Machine Learning Lab Exam

October 2019 Time 4+8 hours Max marks 100

**Note: For the following problems if the data sets are not divided into train and test datasets, divide the datasets in to two parts one for training and other for testing. Show the appropriate performance measures such as error rate, R2 value, ROC/AUC, Confusion matrix etc. clearly on both train and test datasets. Upload the code and screen shots of the results.**

**Evaluation method: A student can double the marks (maximum is 100 marks) what he gets in the 4 hours of exam time by doing remaining problems as an assignment by next day 6-00 pm.**

Download **Computer Hardware Data sets and Breast Cancer Wisconsin (Diagnostic) Data Set**

<http://archive.ics.uci.edu/ml/datasets/Computer+Hardware> and <http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/>. **Use first data set for problem 1 and use second one for all other problems**

1. Estimate the relative performance values of CPU by using a linear regression method. Show clearly in a table learning parameter, number of iterations and cost of the regression at least for three cases. (10 Marks)
2. Use Logistic Regression to classify a tumor is malignant or benign and tabulate results for different threshold input parameters. (10 Marks)
3. Build a Logistic Regression model for each feature to measure its correct classifying power. Select only impactful features to build Logistic regression by using forward feature selection method. (Select the top such feature and add one by one feature from the remaining and build a Logistic Regression model to select a feature which improves correct classifying power most. Stop this process of selecting features once the correct classifying power doesn’t improve). (10 Marks)
4. Build L1 and L2 regularized logistic regressions to determine the tumor is malignant or benign using the features selected by forward feature selection method in above section problem 3. Describe how the variance and bias for these 2 models improve or decrease. (10 Marks)
5. Use Feed Forward Error Back Propagation Algorithm to classify a tumor is malignant or benign. (10 Marks)
6. Using SVM, build a tumor classifier. Use two different kernels and compare results. Upload your code along with a document of results. State the kernels you chose and the corresponding error rates. (10 +10=20Marks)
7. Use K means clustering to segment patients in to groups and find out whether the grouping helps to us identify that a patient in particular groups are more frown to have malignant tumor.(Don't use the label column for this purpose). State the number of centroids you chose and the corresponding malignant rate in each cluster. (10 Marks)
8. Reduce the data set in to 2 dimensional by using PCA. get top 2 principal components. (Don't use the label column for this purpose). State all principal components you got and how you chose top 2. (10 Marks)
9. Find anomalies in the above 2-dimensional data with each column and find out if there is a relation between anomaly and malignant tumor. In other words, can you say anomaly has more probability to be a malignant tumor. State malignant tumor rate in each type of anomaly. (10 Marks)