Final Project: Regularized Logistic Regression for Ozone classification

- **Instructions:** While discussion with classmates are permitted and encouraged, feel free to work on the project in a group or independently (using R or Python). For a beginner R-programmer, a sample R-script is provided in D2L. Please practice the sample code before starting the project.

- This project deals with a ozone prediction in the atmosphere using a machine learning procedure. Consider the **Ozone** dataset available from your D2L account.

  The persistence of highly concentrated ozone levels in the troposphere does harm to humans, animals, and plants. It is therefore vital to detect high levels of ozone early in order to ensure a healthy environment, especially for the elderly, children, and asthmatics. El Paso, which is considered a high ozone affected city in the USA, has a history of very high ozone levels every year. In this project, we will use the data sets of air pollutants and meteorological variables from 2015 to 2019 from the El Paso area to classify the high/low ozone levels. The dataset was collected from the Texas Commission on Environment Quality (TCEQ) ground stations and cleaned for research purposes. The ozone column is the response (dependent) variable that indicates whether a given day had low ozone (0) or high ozone (1). Thus we have a binary classification problem.

  Follow the steps below to conduct your analysis

  1. (**Data Collection**) Bring the data in R. Print the first five lines and last five lines of the data.
  2. (**Data Preparation**)
     (a) Check the variable types, missing values, outlying and possibly wrong records, and other issues.
     (b) Perform some Exploratory Data Analysis (at least three interesting findings). In particular, inspect the frequency distribution of the target variable. Also, explore the associations between target and other attributes.
     (c) Partition the data into three parts, the training data (D1), the validation data (D2), and the test data (D3), with a ratio of 2:1:1.
     (d) Using model.matrix function, create a input feature matrix from train data and delete the first column. Perform the same analysis for D2 and D3 dataset.
  3. (**Model Building**)
     (a) Fit the logistic regression with any regularization (lasso, or, ridge, or Elastic Net) using the training data D1. Explain why you choose that regularization.

(b) Select the best tuning parameter using the validation data D2. Different criteria could be used here for the selection, such as the misclassification rate or the mean square error for the predicted probabilities.

(c) Plot the tuning parameter (λ) vs. any evaluation metric you chose to pick best λ.

(d) Present your final 'best' model fit by pooling D1 and D2 together.

(e) Print the coefficients of the features. Which variables are important predictors?

(f) Find the odd ratio of any three predictors and interpret your analysis.

4. (**Model Evaluation**)

   (a) Apply the final logistic model to the test data D3.

   (b) Present the ROC curve and the area under the curve, i.e., the C-index. Interpret your analysis.

   (c) Find the MSE, miss classification rate, and confusion matrix. Interpret each component of confusion matrix regarding the ozone classification (e.g., interpret sensitivity, specificity, recall, precision, etc.).

- **Timeline:**

  – A final report is due on 11:59 pm Friday, December 03, 2021. All members of a group do not need to individually submit their project. Please submit just one report for the group but ensure that the names of all group members are listed on the report. The report (word/pdf file) should contain the outputs (image, table, etc.), **and interpretation of your analysis**. Please choose any of the following to submit:

    1. R script/Python script and a doc/pdf report, or
    2. A pdf file generated by R-markdown that includes both your analysis and code, or
    3. A notebook file in Jupyter with R/Python that includes both your analysis and code.

– – – – HAPPY CODING – – – –