CSCI-567: Assignment # 2

Due on Friday, October 2, 2015

Saket Choudhary
2170058637

Contents

Todo list

Problem 1

Problem 1: (a)

Linear regression assumes uncertainty in th measurement of dependent variables(X). Y being an independent variable is assumed to be the 'true' value that can be measured with ultimate precision. Any model that relates the independent variable to dependent will assume some kind of modelling error which in case of linear regression is often taken as Gaussian random variable . Additively the 'noise' and the dependent variable predict the independent.

Problem 1: (b)

In order to make linear regression robust to outliers, a näive solution will choose "absolute deviation"($L1$ norm) over "squared error"($L2$ norm) as the criterion for loss function. The reason this might work out in most cases(especially when the outliers belong to a non normal distribution) is that "squared error" will blow up errors when they are large. Thus $L2$ normn will give more weight to large residuals, while the $L1$ norm gives equal weights to all residuals.

Problem 1: (c)

A quick way to realise this is to consider the scale. Say one of the dependent variables is 'time'. Rescaling time from hours to seconds will also rescale its coefficient, but the importance remains the same!

Problem 1: (d)

If the dependent variables are perfect linear combination, the matrix XX^T will be non invertible.

Problem 1: (e)

A simple solution would be to start from '0000'(length being equal to the number of levels of the categorical variables) for one of the categories.

Problem 1: (f)

If the independent variables are highly correlated, the coefficients might still be entirely different. So if a feature

Problem 1: (g)

Using a posterior probability cutoff of 0.5 in linear regression is not same as 0.5 for logistic. A 0.5 rehsold on logistic guarantees that the point all points lying to the right belong to one class. However for a regression problem, this is not true, because the predicted value of y is an 'intepolated or extraplolated' In any case, logistic regression is a better choice

Problem 1: (h)

When the number of variables exceed the number of samples, the system is undetermined. And yes, it can be solved by simply obtaining psuedo-inverse of X which is always defined.

Problem 2

Given