Homework 4. Due: Monday, Nov. 21, 6pm

Problem 1.

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

(b) (**10 points**) Using the k-means algorithm, cluster the data from the attached file **realdata.txt**. Plot X,Y coordinates for the entire dataset. Use different symbols and colors to represent your datapoints 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.

Problem 2.

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

(b) (**10 points**) Using the implemented algorithm, train and test the data from the attached file **realdata1.zip**:
- 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 set as a function of (λ):

$$f\text{ - }measure = \frac{2 \times Pr \times Re}{Pr + Re}$$

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

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

This part of the homework, you will be working on applying methods and datasets from Scikit-learn library.

Problem 3. (**20 points**) Apply three clustering techniques to the handwritten digits dataset. Assume with k=10:

- k-means clustering implemented above
- Agglomerative clustering with Ward linkage (sklearn.cluster.AgglomerativeClustering)
- AffinityPropagation (sklearn.cluster.AffinityPropagation)

The two latter algorithms can be found here:
http://scikit-learn.org/stable/modules/clustering.html#clustering

The primary dataset you will be working with is the handwritten digits **datasets.load_digits** with description available here:

http://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html#sklearn.datasets.load_digits

Assess the algorithm using the following protocol:

- 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” an 25 of “2”, the cluster is labeled as “5”.

- Report the confusion matrix 10x10 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 of confusion matrix corresponding to the missing clusters.

- Calculate the accuracy of each clustering method using Fowlkes and Mallows index: https://en.wikipedia.org/wiki/Fowlkes–Mallows_index

Problem 4. (**20 points**) Apply three classification techniques to the same realdata1.zip dataset as in Problem 2:

- Support Vector Machine with the linear kernel and default parameters (sklearn.svm.SVC)
http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

- Support Vector Machine with the RBF kernel and default parameters (sklearn.svm.SVC)
http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

- Random forest with default parameters (sklearn.ensemble.RandomForestClassifier)
http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Train and test the algorithms using the data realdata1.zip from Problem 2:
- Use 80% of each class to train, 20% to test
- Report the f-measure of the algorithms' performance on the training set:

$$f\text{-}measure = \frac{2 \times Pr \times Re}{Pr + Re}$$

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

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).