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

Debpriya Seal
debseal@indiana.edu

September 5, 2013

All the work herein is solely mine.

Problems

Problem 1

(a) Data Mining
Answer: Data Mining is a relatively new field resulted from confluence of inter-disciplinary fields viz. Database Systems, Machine Learning and, Statistics. It is a process of analysing/studying the huge data and looking for any information(or pattern) which could be helpful w.r.t increase in revenue for company. Data Mining uses sophisticated mathematical algorithms to implement this. Data Mining can even forecast/predict things which cannot not be told otherwise.

(b) Machine Learning
*Answer:*Machine Learning is a branch out of Artificial Intelligence. To put very crudely, it means to make computers do things without explicitly coding for it. And it achieves this by designing/developing algorithms which evolve its behaviour based on the training data.

(c) Probability
*Answer:*Probability is again a branch of Mathematics dealing with likelihood of something happening. Crudely it is the chance/likelihood of an event occurring, represented within a scale of 0 to 1. With 1 representing that event will be occurring for sure and with 0 for possibility of never occurring at all.Mathematically, it is the number of targeted event by total possible outcomes.

(d) Statistics
*Answer:*To say it simply, it means making sense of data.It is a interdisciplinary science to extract inforamation from data using mathematical models. It tries to find something in the data that can be

useful to the company. For example a pattern of causation or correlation can be very fruitful for a company to know.

(e) Pattern
*Answer:*Pattern is a intelligible and recognizable regularity in data(in terms of machine learning). Pattern could be pertinent to anything. And pattern recognition is a common application of machine learning.

(f) Consistency
*Answer:*Literally it means in-agreement.And it is broad term, it is used in various ways like data consistency ,algorithm consistency. However, what it means is that we can rely on it. It is consistent in what it is supposed to do.

(g) Prediction
*Answer:*In world of Data Mining, it means that based on empirical(learning) data, we tell something about future.Now prediction can be of various types. Classification, Forecast, Clustering to name a few.

(h) Feature
Answer: Feature is basically a Machine Learning term. It is also known as attribute, field and, variable. Any observation can have many attribute to it like for the price of apartment. Price of apartment can be attached to many factors like, No. of BHK, beach facing to name a few. Now these can be thought of as the feature of a observation. Which defines about the observation i.e. price of apartment. Feature are generally classified as *quantitave* and *categorical.*Categorical Features would be like Male or Female, Selected or rejected etc.While quantitative is more like any numerical value.

(i) Random Variable
*Answer:*We want the outcome of every experiment to be a number.However, the outcome of every experiment is not a number like, toss of a coin. Random Variable is function which maps every outcome with an unique number.Random variable can be of two type *discrete* and *continuous.*With discrete random variable, we can count the number of values it can take on(*example: Toss of coin*).However, with conintous random variable can take on infinite numbers(*example: Amount of rainfall in a year*).

Problem 2

(a) Description of the ML Algorithm Implemented. *Answer:* Model Used: Logistic Regression using Maximum Likelihood Estimation (MLE)
There are 2 caveats with this:

  i. At times the parameter might be computed to ∞. That is generally due to sparse data. But ofcourse parameter which tends to infinity will not converge. However, there is way to make it converge by skipping those values. But stilll that does not guarentee convergence.
  ii. Based on the learning rate, the model may keep on oscillating. Hence, never converging to the best solution.

(b) 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.

Answer: As per Prof. Dalkilic hint Expected Value would play a key role to help us prove that this random algorithm will converge.The formula foror Expected Value
$E(X) = \sum_{i=1}^{n} x_i$
So let's see what is expected number of times we need pick a X coordinate to get the X on the line of classfier(if it exist). [h]

Table 1: Δ

Number of picks	Probability	$E(X)$
1	$\frac{1}{99}$	$1 * \frac{1}{99}$
2	$\frac{98}{99} * \frac{1}{99}$	$2 * \frac{98}{99} * \frac{1}{99}$
3	$\frac{98}{99}^2 * \frac{1}{99}$	$3 * \frac{98}{99}^2 * \frac{1}{99}$

Let $S = 1 * \frac{1}{99} + 2 * \frac{98}{99} * \frac{1}{99} + 3 * \frac{98}{99}^2 * \frac{1}{99} + ...\infty$
$\frac{99}{98}S = \frac{1}{99} + \frac{2}{99} * \frac{99}{98}^0 + \frac{3}{99} * \frac{98}{99}^1 + ...\infty$
$S(\frac{99}{98} - 1) = \frac{1}{98} + \frac{2}{99} * \frac{99}{98}^0 + \frac{3}{99} * \frac{99}{98}^1 + ...\infty$ -eqn 1.0
Equation 1.0 is a arithmatic -geometric series whose sum for infinity is given by below formula:
$S_n = \frac{a}{1-r} + \frac{dr}{1-r}^2$ when n tends to ∞
Putting the values of a,d and, r: we get $S_n = 99$ when n tends to ∞
Clearly we can see that in certain number of times we can get that point. Similarly we can say for Y coordinate as well.And this algortihm will converge for sure.

(c) Discuss how you are designing and implementing μ and τ.
Answer: Lets start by discussing μ first. For this i am using logsitic/sigmoid function. And my μ looks like below:
$\mu = \frac{1}{m}[\sum_{i=0}^{m} Y_i * \log(h(\theta)) - (1 - Y_i) * \log(1 - h(\theta))]$

where $h(\theta) = \frac{1}{1+e^{ax_1+bx_2+c}}$ and,
Y_i = the training labels given to us. Now clear my hypothesis $h(\theta)$ will always be between 0 and 1 as a property of sigmoid function. Now, when the training data label $Y_i = 0$, the first equation in my μ will become zero and we would be left with only the $2^n d$ equation. And if my hypothesis says it to be 1 which is completely wrong, i will penalize my algorithm and log(0) will become ∞.However, if my hypothesis says it to be 0 then I reeward my algorithm and log(1) will become 0. Similarly for the case when the $Y_i = 1$.
Coming to the τ, I got this value more by seeing it emperically. I ran it for couple of times and found this to be the best fit.

(d) Assume that the label has three distinct values, instead of two. How could you reasonably easily modify this algorithm to distinguish three classes?
*Answer:*As per me the easiest thing would be to implement the algorithm twice. First we will try to classify between first and rest. And then again implement the algorithm for the rest 2 left categories, this time classifying them apart.

(e) Discuss potential problems with the classification of the hyperplane.
Answer: Firstly, it does not scale with increase in dimension.Secondly, it require huge computation power. Otherwise, it gets really slow.

(f) The method of generating data is artificial maybe too much so. Explain

Answer: Clearly, the training data we are generating are sheer random values. However, in real problems we generally have real time data which has some hidden pattern (if any). And for any machine learning algorithm, the training data is of vast importance. This is the very reason that data-cleansing takes the major apart of data mining.Otherwise, the algortihm will learn wrongly resulting into a absurd results.

(g) Imagine increasing δ many fold; perhaps $\delta = \mathbb{N}^{100000}$. A curious phenomenon occurs as the dimensions increase there's less apparent difference between the concepts of near and far.

*Answer:*With high dimension, the error caused by the noise features becomes eventual thing. Also, with distance-based classifier, a point X_i is classified by it is close to any X_i's to class (say) 1. But as the dimension increases, the far and near concept cease to work.