IE506 Machine Learning: Principles and Techniques **Jan-May 2023**

Assignment 1: Due On 4^{th} February 2023 (11:59 PM IST)

1 Instructions

Answer all questions. Write your answers clearly. You can score a maximum of 60 marks in this assignment.

Use different python notebook (.ipynb) files for each question. Use the same random seed value for all questions so that the splits are same in all questions.

Name the .ipynb files as "IE506_rollno_assignment1_q1.ipynb", "IE506_rollno_assignment1_q2.ipynb" and "IE506_rollno_assignment1_q3.ipynb".

Make sure that your answers and plots are clearly visible in .ipynb files.

Create a folder "IE506_rollno_assignment1" and place all your solution .ipynb files in the folder.

Zip the folder "IE506_rollno_assignment1" to create "IE506_rollno_assignment1.zip". Upload the single zip file "IE506_rollno_assignment1.zip" in moodle.

There will be no extensions to the submission deadline.

Note: Submissions not following the instructions will not be evaluated.

2 Questions

1. For the following questions, **do not use** `scikit-learn` package.
   (a) **[1 mark]** Read the dataset in `file1.csv` into a `pandas` dataframe.
   (b) **[1 mark]** Display the corresponding data description file `desc1.txt` and understand the contents of the data in `file1.csv`.
   (c) **[2 marks]** Replace the column names of data frame with meaningful column names, designed by you using the description in `desc1.txt`.
   (d) **[1 mark]** Display the maximum, minimum, median, first quartile, third quartile information for each column in the dataframe. Use an appropriate `pandas` command.
   (e) **[2 marks]** Use an appropriate `pandas` command to check if any column in the dataframe contains any missing value or not. Drop those rows if there are missing values in the row.
   (f) Consider the weight attribute as the response variable and all other attributes as predictor variables.
   (g) **[2 marks]** Split the data into two sets such that 80% of the data is considered as set $\mathcal{T}_1$ and 20% of the data is considered as set $\mathcal{T}_2$.
   (h) **[4 marks]** Using $\mathcal{T}_1$ as training data, solve an ordinary least squares regression problem and obtain the regression coefficients. Using these coefficients, compute the R^2 and mean squared error values on the training set $\mathcal{T}_1$ and test set $\mathcal{T}_2$.

(i) **[3 marks]** Also plot the residual vs fitted response values for the training set $\mathcal{T}_1$ and test set $\mathcal{T}_2$. Based on the plots, indicate if there are outliers in the training or test sets.

(j) **[4 marks]** Based on R^2 and mean squared error values and on the residual vs fit plots, discuss if the linear regression model assumption is good for the data, and if the model generalizes to unseen data. Also indicate if the residual behavior shows any pattern based on the fitted values.

(k) **[6 marks]** Now, to perform ℓ_2 regularized ordinary least squares regression, choose regularization constant $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000, 10000\}$ using a 5-fold cross-validation on the full train set $\mathcal{T}_1$. Use average R^2 scores of the folds as the criterion to choose best regularization parameter. Perform ℓ_2 regularized ordinary least squares regression with best regularization parameter obtained from the cross-validation procedure on the set $\mathcal{T}_1$ and obtain the regression coefficients. Using these coefficients, compute the R^2 and mean squared error values on the training set $\mathcal{T}_1$ and test set $\mathcal{T}_2$.

(l) **[2 marks]** Also plot the residual vs fitted response values for the training set $\mathcal{T}_1$ and test set $\mathcal{T}_2$. Discuss your observations.

(m) **[4 marks]** Compare the R^2, mean squared error values and residual plots for unregularized and regularized linear regression. Discuss your observations.

2. For the following questions, **do not use** `scikit-learn` package.

(a) Read the dataset in `file1.csv` into a `pandas` dataframe.

(b) Replace the column names of data frame with meaningful column names, designed by you using the description in `desc1.txt`.

(c) Use an appropriate `pandas` command to check if any column in the dataframe contains any missing value or not. Drop those rows if there are missing values in the row.

(d) **[3 marks]** Perform standardization of each column in the data frame and create a new data frame.

(e) **[1 mark]** Display the maximum, minimum, median, first quartile, third quartile information for each column in the dataframe. Use an appropriate `pandas` command.

(f) Consider the weight attribute as the response variable and all other attributes as predictor variables.

(g) Split the data into two sets such that 80% of the data is considered as set $\mathcal{T}_1$ and 20% of the data is considered as set $\mathcal{T}_2$.

(h) **[5 marks]** To perform ℓ_2 regularized least squares regression, choose regularization constant $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000, 10000\}$ using a 5-fold cross-validation on the full train set $\mathcal{T}_1$. Use average R^2 scores of the folds as the criterion to choose best regularization parameter $\lambda^\star$. Perform ℓ_2 regularized least squares regression with best regularization parameter $\lambda^\star$ obtained from the cross-validation procedure on the set $\mathcal{T}_1$ and obtain the regression coefficients. From the regression coefficients obtained, indicate the relative importance of each attribute in predicting the response variable using the magnitude and sign of the regression coefficients. Compare the regression coefficients with those obtained in Question 1, discuss your observations.

(i) **[3 marks]** Rank the attributes in the data set (excluding the bias term) based on non-increasing order of magnitude of the regression coefficients and print them in that order.

(j) **[8 marks]** Consider the top 5 attributes in the previous question as d_1, d_2, d_3, d_4, d_5. Perform ℓ_2 regularized least squares regression with best regularization parameter $\lambda^\star$ on the set $\mathcal{T}_1$ containing only attribute d_1 and response variable. Compute R^2 values on the set $\mathcal{T}_1$ containing only attribute d_1 and response variable. Also compute adjusted R^2 using the formula $1 - \frac{(1-R^2)(n-1)}{n-k-1}$

where n denotes the number of samples in the set $\mathcal{T}_1$ and k denotes the number of predictor attributes in $\mathcal{T}_1$ (note that $k = 1$ now). Repeat the process to perform ℓ_2 regularized ordinary least squares regression with best regularization parameter $\lambda^\star$ on:

- $\mathcal{T}_1$ containing only attributes d_1, d_2 and response variable
- $\mathcal{T}_1$ containing only attributes d_1, d_2, d_3 and response variable
- $\mathcal{T}_1$ containing only attributes d_1, d_2, d_3, d_4 and response variable
- $\mathcal{T}_1$ containing only attributes d_1, d_2, d_3, d_4, d_5 and response variable

Compute R^2 and adjusted R^2 values for each case and the corresponding test set $\mathcal{T}_2$ containing only the corresponding attributes and response variable. Print a table of R^2 and adjusted R^2 values and discuss your observations. Discuss the significance of adjusted R^2 value.

3. For the following questions, **you can use** `scikit learn` package.

   (a) Read the dataset in `file1.csv` into a `pandas` dataframe.

   (b) Replace the column names of data frame with meaningful column names, designed by you using the description in `desc1.txt`.

   (c) Use an appropriate `pandas` command to check if any column in the dataframe contains any missing value or not. Drop those rows if there are missing values in the row.

   (d) Perform standardization of each column in the data frame and create a new data frame.

   (e) Consider the weight attribute as the response variable and all other attributes as predictor variables.

   (f) Split the data into two sets such that 80% of the data is considered as set $\mathcal{T}_1$ and 20% of the data is considered as set $\mathcal{T}_2$.

   (g) [**5 marks**] To perform ℓ_1 regularized least squares regression, choose regularization constant $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000, 10000\}$ using a 5-fold cross-validation on the full train set $\mathcal{T}_1$. Use average R^2 scores of the folds as the criterion to choose best regularization parameter $\lambda^\star$. Perform ℓ_1 regularized least squares regression with best regularization parameter $\lambda^\star$ obtained from the cross-validation procedure on the set $\mathcal{T}_1$ and obtain the regression coefficients.

   (h) [**3 marks**] From the regression coefficients obtained, indicate the relative importance of each attribute in predicting the response variable using the magnitude and sign of the regression coefficients. Indicate if any of these coefficients is zero. Also compare the regression coefficients with those obtained in Question 2, and discuss your observations.
