Artificial Intelligence (CS 534)
Assignment 3 (80 points)
Due: November 4th, BEFORE CLASS

Problem 1

a. Implement your own k-means algorithm from the lecture slides using Python. (**10 points**)

b. Using the k-means algorithm, cluster the data from the attached file **cluster_data.txt**.
Plot X, Y coordinates for the entire dataset. Use different symbols and colors to represent your data points for different clusters.
Label X and Y axis as 'Length' and 'Width', correspondingly. Label each cluster as "Cluster 1", "Cluster 2", etc. Explain your findings. (**5 points**)

Problem 2

a. Implement your own logistic regression with regularization algorithm from the lecture slides using Python. (**10 points**)

b. Using the implemented algorithm, train and test the data from the attached file **ckd_data.zip**. (**10 points**)
- Use 80% of each class data to train your classifier and the remaining 20% to test it.
- Run different values of logistic regression regularization parameter (λ). The range of λ is from -2 to 4 and the step is 0.2
- Plot the f-measure of the algorithm's performance on the **training and test sets** as a function of λ:

$$f\text{ - }measure = \frac{2 \times Pre \times Rec}{Pre + Rec}$$

$$\text{where } Pre = \frac{TP}{TP + FP}; \quad Rec = \frac{TP}{TP + FN};$$

and *TP* is the number of true positives (class 1 members predicted as class 1),
TN is the number of true negatives (class 2 members predicted as class 2),
FP is the number of false positives (class 2 members predicted as class 1),
and *FN* is the number of false negatives (class 1 members predicted as class 2).

c. Repeat the procedure in (b) but now using the features normalized with the standardization protocol discussed in the class. (**5 points**)

In the following part of the assignment, you will be working on applying methods and datasets from the scikit-learn library.

Problem 3

Apply three clustering techniques to the handwritten digits dataset. Assume that k = 10. **(25 points)**

a. K-means clustering (implemented in Problem 1).
b. Agglomerative clustering with Ward linkage (sklearn.cluster.AgglomerativeClustering).
c. Affinity Propagation (sklearn.cluster.AffinityPropagation).

The dataset you will be working with is the handwritten digits and the details can be found here.

Assess all three clustering algorithms using the following protocol:

i. Each cluster should be defined by the digit that represents *the majority* of the current cluster. For examples, if in the second cluster, there are 60 data points of digit "5", 40 of "3" and 25 of "2", the cluster is labeled as "5".

ii. Report the 10x10 confusion matrix by comparing the predicted clusters with the actual labels of the datasets. If the clustering procedure resulted in less than 10 clusters, output "-1" in the position to the missing clusters in the confusion matrix.

iii. Calculate the accuracy of each clustering method using the Fowlkes-Mallows index (sklearn.metrics.fowlkes_mallows_score).

Problem 4

Apply three classification algorithms to the same **ckd_data.zip** dataset as in Problem 2. **(15 points)**

a. Support Vector Machine with the linear kernel and default parameters (sklearn.svm.SVC).
b. Support Vector Machine with the RBF kernel and default parameters.
c. Random forest with default parameters (sklearn.ensemble.RandomForestClassifier).

Assess all three classification algorithms using the following protocol:

i. Use 80% of each class data to train your classifier and the remaining 20% to test it.
ii. Report the f-measure of the algorithm's performance on the **training and test sets.**

$$f\text{-}measure = \frac{2 \times Pre \times Rec}{Pre + Rec}$$

where $Pre = \frac{TP}{TP + FP}$; $Rec = \frac{TP}{TP + FN}$;