ENPM703- Assignment-1

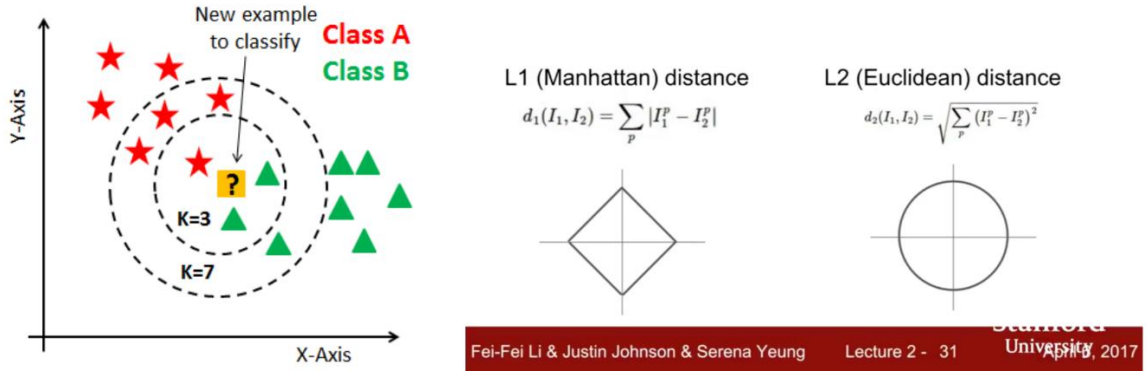
Part1: KNN Classifier

CIFAR-10 Dataset

Contains collection of 60000 RGB images with dimensions 32x32. Thus, loading the training dataset forms an array of size (50000x32x32x3). This data is pre-processed to a row matrix of dimensions (50000x3072) by straightening up the pixel values. Only a sample of 5000 images for training and 500 images for testing are used.

K-NN Classifier

K nearest neighbour is a classification algorithm which classifies dataset according to the similarity with other data points. This similarity can be found by l1-Manhattan or l2-Eucladian distance. The class of the test data is predicted based on the closest distance between the test and k nearest points.



L1 and L2 Distance

L1 distance gives the absolute difference between the pixel values. This type of distance measurement is suited if the dimensional features have particular meaning such as make, number of seats, year of manufacturing, etc. of a car.

L2 distance gives the straight-line distance between two points. This can be used if the features do not have any specific meaning, like in this case where they are the pixel values of the images.

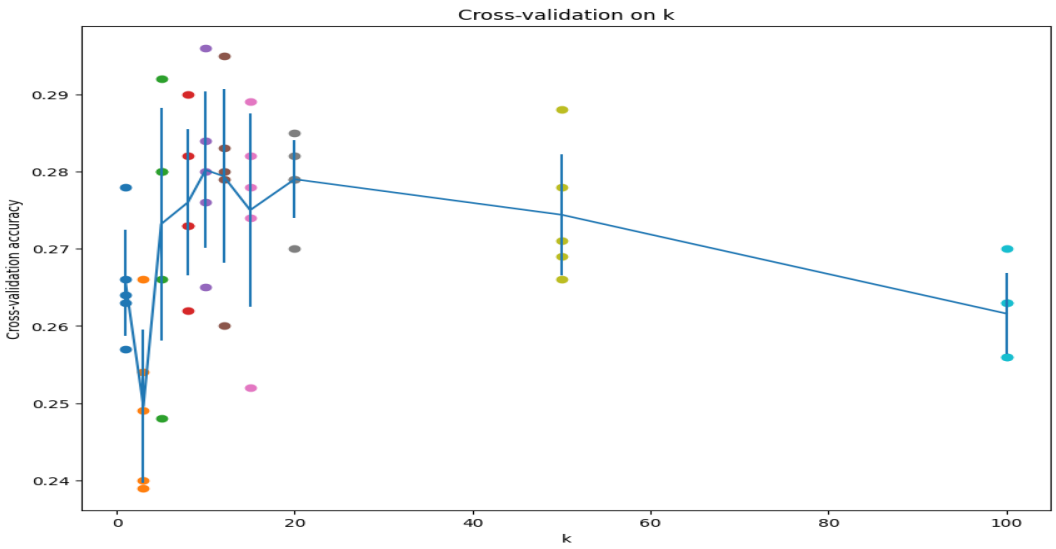
In this algorithm, we have computed l2 distance with two, one and zero loop. Since iterating once is an efficient vectorized code, by using basic matrix algebra, the computation of l2 distance becomes much quicker.

Cross Validation

Cross Validation is used to determine the best k value such that the algorithm has maximum accuracy. In cross validation, we split the training data into five folds and out of five one of the folds is kept as validating set. The algorithm is trained using the rest of the four folds and the k value is validated using the validating set.

Also, the validating set get changed in each iteration, so that there is no bias in the training. Thus, the training and prediction is done for all five iterations and the accuracy values are recorded in a dictionary.

Finaly, the accuracy values for each k are plotted and the k with maximum accuracy is chosen.



Here both k=10 and k=12 has similar accuracy values, but k=10 has the maximum accuracy and is the suitable value of k for this data set.

Is K-NN best?

The maximum accuracy in this classification algorithm is 28.7%, which is slightly higher than the baseline accuracy of 10%. One of the reasons for the minimal performance is that knn is sensitive to redundant features such as background of an image. For example,

in these images the background is mostly similar and the distance value will be minimal, since there is no much variation un the pixel intensities.

Moreover, knn does not perform well if the data is not well sampled. Since knn predicts the class based on the neighbours, if the data is not sampled properly the algorithm might predict the class whichever is closer to the test data. Thus, might not be suitable for minority class samples.