Assignment 1

Foundations of Machine Learning (**CS564**)

Department of CSE, IIT Patna

Date:- 19-Aug-2018
Marks:- 20 Marks

Instructions:

1. All the assignments should be completed and uploaded by **26-Aug-2018, 11.00 pm.**
2. Markings will be based on the correctness and soundness of the outputs. Marks will be ***deducted in case of plagiarism***.
3. Be precise for your explanations in the report. Unnecessary verbosity will be penalized. Prepare a Detailed report of the assignment.
4. Code should be done in ***Python*** or ***R***.
5. You should zip all the required files and name the zip file as ***Group_no*.zip**, eg. **Group_13.zip.**
6. Upload your assignment (**the zip file**) in the following link: https://www.dropbox.com/request/ZJ0hsBb14A93EYx8wMi7

• The goal of this assignment is to experiment with feature extraction methods, linear methods for regression and logistic-regression.

Brief Description of dataset:
Dataset Download Link: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/
Download winequality-red.csv

Dataset Name	Wine Quality	Dataset Characteristic	Multivariate
Number of instance	1600	Attribute Characteristic	Real
Number of Attribute	12		

Attribute Description:

Attribute No	Description
1	fixed acidity
2	volatile acidity
3	citric acid
4	residual sugar
5	chlorides
6	free sulfur dioxide
7	total sulfur dioxide
8	density
9	pH
10	sulphates
11	alcohol Output variable (based on sensory data)
12	quality (score between 0 and 10)

Instruction Regarding Dataset:
Apply 5-fold cross validation on the dataset.
(https://en.wikipedia.org/wiki/Cross-validation_(statistics))

Questions
Linear Regression
1]Learn a linear classifier on the above dataset by using regression on **alcohol variable (feature no 11)**. Report the
a)Predicted alcohol content from learning(individual test data)
b)Report the average **residual sum of squares (RSS)** over these 5 folds.
(https://en.wikipedia.org/wiki/Residual_sum_of_squares)

Regularized Linear Regression

2] Use Ridge-regression on the above data on **alcohol variable (feature no 11)**. Repeat the experiment for different values of λ(parameter).

a) Report the residual error for each fold

b) Which value of λ gives the best fit?

(The value of lambda determines the importance of this penalty term. When lambda is zero, the result will be same as conventional regression; when the value of lambda is large, the coefficients will approach zero.)

Logistic Regression

3] Perform multi-class(0-10 classes) Logistic Regression on **variable quality (feature no 12)**.

a) Report per-class precision, recall and f-measure for each fold.
b) Report the average per-class precision, recall and f-measure for 5 fold cross-validation.
c) Report the misclassification (provide confusion matrix) and also carry out the error analysis on few misclassified instances.