CSCE 478/878 Recitation 8 Handout

Linear Regression: Ordinary Least Squares (OLS) Method and The Singularity Problem

March 5, 2019

- This is a **graded recitation**. Submit your work for grading.
- Your Jupyter notebook should be submitted via webhandin by 4:45PM, **March 5**.
- Use the following naming convention: `<lastname>_<firstname>_8.ipynb`

Part 1: OLS Linear Regression Using Python (90 pts)

1. Read the "*Linear Regression-1-OLS*" jupyter notebook (posted on github) carefully to understand how to use the ordinary least squares (OLS) method for solving a Linear Regression problem. In particular observe how to fix the **singularity** issue.
2. Then, implement the marked sections in the "*Recitation-Linear Regression-OLS-Singularity Problem*" jupyter notebook. You will need to download the file "OLS_Data.csv" to create your dataset for this recitation.

Note: "train_test_split" Function:

You should set the "random_state" attribute in the "train_test_split" function as shown below. If you don't specify it, then every time you run your code a new random value is generated and the train and test datasets would have different values each time.

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

Score Distribution:

- Manually Coded OLS Solution (**pts 35**)

Expected Output:

```
Determinant of (X_bias^T.X_bias):  0.0
LinAlgError: Singular matrix
```

- Applying OLS Method on Data Matrix With Colinearity in Columns (**pts 50**)

Expected Output:

```
Mean squared error: 21.64
```

```
Coefficient of determination r^2 variance score [1 is perfect predic
tion]: 0.75
```

Note: your output might vary slightly because of the randomness of the train-test slip.

- Evaluate the Model Using Test Data - OLS Linear Regression (**pts 5**)

Expected output:

```
Mean squared error: 24.29
Coefficient of determination r^2 variance score [1 is perfect predic
tion]: 0.67
```

Note: your output might vary slightly because of the randomness of the train-test slip.

Part 2: Understanding the Singularity Issue and its Solution (10 pts)

Answer these questions in the marked sections of the "*Recitation-Linear Regression-OLS- Singularity Problem*" jupyter notebook.

1. Why do you think the singularity matrix error occur while using OLS method on the "OLS_Data.csv" dataset?
2. To fix the singularity problem of the $X_{bias}^{T}X_{bias}$ matrix what non-zero positive number did you add on its diagonal?
3. Add 100000 on the diagonal of the $X_{bias}^{T}X_{bias}$ matrix and report the MSE and the R^2 values for the training data set. Explain these results.
4. After adding 100000 on the diagonal of the $X_{bias}^{T}X_{bias}$ matrix what change did you notice in the weights of the model?