ECE 521 Assignment 1 k-NN

By: Logan Rooks, Krist Papadopoulos, Yu Liu

Due: February 2, 2018

1. **Euclidean Distance Function**

The squared Euclidean distance function for the input of two matrices is given below in vectorized Tensorflow.

def **dist\_euclid**(X, Z):

'''

Compute squared Euclidean L2 distance between tensor X and tensor Z

Input: X is an N1 X D tensor, Z is is a N2 X D tensor

Output: ||X-Z||2/2 distance is a N1 X N2 tensor containing pairwise squared Euclidean distances

X = tf.constant([[3, 4, 5], [5, 1, 1]])

Z = tf.constant([[2, 4, 5], [1, 1, 1], [6,1,8]])

D\_euc(X,Z) = [[ 1 29 27]

[34 16 50]]

'''

X\_norm = tf.reshape(tf.reduce\_sum(tf.square(X),axis=1), [-1,1])

Z\_norm = tf.reshape(tf.reduce\_sum(tf.square(Z), axis=1), [1,-1])

dist = X\_norm + Z\_norm - 2\*tf.matmul(X,tf.transpose(Z))

return dist

1. **Making Predictions for Regression**
   1. **Choosing the Nearest Neighbours**

A vectorized Tensorflow function was created to take a test point and compare to the training points to get the pairwise distances. A parameter *k,* for the # of closest training points (based on the pairwise Euclidean distance) to the test point is used with the tensorflow function *tf.nn.top\_k,* to select the indices for these closest training points.

Need to add code here – to discuss further with TA

* 1. **Prediction**

**For k = 1:**

The train set MSE is:

0.0

The validation set MSE is:

0.543099

The test set MSE is:

0.293814

**For k = 3:**

The train set MSE is:

0.211613

The validation set MSE is:

0.648607

The test set MSE is:

0.264381

**For k = 5:**

The train set MSE is:

0.260886

The validation set MSE is:

0.583462

The test set MSE is:

0.356484

**For k = 50:**

The train set MSE is:

2.49342

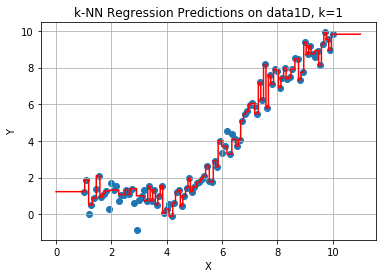
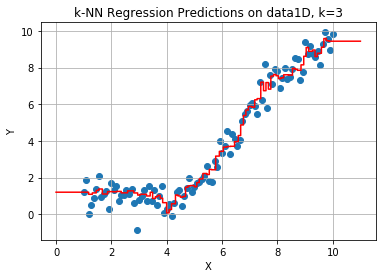
The validation set MSE is:

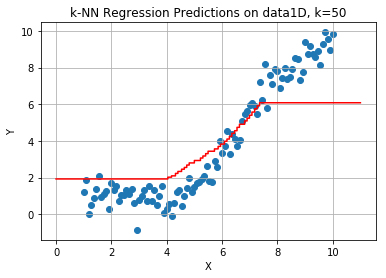
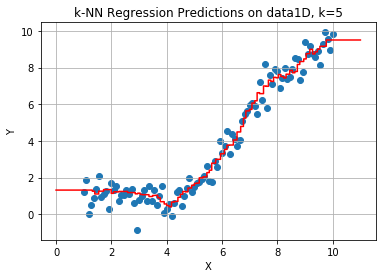
2.44609

The test set MSE is:

1.40526

Therefore, the best k is k = 1 based on the lowest validation MSE.





The plots show how the data is fit by k-NN, where for k=1 there is overfitting to the data, and for k=50 there is underfitting. From the plots the best fit appears to be k=5, where the fit is likely to generalize better since it captures the trend without significant overfitting or underfitting.

1. **Making Predictions for Classification**
   1. **Predicting Class Label**
   2. **Face Recognition using k-NN**
   3. **Gender Recognition using k-NN**