CSE 572
Data Mining
Instructor- Prof. Arunabha Sen

Assignment 4

Total marks : 20

Submission Deadline : 04.10.2019 11:59 pm AZ time.

- For submission, you should submit codes and a PDF report containing the results in a zipped file (only one submission per group). The PDF file should contain names of all the members. The zipped file name should be in the following format:

  GroupName_GroupID.zip [eg.- DM_12.zip]

- Refer to the 'group formation sign-up' sheet in the blackboard for group ID and GroupName (Group name should be the First name of Member 1. Group ID can be obtained from the first column).

- For coding you can use both Matlab and Python. For Matlab and Python codes, include .m and .py files in the zipped folder respectively.

In this Assignment, you are required to build models for (a) regression and (b) classification

You are allowed to use linear regression and Decision Tree libraries.

Task 1) [5 marks]

Regression

Datasets "PB1_train.csv" and "PB1_test.csv" have three columns- first two columns are the **features** (**x**=[x_1,x_2]) and the last column is the prediction value (y). The hypotheses formula is given as:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

Train a linear regression model, *M*, over **x** and y values from "PB1_train.csv" and report the corresponding model parameters (θ_0, θ_1 and θ_2).

Test model *M* on “PB1_test.csv” and report the predicted values ($\bar{y}$) for each row. Calculate the mean-squared-error between the predicted values and original values (third column in “PB1_test.csv”).

Generate a 3-Dimensional plot from the (**x**,y) values of “PB1_test.csv” and add the best fit plane (regression plane generated by *M*) to the plot.

Now given a point (46, 53), what is its corresponding predicted y-value?

Deliverables:

[1] Model parameters (θ values), **[2]** Predicted Values on “PB1_test.csv”, **[3]** Plot,

[4] mean-squared-error on the test set, **[5]** y-value for **x** = [46,53].

Task 2) [5 marks]

Regression

This task is similar to Task 3, but with a different dataset.

Similar to the first Dataset, datasets “PB2_train.csv” and “PB2_test.csv” have three columns, first two columns are the features (**x**=[x_1,x_2]) and the last column is the prediction value (y). The hypotheses formula is given as:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

Train a linear regression model, *M*, over **x** and y values from “PB2_train.csv” and report the corresponding model parameters (θ_0, θ_1 and θ_2).

Test model M on “PB2_test.csv” and report the predicted values ($\bar{y}$) for each row. Calculate the mean-squared-error between the predicted values and original values (third column in “PB2_test.csv”).

Generate a 3-Dimensional plot from the (**x**,y) values of “PB2_test.csv” and add the best fit plane (regression plane generated by *M*) to the plot.

Now given a point (19, 76), what is its corresponding y-value?

Deliverables:

[1] Model parameters (θ values), **[2]** Predicted Values ($\bar{y}$) on “PB2_test.csv”, **[3]** Plot.

[4] mean-squared-error on the test set, **[5]** y-value for **x** = [19,76].

Task 3) [5 marks]

Classification

In this classification problem, you are required to train a Decision-tree model that predicts whether a person is male (represented as 0) or female (represented as 1), given three features: height (in centimeters), age and weight (in kilograms).

Use “PB3_train.csv” and “PB3_test.csv” for this task, where the first three columns represent three features (height, age, weight), and the fourth column represent class label (0/1). Train a decision tree DT (use **Gini-index** metric) on “PB3_train.csv” data that learns to map the mentioned features to their corresponding class values.
Report the predicted values ($\bar{y}$) and accuracy percentage (percentage of matches) of the model DT by testing it on “PB3_test.csv” data.

[1] Accuracy (in percentage) , **[2]** Predicted Values ($\bar{y}$) on “PB3_test.csv”.

Task 4) [5 marks]

Classification

The task is similar to Task 3, but with a different dataset.

Use “PB4_train.csv” and “PB4_test.csv” for this task, where the first three columns represent three features (height, age, weight), and the fourth column represent class label (0/1). Train a decision tree DT (use **Gini-index** metric) on “PB4_train.csv” data that learns to map the mentioned features to their corresponding class values.
Report the predicted values ($\bar{y}$) and accuracy percentage (percentage of matches) of the model DT by testing it on “PB4_test.csv” data.

[1] Accuracy (in percentage), **[2]** Predicted Values ($\bar{y}$) on “PB4_test.csv”.