QUESTIONS:

Dataset:

Optdigits data by Alpaydin and Kaynak, from UCI Machine Learning Repository:
ftp://ftp.ics.uci.edu/pub/ml-repos/machine-learning-databases/optdigits/

You need the files:

optdigits.names	explanation of data
optdigits.tra	training data
optdigits.tes	test data

Q1) [4 points] [Clustering]

Write down the code for ~~knn~~ algorithm in python and cluster the training data set into K clusters. As the distance measure two instances use two similarities:

1: **Euclidean Distance,**

2: **Weighted Euclidean distance** where the pixels at the center of the image are twice more weight than the pixels around the borders of the image and the weight decreases linearly from the center to the borders.

Measure clustering performance using 2 methods:

1: **Reconstruction error** of each cluster and its average

2: **Gini impurity** (use Gini index for multiple classes) of each cluster and its average.

Plot **K=10, 20, 30** vs the average impurity for each distance measure on the training set.

Q2) [2 points] [Nonparametric Methods]

[K-NN] Classify the test set, using 1-NN and 5-NN algorithms and the two distance measures (Euclidean Distance and Weighted Euclidean Distance) to the training instances.

[Centroid Based] Classify the test set using the distance to the cluster centroids for K=30 clusters, the two distance measures and the cluster majority as the class label.

Compare the performances of both methods.

NCC (nearest cluster classifier)