CSCI-B 565 DATA MINING
Homework 1
Morning class
Computer Science Core
Fall 2013
Indiana University

Ramakant Khandel
khandelr@indiana.edu

September 5, 2013

1. Definitions

Data Mining:

Data Mining is known as the process of finding pattern in the data.

Machine Learning:

Machine Learning is known as the process in which computer generates rules based on data to improve performance.

Probability:

Probablity is defined as likelihood that the event will occur.

Statistics:

Statistics is used for interpreting and analysing large data.

Pattern:

The repeated design in the data is known as pattern.

Consistency:

regular, uniform

Prediction:

forecast

Feature:

distinct attribute

Random Variable:

A quantity or variable having numerical value.

2. (a) **Assume the data are linearly separable; that is, a hyperplane exists that has 0s on one side and 1s on the other. Give a proof that this random ML algorithm will eventually converge to a correct classifier.**
The algorithm tries to divide the data sets in two parts and tries to find the solution with zero error point or minimal error point.

   (b) **Discuss how you are designing and implementing 'mu' and threshold of error.**
The measure of correctness for a particular function is calculated by subtracting the error count from the total number of data sets and dividing by same. The threshold of error is defined as 0.3.

   (c) **Assume that the label has three distinct values, instead of two. How could you reasonably easily modify this algorithm to distinguish three classes?**
The hyperplane divides the data sets in two parts. This algorithm can be modified to distinguish three classes by assigning values to data sets in 2,3 and 4 quadrant as -1.

   (d) **Discuss potential problems with the classification of the hyperplane.**
The algorithm is used to divide the data sets in two parts and ideally finds the solution with zero error points. This desired result is not possible each time. There are situation in which error count for the algorithm is too high ie it fails to classify the data sets in two parts.

   (e) **The method of generating data is artificial maybe too much so. Explain**
The algorithm is used to divide the data sets in two parts. This algorithm is independent of the data sets and tries to divide the data sets in two parts with minimum error points. The algorithm is designed in such a way that it works irrespective of the data sets.

   (f) **Imagine increasing many fold; perhaps = N100000. A curious phenomenon occursas the dimensions increase theres less apparent difference between the concepts of near and far.**
A line is a hyperplane in two dimension and a plane is hyperplane in three dimension. As the dimension increases the equation of the hyperplane will change in order to divide the data sets in two parts. As the dimension increses the distance between the data sets will change.