IE 582 Statistical Learning for Data Mining
Homework 4, due December 27th, 2021

Instructions: Please solve the following exercises using R (http://www.r-project.org/) or Python (https://www.python.org/). You are expected to use GitHub Classroom and present your work as an html file (i.e. web page) on your progress journals. There are alternative ways to generate an html page for you work:

- A Jupyter Notebook including your codes and comments. This works for R and Python, to enable using R scripts in notebooks, please check:
  - https://docs.anaconda.com/anaconda/navigator/tutorials/r-lang/
  - https://medium.com/@kyleake/how-to-install-r-in-jupyter-with-irkernel-in-3-steps-917519326e41

  Things are little easier if you install Anaconda (https://www.anaconda.com/). Please export your work to an html file. Please provide your *. ipynb file in your repository and a link to this file in your html report will help us a lot.

- A Markdown html document. This can be created using RMarkdown for R and Python Markdown for Python

Note that html pages are just to describe how you approach to the exercises in the homework. They should include your codes. You are also required to provide your R/Python codes separately in the repository so that anybody can run it with minimal change in the code. This can be presented as the script file itself or your notebook file (the one with *.ipynb file extension).

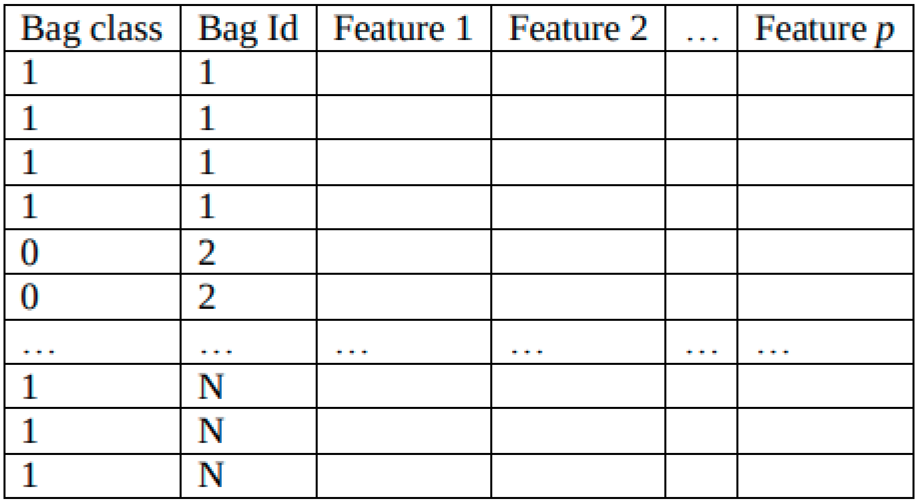
The last and the most important thing to mention is that academic integrity is expected! Do not share your code (except the one in your progress journals). You are always free to discuss about tasks but your work must be implemented by yourself. As a fundamental principle for any educational institution, academic integrity is highly valued and seriously regarded at Boğaziçi University.

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 [Küçükaşcı, Emel Şeyma, and Mustafa Gökçe Baydoğan. "Bag encoding strategies in multiple instance learning problems." Information Sciences 467 (2018): 559-578.].

In this task, 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 illustrated in Figure 1



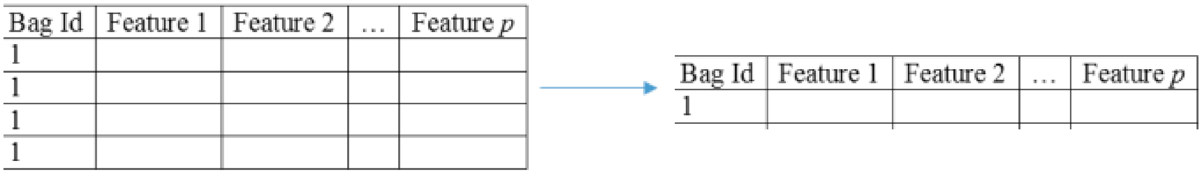
Bag class	Bag Id	Feature 1	Feature 2	…	Feature *p*
1	1				
1	1				
1	1				
1	1				
0	2				
0	2				
…	…	…	…	…	…
1	N				
1	N				
1	N				

Figure 1. Structure of the provided data

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?). One easy way to represent the bag is to take the average of the instance features so that the bag is represented by the center of its instances. 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 p features). Figure 2 illustrates the idea.



Bag Id	Feature 1	Feature 2	…	Feature *p*
1				
1				
1				
1				

Bag Id	Feature 1	Feature 2	…	Feature *p*
1				

Figure 2. Summarizing instance level information to bag level information. The column means can be used as the bag-level representation. This representation does not benefit from the labels. Hence, it is unsupervised.

Task:

Suggest two alternative bag-level representations for the given multiple instance learning problem. Based on the proposed bag-level representations for Musk1 dataset, evaluate at least two reasonable classifiers (of your choice). The example representation is very simple and introduced for illustration purposes. Please do not consider it as a valid alternative. You are expected to justify your bag-level representation approach. This may be achieved by some descriptive analysis you perform (to motivate your approaches). You can benefit from the literature as long as you provide your references accordingly. Please note that if the proposed representation approach has certain parameters, these parameters are also part of your algorithm. Hence, representation parameters (if there is any) should also be tuned together with the parameters of your proposed approach. Specify the best set of parameter combination for the proposed representations and your classifier. Use the accuracy based on 10-fold cross-validation on the training data as your primary performance metric in your evaluations.