**CS 484 Data Mining**

**Fall 2015**

**Professor Jessica Lin**

**HW 1 – Due 9/15**

**Total: 90 points**

**Question 1 (60 points)**

For the data set described below, give an example of the types of data mining questions that can be asked (one for each classification, clustering, association rule mining, and anomaly detection task) and the description of the data matrix (what are the rows and columns). If necessary, briefly explain the features that need to be constructed. Note that, depending on your data-mining question, the row and column definitions may be different.

Example data: telecom company database

|  |
| --- |
| DM Task: Classification |
| **Data Mining Question**: Will the customer churn soon after his/her contract expires?  **Row**: A customer  **Column**: Customer account attributes, such as age, income, length with the company, number of calls to customer service, overage charges, data usage, and a target variable of whether the customer churned or stayed with the company. |

|  |
| --- |
| DM Task: Clustering |
| **Data Mining Question**: Do the customers form natural groupings?  **Row**: A customer  **Column**: Customer information such as age and income.  (Note: We can also cluster customers based on their calling behaviors, in which case we would need call data such as: number of outgoing calls, number of international outgoing calls, average call duration, total call duration, number of calls on the weekend, number of calls at night, number of calls during work hours, etc.) |

|  |
| --- |
| DM Task: Association rule mining |
| **Data Mining Question**: What plans, add-ons or services are frequently purchased together, by what kind of customers?  **Row**: A customer  **Column**: Customer account attributes, as well as different services that the customer has in his/her contract, such as data plan type, text messages, international calls, etc. |

|  |
| --- |
| DM Task: Anomaly detection |
| **Data Mining Question**: Telecommunication fraud detection (is the call fraudulent?)  **Row**: A call  **Column**: Call data record such as caller and called identification numbers, date and time, call type (voice, text), destination, dropped call, etc., and potentially a column indicating whether the call was fraudulent (if using supervised learning) |

1. Web pages
2. Online movie rental store database
3. From a past job, hobby, or interest, come up with your own dataset example.

**Question 2 (15 points)**

A store finds that, through association rule mining, that customers who buy diapers are likely to also buy beer. What can the store do with this information to boost their sales? Provide 3 suggestions. Be as specific as you can.

**Question 3 (15 points)**

For each of the three datasets described in Question 1 (including the last one that you come up with), in order to cluster the objects, you need to measure the similarity or distance between them. For each dataset, name the appropriate distance or similarity measure you will use. The answer could be as simple as “Euclidean distance” (or other more appropriate distance measures depending on the data) but you should also describe if any pre-processing is needed on the row vectors before computing the distance/similarity.

Sample answer on the “example data” (telecom database):

(From Question 1)

|  |
| --- |
| DM Task: Clustering |
| **Data Mining Question**: Do the customers form natural groupings?  **Row**: A customer  **Column**: Customer information such as age and income. |

We can use Euclidean distance to compute the distances between the customer vectors. We need to normalize each attribute, e.g. to the [0, 1] range, otherwise the income attribute would dominate the distance value. If we have calling behavior data, the same approach can be used.