KNN ASSIGNMENT

KNN

The kNN classifier consists of two stages:

1. During training, the classifier takes the training data and simply remembers it
2. During testing, kNN classifies every test sample by comparing to all training samples and transferring the labels of the k most similar training examples.

We would now like to classify the test data with the kNN classifier. Recall that we can break down this process into two steps:

1. First we must compute the distances between all test examples and all train examples.
2. Given these distances, for each test example we find the k nearest examples and have them vote for the label

Now for the given dataset that captures essential medical parameters of a person with and without diabetics.

Dataset link:
https://raw.githubusercontent.com/MSPawanRanjith/FileTransfer/master/diabetes.csv

Objective:

Build a KNN model for prediction whether a person will have diabetes or not with a high accuracy score.

1) Perform some appropriate Pre-Processing steps on the given dataset for better results (Ex, try converting categorical to numerical)
2) Implement the KNN algorithm on your own. (Don't use any pre built code/lib)
3) Try other possible processes that can be done to dataset and tuning the model to increase accuracy.
   - Increase K value
   - Normalisation
   - Different Distance Metrics
4) Perform Feature Ablation Study

Additional Tries: Weight the features before doing KNN prediction.

K-MEANS ASSIGNMENT:

K-MEANS

Clustering is one of the most common exploratory data analysis techniques used to get an intuition about the structure of the data. K-Means algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group, It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.

BIC (Bayesian Information Criterion):

Read thru this for better understanding:

https://github.com/MSPawanRanjith/FileTransfer/blob/master/BIC_notes.pdf

Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models. This criterion is similar to likelihood, but it places a penalty on the complexity of the model.

$$BIC(\phi) = \hat{l}_{\phi}(D) - \frac{p_{\phi}}{2} \cdot \log R$$

$\hat{l}_{\phi}(D)$ → Is the likelihood of a model (Φ)

p_{ϕ} → is the number of parameters in the model (Φ)

R → is the total number of points belonging to a centroid

Dataset Link:

https://raw.githubusercontent.com/MSPawanRanjith/FileTransfer/master/kmean_dataset.csv

Objective:

Build a K-Means Model for the given dataset. So in K-Means choosing the K value that gives a better model is always a challenge, As we increase value of K with dataset having n points, the likelihood of the model increases, and obviously K<N, so rank or maximize the likelihood we use BIC(read about Bayesian Information Criterion for better understanding, before attempting the question)

Now,

1) Build a K-Means Model for the given Dataset (You can use the library funct.)
2) Implement the BIC function that takes the cluster and data points and returns BIC value
3) Implement a function to pick the best K value, that is maximize the BIC.
4) Visualize the pattern found by plotting K v/s BIC.