Universita' degli studi dell'Aquila

Introduction to Statistical Learning

Dr. Alessandro Giovannelli

1 Exercise 1

1.1 Introduction

The data `flightdelays.csv` reports airplane flights in January 2004, flying from the Washington DC area into the NYC area.

The variable of interest (the response) is if a flight has been delayed by more than 15 min or not (coded as 0 for no delay, and 1 for delay). The explanatory variables include three different arrival airports, John F Kennedy (JFK), Newark (EWR), and LaGuardia (LGA). The three different departure airports are Reagan, Dulles, and Baltimore, along with eight carriers, and a categorical variable for 59 different hours of departure (6 am to 10 pm). The other variables include weather conditions (0 = good/1 = bad) and day of the week (1 for Sunday and Monday and 0 for all other days).

It is required to construct the best model for this data. This can be done by selecting 70% of the cases of data set for the fitting (training) data set, wherein the remaining 30% of the cases become the evaluation data set.

If you want to partecipate to the final competition, you should provide in a CSV format the output obtained on the test set `flightdelays_test.csv`. It is important to note that in this case the response is not given since it will be evaluated successively.

1.2 Main Task

Your objective is to build an accurate and reliable classification method for predicting the assigned variable.

You should address the following points:

a. Explore the dataset and provide a description of the essential features of the data. Use (parsimoniously and effectively) graphs and tables.

b. Use the training sample for the selection of a prediction rule. Select the appropriate model and highlight the main estimation results for all the methods considered. You should consider different classifiers, e.g. following the exercise done during the lecture,

c. Decide which is most suitable method, i.e. the one that yields the most accurate predictions of the outcome variable in the validation sample. Discuss the results based on Confusion Matrix and ROC.

2 Exercise 2

Generate a data set with $p = 8$ features, $n = 100$ observations, and an associated quantitative response vector generated according to the model

$$y_i = \sum_{j=1}^{N} \beta_j X_{ij} + \varepsilon_i.$$

It is assumed that the covariance matrix of the X-variables has the following form

$$\Sigma_X = \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \ddots & & \vdots \\ & & \ddots & & \\ \vdots & & \ddots & 1 & \rho \\ \rho & & \dots & \rho & 1 \end{pmatrix}$$

where $\rho \in \{0, 0.5, 0.95\}$. Two design are considered for the exercise:

- Case 1: $b = (1, 1, 1, 1, 1, 1, 1, 1)$;
- Case 2: $b = (1, 1, 1, 1, 0, 0, 0, 0)$.

We generate $X \sim \mathcal{N}(0, \Sigma_X)$ draw of normally distributed random variables with mean zero and variance Σ_X and $\varepsilon \sim \mathcal{N}(0, \sqrt{3})$

Split your dataset into a training set containing 70% of observations and a test set containing 30% of observations and performs 200 simulations Monte Carlo.

1. Perform LASSO, RIDGE, Principal Component and Subset Regression on the training set, and report the MSE on test set associated with each model considereds.
2. Report also the BoxPlot for all the MSE obtained for each simulation.
3. Repeat the exercise in (1)-(3) when $N = 500$ and $N = 1000$.

Comment the results.

Hint: In the first experiment, all predictors are relevant and matter equally; in the second experiment only the first four predictors matter to the outcome.

The deliverable is a written report of maximum 6 pages, including figures and tables. It should be sent to the email address `alessandro.giovannelli@univaq.it` by May 30, 2023 6:00 p.m.