ECE 562 Machine Learning, Fall 2019
Instructor: Zafer Aydın
Computer Homework 4

Introduction

Download the template codes for Udacity's Intro to Machine Learning course by running the following command in a terminal

git clone https://github.com/udacity/ud120-projects.git

Submit your scripts as q1.py and/or a text document that includes your answers to the questions below through Canvas.

Assignment

1. Fill the missing code lines in `regression/finance_regression.py` by fitting a linear regressor and changing the color of the test samples to red. The script will draw a scatterplot, with all the data points drawn in. Include this plot into your report.

What target are you trying to predict? What is the input feature being used to predict it?

2. What are the slope and intercept of the line that you fitted? Hint: the slope is the reg.coef_ attribute and the intercept is the reg.intercept_ attribute.

3. What accuracy score (i.e. R^2 score) do you find when you test your regressor on train set?

4. What's the R^2 score on the testing data?

5. There are lots of finance features available, some of which might be more powerful than others in terms of predicting a person's bonus. For example, suppose you thought about the data a bit and guess that the "long_term_incentive" feature, which is supposed to reward employees for contributing to the long-term health of the company, might be more closely related to a person's bonus than their salary is.

A way to confirm that you're right in this hypothesis is to regress the bonus against the long term incentive, and see if the regression score is significantly higher than regressing the bonus against the salary. Perform the regression of bonus against long term incentive. Include the scatter plot into your report. What's the R^2 score on the test data? Compare this R^2 score to the score you obtained in question 4. Which one is better?

6. In this question you will explore the effect of having an outlier on train set. Go back to the setting where you predict bonus from salary. Train a linear regression model using test set, which includes an outlier point as well (see the red data point in Figure 1).

Add these two lines near the bottom of finance_regression.py, right before plt.xlabel(features_list[1]):

```
reg.fit(feature_test, target_test)
plt.plot(feature_train, reg.predict(feature_train), color="b")
```

Now we'll be drawing two regression lines, one fit on the test data (with outlier) and one fit on the training data (no outlier). Include the plot to your report. That single outlier is driving most of the difference. What's the slope of the new regression line?