CS60050: Machine Learning
Autumn 2015, CSE, IIT Kharagpur
Assignment 1

All questions carry 10 marks each.

Question 1: Consider a dataset where each data point is associated with the following:

- t_n : true label
- x_n : feature vector and
- $r_n > 0$: the importance of n^{th} example.

Consider the error function:

$$E(w) = \frac{1}{2}\sum_{n=1}^{N} r_n\left(t_n - w^T\phi(x_n)\right)^2$$

Derive an expression for optimal w.

Question 2: Download the Boston house price data from:
http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression/housing
Split the dataset randomly into training and test sets in 60:40 ratio. Use the formula derived in the class to learn a regularized linear regression model for various values of regularization parameter: $\lambda = \{\, 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000 \,\}$. Report the training set and test set accuracies for these values.

Question 3: Do the above exercise using regularized least squares regression in Weka.

Question 4: Show that the predictive distribution $P(w|t, x, \alpha)$ of weight vector w, for the MAP inference of linear regression model is Gaussian.

Question 5: Show that the number of possible terms in a D^{th} degree polynomial of M variables is:

$$\sum_{d=1}^{D} \frac{(M+d-1)!}{(M-1)!\,d!}$$

Question 6: Calculate the probability that k coefficients will be non-zero in the optimal M dimensional LASSO solution, when the un-regularized least squares solution is uniformly distributed over an M-dimensional box, $-l \leq w_i \leq l, \forall i = 1 \ldots M$, of very large size l.

Question 7: Generate datasets as follows:

- Compute f(x) = sin(x) varying x from -pi to pi in steps of 0.1.
- Compute t(x) = f(x) + a*\epsilon, where epsilon is a Gaussian random number with zero mean and unit variance. Generate 10 such instance for each x.

- Predict t(x) as a 9^{th} degree polynomial of x using regularized least squares, varying $\lambda = 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000$.
- Repeat the above experiment for $a = 0.1, 0.5, 1, 2, 10$.

Report test set errors for the above experiments, for each combination of (λ, a).

Question 8: Generate 1000 numbers sampled from Gaussian distribution with mean 5 and variance 2. Generate 100 subsets of size 10 and estimate the mean and variance from each subset. Calculate the bias and variance of mean and variance estimators.