COMP 135 – Machine Learning – Fall 2015

Empirical/Programming Assignment 3

Due date: Thursday, November 5 (by the beginning of class, both paper and electronically)

1 Introduction

In this assignment you will implement and test properties of the k-means algorithm.

2 Data

For this assignment we use 4 datasets. The first, generated artificially, was constructed to have 3 underlying distinguishable classes but a high level of noise. The other 3, *iris*, *ionosphere* and *seeds* are from the UCI machine learning repository. The features in all files have been normalized using the Z-score method so that no further pre-processssing is required. We explicitly provide the class labels for the purpose of calculating NMI with the labels, but the class feature should not be used for the purpose of clustering. The data files are accessible through the course web page.

3 Your Tasks

3.1 Implementing k-Means

Implement the k-means algorithm in order to cluster the data mentioned above. Here you can reuse your arff parser from project 1 to read the arff files. We highlight two points for the implementation: (1) To ensure modularity, write a function $dist(i, j)$ to calculate the distance between examples i and j in the data file and use it in the k-means code. As noted above, although this information is given in the same file, please make sure *not* to use the class label for distance calculation. (2) To initialize k-means please choose a random selection of k points from the dataset as the initial means. Make sure to seed the random number generator so that the results are easy to reproduce.

3.2 Sensitivity of k-means to initialization

For each of the 4 datasets, repeat the following 10 times with different random initializations: run k-means with k set to the number of classes in the dataset. Then calculate both the cluster scatter CS $= \sum_j \sum_{x \in C_j} \|x - \mu_j\|^2$ and the NMI of the computed clustering to the class labels. Explanation and equations for NMI are given in the lecture slides. For each dataset, plot a bar graph for CS and NMI across the 10 runs.

Both CS and NMI can be used as quality criteria for the clustering (except that in a real application we will not have access to labels and hence cannot calculate NMI). Write a short report on the results: are CS and NMI stable across multiple initializations? are their quality judgements in agreement? What other observations can you make from the results?

3.3 Selecting k

Here we attempt to check the “knee criterion” for selecting the number of clusters. In particular, for each dataset run k-means with $k = \{1, \ldots, 15\}$. To avoid the variability observed in the previous part, for each value of k run the algorithm 10 times and pick the clustering result with the smallest CS from these runs. Now plot CS as a function of k.

Write a short report on the results: is there “visual evidence” suggesting that CS be used as a criterion for selecting k? Are the results consistent across the 4 datasets? if not, what observations can you make from the results?

4 Submitting your assignment

- You should submit the following items both electronically and in hardcopy:
  (1) All your code for data processing, learning algorithms, and the experiments. Please write clear code and document it as needed.
  (2) A short report with the plots as requested and a discussion with your observations from these plots.
- **Please submit a hardcopy** in class.
- **Please submit electronically using provide** by 1:30 (class time). Put all the files from the previous item into a zip or tar archive, for example call it `myfile.zip`. Then submit using
  `provide comp135 a3 myfile.zip`.

Your assignment will be graded based on the code, its clarity, documentation and correctness, the presentation of the plots, and their discussion.