CMPT 459 Fall 2020
Data Mining
Martin Ester
TAs: Madana Krishnan Vadakandara Krishnan and Rhea Rodrigues

Assignment 2

Total Marks: 100

Assignment 2.1 (55 marks)

We want to cluster categorical data, i.e. data that have categorical attribute domains. The task of this assignment is to develop the equivalent of the *k*-means algorithm for categorical data, called the *k-modes* algorithm. We assume the following distance function (Hamming distance) for pairs of categorical objects:

$$dist(x,y) = \sum_{i=1}^{d} \delta(x_i, y_i) \;\; with \;\; \delta(x_i, y_i) = \begin{cases} 0 \;\; if \;\; x_i = y_i \\ 1 \;\; else \end{cases}$$

Instead of the means, we choose the mode *m* as the representative of a cluster *C*. In every attribute, the mode takes the attribute value that is most frequent within the cluster. More formally, the mode *m* of cluster *C* is defined as follows:

$$m = (m_1, ..., m_d) \;\; \text{such that} \;\; \forall\, j, 1 \le j \le d, \; \forall a \in A_j : |\{x \in C \mid x_j = m_j\}| \ge |\{x \in C \mid x_j = a\}|$$

a) (30 marks)

Prove that *m* is the object minimizing the cluster cost

$$TD(C,m) = \sum_{p \in C} dist(p,m)$$

Hint: first formulate the intuition of the proof, then formalize it. The proof can be performed by contradiction.

b) (25 marks)

Provide a pseudo-code for the *k*-modes algorithm with the following header:

k-modes(dataset, numberOfClusters)

To initialize the *k* cluster representatives, take a random sample of *k* objects from the dataset.

Assignment 2.2 (45 marks)

Implement the K-modes algorithm in python according to your pseudo-code.

Your program will accept two inputs: the dataset and numberOfClusters.

As output, write a new csv file in which every row consists of the example and the corresponding cluster number which was assigned to it by your algorithm.

You will be using the Mushroom Data set which you can download here. Read the Data Set Description found on the website to understand the data. Note that the first column is the class label, followed by 22 columns of categorical attributes.

You must handle missing values in the dataset. Mention how you handled the missing values in your PDF Report (see submission details below).

Follow these instructions for your implementation:

- Prepare the data
    - Handling the missing values (5 marks)
    - Brief description of the above in the report (5 marks)
- Initialization of centroids (5 marks)
- Assignment of observations to a cluster (5 marks)
- Compute new centroids (10 marks)
- Terminating criteria for iterations (10 marks)
- Saving results to a file (5 marks)

You can use libraries including math,numpy, scipy, random, etc. Do not use the scikit-learn libraries. Implement your own logic to compute the distance between observations. You will be marked on the correctness of your implementation.

Submission details

1. Submit a PDF Report for the solutions to Assignment 2.1. In your report, for Assignment 2.2, briefly explain how you decided to handle the missing data and why you handled it in this way.
2. For Assignment 2.2 submit a python program that reads the data set as input (assume the data is in the same directory as the program).