IIT Bhilai

Department of EE & CS

Machine Learning: CS550, 2018-19, M

Assignment 1 **Due date:** 28 Aug 2018

Note: Install anaconda environment and use spyder/jupyter to work on your programs. You can also work on H2O environment.

Question 1

You will use data https://www.kaggle.com/mirichoi0218/insurance/home available at Kaggle. The problem is to predict medical insurance cost using a regression model.

- Divide the data into training, cross-validation and testing data using the following methods.
    - Holdout method with random sampling
    - Stratified holdout method.
    - Stratified k-folds cross-validation.
    - Bootstrapping
- Write a function to make predictions of the output given the input feature. Your objective is to find the simplest model with least error. The data contains 5 features and one output feature- insurance cost.
    - Using different features combination, find the one that gives you the least error. What measure would you use to compute error: MSE, RMSE, Sum-of-Error-Squares.

Plot errors for different models that you try for one chosen error measure. Label your graph properly. Submit your final model and the error graph.

Question 2

You will use data on telco customer churn data available at Kaggle. The problem is to predict if a customer with a profile will stay or leave using a classification model.

- Divide the data into training, cross-validation and testing data using the following methods.
    - Holdout method with random sampling
    - Stratified holdout method.
    - Stratified k-folds cross-validation.
    - Bootstrapping
- Write a function to make predictions of the output given the input feature. Your objective is to find the simplest model with least error. The data contains 20 features and one output feature- churn or not.

  write function to give prediction in end

  Reduce the no. of features
    - Using different features combination, find the one that gives you the least error. What measure would you use to compute error: misclassification rate, tp rate, fp rate and ROC curve.

Plot errors for different models that you try for one chosen error measure. Label your graph properly. Submit your final model and the error graph.

Save the model