Lab 4

Sklearn

- First define the steps

```
from sklearn.svm import SVC
from sklearn.preprocessing import StandardScaler
steps = [('scaler', StandardScaler()), ('SVM', SVC())]
```

- Create pipeline object

```
from sklearn.pipeline import Pipeline
pipeline = Pipeline(steps) # define the pipeline object.
```

- Split the data into train and test

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,Y,test_size=0.2)
```

- Fit and predict the regressor

```
pipeline.fit(x_train, y_train)
predicted = pipeline.predict(x_test)
```

- Calculate the mean squared error

```
from sklearn.metrics import accuracy_score, mean_squared_error
print(mean_squared_error(y_test, predicted))
```

Sklearn

- Sklearn FunctionTransformer
- This class can be useful if you're working with a *Pipeline* in *sklearn*

```
def all_but_first_column(X):
    return X[:, 1:]
pipeline = Pipeline([ ('pca':PCA(), "fselect":(FunctionTransformer(all_but_first_column))])
```

Exercise 1

- Open data.csv. First column is the value of X i.e., feature and the second column represents corresponding Y value.
- Load data.csv into pandas
- Fit linear regression on the data
  - Using normal equation
  - Using gradient descent
    - With learning rate L = 0.001
    - Updated m (B in normal equation) and c (a in normal equation) for 100 iterations.
    - Stop the algorithm when there are very small updates for m and c. This happens when the algorithm converges.

Hint: use the equations from the class lecture.

Normal Equation

- Slide 18 and 22 in the lecture. Find value of a and B.

Gradient Descent

- See slide 32.
- Calculate D_c by doing partial derivative.

Exercise 2

1. Read chicago_hotel_reviews.csv file
2. Split the data into 80% train and 20% test.
3. Calculate tf-idf features of “review”.
4. Predict “rating” of each “review” using **Sklearn’s** linear regression
5. Do the same but this time, apply a feature selection technique on the tf-idf features that you computed in step 3. Use any of feature selection techniques available in sklearn.
6. Calculate mean squared error to evaluate your regressor.

Hint: You may use these libraries from sklearn to calculate tf-idf scores and select features

CountVectorizer, SelectPercentile