick EDA on the Logan_housing data set. The data are 4110 observations of Logan housing prices from 2018 to 2020. All the variable names are self explanatory except the **DOM** which stands for **Days On the Market**. Please do not share this data set with anyone outside of the class.

Show me what you have learned from the previous EDAs you did in HW2 and HW3. Try to come up with an interesting story (hypothesis) using this data set. For example, that would be interesting to see the breakdowns of Logan house prices by year, location etc. Treat this exercise as a real world project. Many times the managers have no idea what they want from the data!! your job is to be as creative as possible and come up with informative charts and tables.

Question 2 SVM Regression (60 points)

Do the followings:

1. Define the categorical variables and transform them into dummy variables (if you haven't done this already in Question 1). **(5 points)**
2. Scale all the variables using standardization. **(5 points)**
3. Define the feature space and target variables. Split the data into test (20%) and train set (80%) **(5 points)**
4. From sklearn.svm import the relevant function for SVM regression. Do the followings: **(25 points)**
    1. Train the regression model using its default inputs. (5 points)
    2. Make predictions on the test set and save them as y_hat (5 points)
    3. Construct a data frame named df_predictions with 2 columns. y_test, and y_hat from previous part (5 points)
    4. Visualize actual vs predicted prices in the test set using an scatter plot. Are you visually satisfied with the regression model? (5 points)
    5. Report the R squared and the RMSE in the test set. (5 points)
5. Tuning the hyperparameters using gridsearchCV. I want you to specifically use the following param_grid.

    my_param_grid = *'C': [1,10,100], 'gamma': ['scaled',0.1,0.01], 'kernel': ['rbf']*

    (5 points)
6. Re-estimate (Re-fit) the SVM regression model with the optimal parameters from the gridsearch method. Save the predictions as y_hat_optimized and add this column to the df_predictions data frame from part 4.3 in Question 2. **(5 points)**

7. Report the optimized R-squared and RMSE in the test set and compare them with the outputs from part 4.5 in Question 2. **(5 points)**
8. Estimate the optimized RMSE_test using 5 fold cross validation. **(5 points)**

Question 3 SVM Classification (65 points)

Jeff (the owner of the data set) is specifically interested in the DOM variable (Days on the Market) and he wants to make a classifier that predicts if a new listing will be liquid enough or not? i.e. the number of days on the market (for a new listing) is above or below the average DOM. Unfortunately Jeff has not taken the machine learning course yet and he doesn't know how to apply SVM classification. Could you help him out with the following tasks?

1. Define a binary target variable liquid. $liquid = 1$ if $DOM < average(DOM)$ and 0 otherwise. What are the proportions of liquid listings vs illiquid ones in the data set? Is the target variable (relatively) balanced or (relatively) imbalanced? **(5 points)**
2. Along with the target variable, define your feature space (X) and split the data into test (20%) and train set (80%) **(5 points)**
3. From sklearn.svm import the relevant function for SVM classification. Do the followings: **(30 points)**
    1. Train the SVM classification model using its default inputs. (5 points)
    2. Make classifications on the test set and save them as y_hat (5 points)
    3. Construct a data frame named df_classifications with 2 columns. y_test, and y_hat from previous part (5 points)
    4. Borrow my_SVM_report() function from the python notebook of the SVM lecture. Report the Accuracy, precision, recall and f1 score along with the confusion matrix. Interpret all of these statistics. Do you trust the accuracy of the model? why? (10 points)
    5. Can you plot the ROC curve and report the AUC score in SVM classification? why? (5 points)
4. Tuning the hyperparameters using gridsearchCV. I want you to specifically use the following param_grid.

    my_param_grid = *'C': [1,100,1000], 'gamma': ['scaled',0.01,0.001], 'kernel': ['rbf']*
    (5 points)
5. Re-estimate (Re-fit) the SVM classification model with the optimal parameters from the gridsearch method. Save the predictions as y_hat_optimized and add this column to the df_classifications data frame from part 3.3 in Question 3.

    (5 points)

6. Report the optimized classification metrics using my_SVM_report() function and compare them with the outputs from part 3.4 in Question3. **(5 points)**
7. Estimate the optimized accuracy_test using 5 fold cross validation. **(5 points)**
8. Why do you think Jeff is interested in predicting the liquidity premium (either positive or negative) of the houses? There may be multiple correct answers to this question. Just list whatever reason seems appropriate to you. **(5 points)**

Good luck and enjoy machine learning!