**CS 584 Data Mining**

**Spring 2016**

**Professor Jessica Lin**

**HW 1 – Part 1**

**Due 2/9/16**

**Question 1 (10 points)**

Plumbing Inc. has been selling plumbing supplies for the last 20 years. The owner, Joe, decides that next year it is time to diversify by adding gardening tools to the products. Having had success using customer data to build predictive models to guide direct mail campaigns for special plumbing offers, he considers that data mining could help him to identify a subset of customers who should be good prospects for his new set of products. Is Joe ready to solve this as a supervised learning problem? If yes – how? If no – why not?

**Question 2 (20 points)**

The sinking of RMS *Titanic* is one of the most well known shipwrecks in history (thanks in part to the 1997 film directed by James Cameron). The sinking resulted in more than 1,500 deaths of crews and passengers. One of the reasons for such a large number of deaths is that there were not enough life boats on the ship. It is known that certain groups of passengers (e.g. women, children, and upper-class passengers) were more likely to survive than others. Answer the following questions.

1. Suppose you want to do an analysis on the passengers and their chance of survival. More specifically, you want to predict whether a passenger on Titanic would survive using the Titanic passenger information. Describe how you would solve this problem, similar to what we did on slides #34-38 from the first lecture (Introduction). First describe the “goal” using one sentence, and then describe your approach, e.g. what kind of data you need, how to construct the training data, etc.
2. Instead of using supervised learning (classification), which other data mining task that we discussed in class (clustering, association rule mining, and anomaly detection) can you use to find out what kind of passengers were likely to survive (or not survive)? Describe the approach.

**Question 3 (48 points)**

For each of the data sets described below, give an example of data mining questions that can be asked on the data set (one for classification, one for clustering, one for association rule mining, and one for 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.

1. Online movie rental store database

|  |
| --- |
| DM Task: Classification |
| **Data Mining Question**: **Row**: **Column**:  |

|  |
| --- |
| DM Task: Clustering |
| **Data Mining Question**: **Row**: **Column**:  |

|  |
| --- |
| DM Task: Association Rule Mining |
| **Data Mining Question**: **Row**: **Column**:  |

|  |
| --- |
| DM Task: Anomaly Detection |
| **Data Mining Question**: **Row**: **Column**:  |

(b) News articles

|  |
| --- |
| DM Task: Classification |
| **Data Mining Question**: **Row**: **Column**:  |

|  |
| --- |
| DM Task: Clustering |
| **Data Mining Question**: **Row**: **Column**:  |

|  |
| --- |
| DM Task: Association Rule Mining |
| **Data Mining Question**: **Row**: **Column**:  |

|  |
| --- |
| DM Task: Anomaly Detection |
| **Data Mining Question**: **Row**: **Column**:  |

Here is an example provided for you.

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) |