IE 582 Statistical Learning for Data Mining
Homework 5, due Final Exam

Instructions: Please solve the following exercises using R (http://www.r-project.org/). Also, you are requested to use certain R packages for this particular homework. You are expected to use GitHub Classroom and present your work as an html file (i.e. web page) on your progress journals (as you did for Homework 0).

Do not share your code (except the one in your progress journals)! As a fundamental principle for any educational institution, academic integrity is highly valued and seriously regarded at Boğaziçi University.

Exercise: Multiple Instance Learning

Wikipedia definition: In machine learning, multiple-instance learning (MIL) is a variation on supervised learning. Instead of receiving a set of instances which are individually labeled, the learner receives a set of labeled bags, each containing many instances. In the simple case of multiple-instance binary classification, a bag may be labeled negative if all the instances in it are negative. On the other hand, a bag is labeled positive if there is at least one instance in it which is positive. From a collection of labeled bags, the learner tries to either (i) induce a concept that will label individual instances correctly or (ii) learn how to label bags without inducing the concept.

In other words, multiple instance learning problems differ from regular learning problems. In traditional classification tasks, each object is represented with a feature vector and the aim is to predict the label of the object given some training data. However this modest approach becomes weak when the data has a certain structure. For example, in image classification, images are segmented into patches and instead of a single feature vector, each image is represented by a set of feature vectors derived from the patches. This type of applications fits well to Multiple Instance Learning (MIL) setting where each object is referred to as bag and each bag contains certain number of instances.

In this exercise, you are given a dataset, namely Musk1, dataset description is available on
https://archive.ics.uci.edu/ml/datasets/Musk+(Version+1)
and it is also uploaded to Moodle. Please use the uploaded version. The structure of the file is as follows:

Bag class	Bag Id	Feature 1	Feature 2	…	Feature *p*
1	1				
1	1				
1	1				
1	1				
0	2				
0	2				
…	…	…	…	…	…
1	N				
1	N				
1	N				

For example, there are *N* bags and *p* features in the dataset illustrated above. First bag (The bag with id 1) is from first class and has 4 instances. Similarly second bag (the bag with id 2) has 2 instances and the bag is from class

0. The aim of multiple instance learning is to classify a bag given its instance characteristics. A typical example is smell classification.

From dataset definition: Musk1 describes a set of 92 molecules of which 47 are judged by human experts to be musks and the remaining 45 molecules are judged to be non-musks. The goal is to learn to predict whether new molecules will be musks or non-musks. However, the 166 features that describe these molecules depend upon the exact shape, or conformation, of the molecule. Because bonds can rotate, a single molecule can adopt many different shapes. To generate this data set, the low-energy conformations of the molecules were generated and then filtered to remove highly similar conformations. This left 476 conformations. Then, a feature vector was extracted that describes each conformation.

Normally, instance labels are not known in multiple instance learning. We only know the labels of bags. The structure of the dataset assumes that each instance has the same label as its bag (not a realistic assumption in the context of multiple instance learning). Most of the approaches in MIL literature aim at summarizing the instance level information to bag level information (i.e. there are 4 instances in first bag, how can I represent the first bag as a single feature vector). Suppose we use the following algorithm to transform the instance-level information to bag-level representation.

Step 1: Cluster instances with some clustering algorithm
Step 2: For each bag: Calculate the distance of its instances to each cluster centroid (i.e. each instance of the bag is represented by a feature vector of length *r*)
Step 3: Represent the bag as the average of its instance distances to the cluster centroids (a bag is represented by a feature vector of length *r).*

If I represent all bags in the same manner, the problem can be considered as a regular learning problem (i.e. I have *N* data points with *r* features). Figure 1 illustrates the idea.

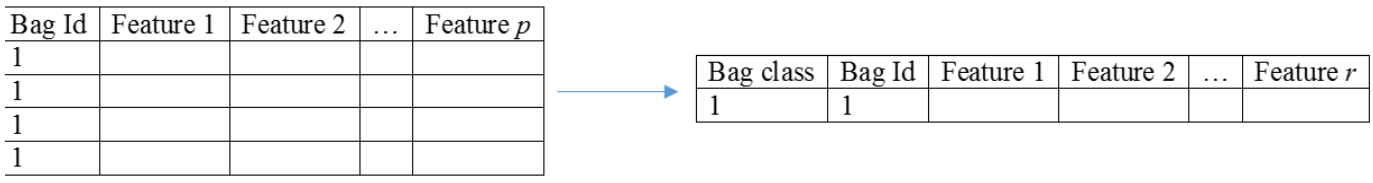

Figure 1. Summarizing instance level information to bag level information. Part (b) proposes to take the column means as the summary.

Task:
Step 1 requires a distance measure and clustering algorithm selection. Propose two distance measures for computing similarity between the instances. For each distance measure alternative, you are expected to evaluate the performance of two clustering approaches: *k*-means and hierarchical clustering. Since you are asked to work with distances, you need to use a modified version of *k*-means which is called the *k*-medoids approach. After clustering, perform the remaining steps to end up with the bag-level representation.

Suppose we decided to train lasso logistic regression on the bag-level representation for classifying bags for Musk1 dataset. Specify the best set of parameters for representation learning (i.e. distance measure, clustering approach and number of clusters) and summarize the area under the ROC curve (AUC) performance based on 10-fold cross-validation on the training data.