Machine Learning
Homework 5 : KNN
(due Midnight March 15)

Instructions

1. In case you are unfamiliar with the Python data ecosystem (NumPy, Pandas), you are recommended to study the first four chapters of the Python data science handbook. A doubt clearing session would be organised in case you have any difficulties in the data science ecosystem.
2. The deadline for full score is Midnight March 15. You can get 50% credit for late submission (Midnight March 17).
3. Total marks = 10
4. You have to type the assignment using a word processing engine, create a pdf and upload on the form. Please note that only pdf files will be accepted.
5. All code/Jupyter notebooks must be put up as **secret gists** and linked in the created pdf submission. Again, only secret gists. Not public ones.
6. Any instances of cheating/plagiarism will not be tolerated at all.
7. Cite all the pertinent references in IEEE format.
8. The least count of grading would be 0.5 marks.

1. (a) Implement KNN classification and regression using Numpy and Pandas. Keep the distance metric as a function argument which can take: 'Euclidean', 'Manhattan' or 'Cosine' **[2 marks]**.

```
def KNN_predict(type='classification', train_X, train_Y, test_X, K, distance_metric='Euclidean'):
```

   (b) Vary the dimension (number of features of X) and the number of train instances and empirically show the runtime of your algorithm as function of dimension and number of train instances. Feel free to use randomly generated data. How does this compare with the theoretical time complexity of KNN? **[2 marks]**.

2. Show the usage of scikit learn's KNN module for the real estate price prediction regression problem.

   (a) Using 5-fold cross-validation report the optimal K value for each fold, alongside, Train, Validation and Test RMSE error. How does the test error compare with the test errors for the different folds you obtained in eariler assignments. **[1 mark]**

   (b) Are all the features on the same scale in the above solution? Does that impact KNN? How? Now, scale the features between 0 and 1. You may also want to ensure that when you're predicting, you scale the input features of the test set using the same function used to transform the train input features. Using 5-fold cross-validation report the optimal K value for each fold, alongside, Train, Validation and Test RMSE error. **[2 marks]**

   (c) Let us pick up a single fold (train on first 80% data and last 20% as test data) and plot train, test error as a function of K. Comment on the shape of the train error curve. **[1 mark]**

      i. On this test set, find the home which gives the maximum RMSE error. What can you say about KNN performance on this home. Why is it poor? Is there something you could do to improve the prediction for this home? **[1 mark]**

3. Draw a Voronoi diagram for the IRIS dataset for 1-NN, where we are only using 'sepal-length' and 'sepal-width' as the two fearures. You could use 3 different colours for the 3 different classes. What can you infer from this diagram? **[1 mark]**

   Questions below will not contribute to your score. They are presented here for the interested reader

4. Implement KD-Tree algorithm for KNN.

5. Question 4.7.4 from ISLR

6. Implement cosine based KNN for the movie recommendation problem. Use MovieLens 100k dataset for the same.

7. We saw in the lectures that Euclidean norm is not well siuted for high dimensional problems. Repeat the experiments for L_p norm where $p <= 1$. What can you conclude from this analysis.