IE506 Machine Learning: Principles and Techniques **Jan-May 2024**

Assignment 1: Due On 30^{th} January 2024 (11:59 PM IST)

1 Instructions

Answer all questions. Write your answers clearly. You can score a maximum of 50 marks in this assignment.

Answers for Question 1 should be provided in a single pdf file.

Name the pdf file as "IE506_yourrollno_assignment1_q1.pdf".

Use different python notebook (.ipynb) files for each programming based question (questions 2 and 3).

Name the .ipynb files as "IE506_rollno_assignment1_q2.ipynb", "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 `.pdf` and `.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. Recall that the conditional mean of (parametrized model-based estimate of) response variable conditioned on an input vector $\mathbf{x}$ of d attributes is given by $E[Y|X = (x_1, x_2, \ldots, x_d)] = \beta_0 + \sum_{j=1}^{d} \beta_j x_j$. Letting $\mathbf{x} = (x_1\ x_2\ \ldots\ x_d\ 1)^\top$, $\boldsymbol{\beta} = (\beta_1\ \beta_2\ \ldots\ \beta_0)^\top$, note that $E[Y|X = (x_1, x_2, \ldots, x_d)] = \boldsymbol{\beta}^\top \mathbf{x}$. Let us denote $E[Y|X = (x_1, x_2, \ldots, x_d)]$ by $\widehat{y}$ (predicted value).
   Consider a data set $D = \{(\mathbf{x}^i, y^i)\}_{i=1}^n$. Let $\mathbf{y} = (y^1\ y^2\ \ldots\ y^n)^\top$ and let $\widehat{\mathbf{y}} = (y^1\ y^2\ \ldots\ y^n)^\top$ denote the vectors containing actual values and predicted values of the response variable for the n samples in data set D. Consider the OLS objective function to determine $\boldsymbol{\beta}$ values by solving $\min_{\boldsymbol{\beta}} J(\beta) = \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$ where X is a feature matrix whose construction is given in the notebook shared in class.
   (a) **[2 marks]** Using the zero-gradient condition $\nabla_{\boldsymbol{\beta}} J = \mathbf{0}$ discussed in class, use appropriate assumptions to find a suitable matrix A such that $\widehat{\mathbf{y}} = A\mathbf{y}$. State the assumptions you used.
   (b) **[4 marks]** In the matrix A in part 1a, denote the i-th diagonal entry by a_{ii}. Verify if $\sum_i a_{ii} = (d+1)$. Also find suitable p and q such that $p \leq a_{ii} \leq q$.
   (c) In the notebooks shared in class, write codes to compute a_{ii}.
   (d) **[3 marks]** Check if you can represent $a_{ii} = \frac{\partial \hat{y}^i}{\partial y^i}$. Using this relation, explain the meaning of a_{ii}.
   (e) **[2 marks]** Explore the other possible meanings of a_{ii} and explain the importance of a_{ii} based on your investigations.

(f) **[3 marks]** Recall that the residual for i-th sample is given by $e^i = y^i - \hat{y}^i$. In the notebooks shared, write code to compute the standardized residuals $e^i_\star = e^i/\sigma$ where $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(e^i)^2}{n-(d+1)}}$. Explain a possible reason for using $1/(n-(d+1))$ as a scaling factor to compute the standard deviation σ of the residuals.

2. For the following questions, **do not use** any Python package. Write the complete code yourself. You must reuse code provided in the notebooks used for class lectures.

   (a) **[1 mark]** Read the dataset in `data1.txt` into a `pandas` dataframe.

   (b) **[1 mark]** Display the corresponding data description and understand the contents of the data in `data1.txt`.

   (c) **[1 mark]** Display the number of samples and number of attributes.

   (d) **[1 mark]** Replace the column names of data frame with meaningful column names, designed by you using the description in `data1.txt`.

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

   (f) **[1 mark]** 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. If not, clearly indicate that there are no missing values.

   (g) **[2 marks]** Find the regression coefficients for the data in `data1.txt` and plot the regression line. Compute R^2. Explain your observations.

   (h) **[2 marks]** Plot the residual vs fitted values. Explain your observations.

   (i) **[2 marks]** Plot the standardized residual (discussed in question 1) vs fitted values. Compare this plot with the residual vs fitted values plot. Explain your observations.

   (j) **[2 marks]** Compute a_{ii} (discussed in question 1) for every i-th sample in the data set. Find the set I of indices of the samples for which $a_{ii} > (2/n)\sum_i a_{ii}$. Rerun the regression to find the regression coefficients based on those samples whose indices are **not** in I. Using the new coefficients, plot the residual vs fitted values, standardized residual vs fitted values plots and compute R^2. Comment on your observations. Can the samples whose indices are in I be called outliers?

   (k) **[2 marks]** Compute a_{ii} for every i-th sample in the data set. Find the set I of indices of the samples for which $a_{ii} > (3/n)\sum_i a_{ii}$. Rerun the regression to find the regression coefficients based on those samples whose indices are **not** in I. Using the new coefficients, plot the residual vs fitted values, standardized residual vs fitted values plots and compute R^2. Comment on your observations. Can the samples whose indices are in I be called outliers? Compare and contrast the observations in parts 2j and 2k.

   (l) **[2 marks]** In parts 2j, 2k, we have used the condition $a_{ii} > (p/n)\sum_i a_{ii}$ where $p \in \{2, 3\}$. Explain why such a condition might be useful to segregate problematic samples.

3. **[18 marks]** Repeat the analysis in Question 2 for the data provided in file `data2.txt`. Write all your observations clearly.