Homework 3

COMP 7150/8150

FALL 2016

Instructor: Deepak Venugopal

Due Date: November 3 2016 (submit code to ecourseware)

1. Clustering for document classification (65 points)

In this assignment, you will use clustering algorithms to classify documents. Load the 20 newsgroups dataset from sklearn using the following package

from sklearn.datasets import fetch_20newsgroups

Load the training and test data sets using
newsgroups_train = fetch_20newsgroups(subset='train')
newsgroups_test = fetch_20newsgroups(subset='test')

Each instance is a document and has a specific topic. The list of topics can be listed using newsgroups_train.target_names

The topic corresponding to each training (or test) instance in the dataset can be obtained using
newsgroups_train.target

Your task is to use clustering algorithms to cluster the documents based on the bag of words model. Specifically, you will convert each document to a TFIDF vector and then run the K-Means and Gaussian Mixture Models algorithms.

For evaluation, you will compute the weighted F-1 score on the test set. Specifically, train the clustering algorithm on the training set. For each test instance, you will predict the cluster to which it belongs and assign the predicted topic to the test instance as follows. The topic corresponding to the test instance is equal to the majority topic of the cluster. For instance, after training, suppose the cluster has 10 elements, and 8 of them are of type 0, then we will say the cluster's type is 0, and any test instance that will predicted to belong to this cluster will be of type 0.

Perform the above evaluation for K-Means and Gaussian Mixture Models. Report your results for at least 5 different parameter settings (varying k for k-means, varying the number of mixture components for GMM).

2. Linear Regression (35 points)

   In this question, you will experiment with Lasso and Ridge regression.

   Load the Boston housing prices dataset from sklearn as,

   from sklearn.datasets import load_boston

   Perform Lasso and Ridge Regression on this dataset. Compare the performance with tuning the parameters and without tuning the parameters.
   Specifically, use a subset of the training set as a validation set. Tune the parameter alpha (regularization strength) for Ridge and Lasso using the validation set. Report the Mean-Squared-Error on the test set using the tuned parameters. Compare this result with the results you get without tuning the parameters, i.e., use the default parameters for training on the full training data set and test on the test set.