Assignment

In this assignment, you will implement k-means clustering and use it to construct a multi-label classifier to determine the variery of wheat. For the dataset, we use "seeds" dataset from the machine Learning depository at UCI:
`https://archive.ics.uci.edu/ml/datasets/seeds`

Dataset Description: From the website: "... The examined group comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected for the experiment..."

There are 7 (continuous) features) $F = \{f_1, \ldots, f_7\}$ and a class label L (Kama: 1, Rosa: 2, Canadian: 3).

1. f_1: area A
2. f_2: perimeter P

3. f_3: compactness $C = 4\pi A/P^2$
4. f_4: length of kernel,
5. f_5: width of kernel,
6. f_6: asymmetry coefficient
7. f_7: length of kernel groove.
8. L: class (Kama: 1, Rosa: 2, Canadian: 3)

For the first question, you will choose 2 class labels as follows. Take the last digit in your buid and divide it by 3. Choose the following 2 classes depending on the remainder R:

1. $R = 0$: class $L = 1$ (negative) and $L = 2$ (positive)
2. $R = 1$: class $L = 2$ (negative) and $L = 3$ (positive)
3. $R = 2$: class $L = 1$ (negative) and $L = 3$ (positive)

Question 1: Take the subset of the dataset containing your two class labels. You will use random 50/50 splits for training and testing data.

1. implement a linear kernel SVM. What is your accuracy and confusion matrix?
2. implement a Gaussian kernel SVM. What is your accuracy and confusion matrix?

3. implement a polynomial kernel SVM of degree 3. What is your accuracy and confusion matrix?

Question 2: Pick up any classifier for supervised learning (e.g. kNN, logistic regression, Naive Bayesian, etc).

1. use this classifier to your dataset. What is your accuracy and confusion matrix?

2. summarize your findings in a table below and discuss your results

Model	TP	FP	TN	FN	accuracy	TPR	TNR
linear SVM Gaussian SVM polynomial SVM your classifier							

Question 3: Take the original dataset with all 3 class labels.

1. for $k = 1, 2, \ldots, 8$ use k-means clustering with random initialization and defaults. Compute and plot distortion vs k. Use the "knee" method to find the best k.

2. re-run your clustering with best k clusters. Pick two features f_i and f_j at random (using python, of course) and

plot your datapoints (different color for each class and centroids) using f_i and f_j as axis. Examine your plot. Are there any interesting patterns?

3. for each cluster, assign a cluster label based on the majority class of items. For example, if cluster C_i contains 45% of class 1 ("Kama" wheat), 35% of class 2 ("Rosa" wheat) and 20% of class 3 ("Canadian" wheat), then this cluster C_i is assigned label 1. For each cluster, print out its centroid and assigned label.

4. consider the following multi-label classifier. Take the largest 3 clusters with label 1, 2 and 3 respectively. Let us call these clusters A, B and C. For each of these clusters, you know their means (centroids): $\mu(A)$, $\mu(B)$ and $\mu(C)$. We now consider the following procedure (conceptually analogous to nearest neighbor with $k = 1$): for every point x in your dataset, assign a label based on the label on the nearest (using Euclidean distance) centroid of A, B or C. In other words, if x is closest to center of cluster A, you assign it label 1. If x is closest to center of cluster B, you assign it class 2. Finally, if x is closest to center of cluster C, you assign it class 3. What is the overall accuracy of this new classifier when applied to the complete data set?

5. take this new classifier and consider the same two labels that

you used for SVM. What is your accuracy and confusion matrix? How does your new classifier (from task 4) compare with any classifiers listed in the table for question 2 above?