Machine Learning-I (CS/DS 706)

Assignment - I

Prof.G.Srinivasaraghavan

Date Posted: Aug 14, 2017	**Submit By**: Aug 24, 2017, Midnight	**Max. Marks**: 20

1. **Sampling (R & Python)**: Pick any dataset with a numeric column C and a couple of statistical aggregates on the column value (mean, median, variance, etc.) — denoted as say τ. Carry out the following experiment on the dataset. Compute τ on C across the entire dataset — call it τ_{c*}. Let n denote the number of rows in the dataset. For this experiment to work well, you would need a fairly large dataset.

   a. Pick a fraction in the range $\rho \in (0, 1)$. Randomly (uniform) sample $n\rho$ elements from the dataset and compute the sample statistic τ_{ci} for C in the sample, where i is the sample index. Repeat this m times, each time with a different sample. Compute the average of the sample statistic

   $$\tau'_{c*} = \frac{\sum_{i=1}^{m} \tau_{ci}}{m}$$

   . Generate plots to illustrate how τ'_{c*} converges to τ_{c*} for different values of ρ and m.

   b. Repeat the above experiment but for varying sample sizes in the same 'run'. Specifically, pick m samples as above but the sample sizes are in turn randomly sampled from a Gaussian. Generate plots showing how convergence changes with μ, σ and m where μ, σ^2 are the mean and variance of the Gaussian from which the sample sizes are picked.

2. **Effect of Noise (R & Python)**: Split the IRIS dataset (one of the readily available datasets in R and Scikit Learn) into a 'training' set consisting of 40 samples each from all the three classes *setosa*, *versicolor*, *virginica*. Keep the rest (10 each) as test data. Run `randomforest` (you can keep all the parameters of the call other than the data itself as the default values) on the training data and measure its accuracy on the test data. Repeat the `randomforest` run on two 'corrupted' versions of the training data and compare the performance of the models generated from the corrupted training data with the those obtained from the original training set. The 'corrupted' versions are:

   a. Training data with the 4 explanatory attributes corrupted with some noise (arbitrarily pick a few data points and add a noise, say sampled from a standard Gaussian, to one of the attributes). Do not tamper with the class labels.

   b. Training data with some of the class labels corrupted (changed to something else) without changing any of the explanatory attributes.