

1. Consider the “Data1” data set. Let  $y$  be the response variable and  $V_1, V_2, V_3, V_4, V_5, V_6$  be the predictors. Please answer the following questions.
  - (a) Use the training set to find the best linear regression models using respectively the best subset selection, forward stepwise selection and backward stepwise selection methods with the  $C_P$  criterion. What are the variables selected by the three methods? Are they the same?
  - (b) Use the test set to compute the test MSEs for the best subset selection models with predictor number =  $1, 2, \dots, 6$ . Plot the test MSEs v.s. the predictor number. (Hint : You can use the function `model.matrix()` to convert the test set into a matrix with a constant column of value 1. Then you can extract the model coefficients and multiply them with the converted matrix form of the test set to compute the test MSE.)
  - (c) Include the interaction terms of the predictors found by the best subset selection in (a), and rebuild the linear regression model using the best subset selection method with  $C_p$  criterion. Find the best fitted regression model and compute its test MSE.
  - (d) Based on the results of (b) and (c) which model (model with or without interaction) would you choose for the future prediction?
2. Consider the “Data2” data set. Let  $y$  be the response variable and  $A, B, V_1, V_2, \dots, V_{30}$  be the predictors. Perform the following preprocessing before answering the questions.
  - (i) Transform the predictors  $A$  and  $B$  as **factor**.
  - (ii) Use `dummy.data.frame()` in dummies package to do one-hot encoding for the predictors  $A$  and  $B$ .

Make sure to use `set.seed(10)` prior to any modeling.

- (a) Fit the lasso regression model using the tuning parameter  $\lambda$  which minimizes the 10-fold cross validation error. What are the training and test mean square errors?
- (b) Fit the ridge regression model using the tuning parameter  $\lambda$  which minimizes the 10-fold cross validation error. What are the training and test mean square errors?
- (c) Fit the principal components regression model with minimal 10-fold cross validation MSE. Remember to standardize the predictors before modelling. What are the training and test mean square errors? (Hint : You can use `MSEP(pcr.fit)$val[1,,]` to find the cv MSE.)
- (d) Based on the results of (a)-(c), which model will you suggest for prediction? Why? And give reasoning for the order of the three test MSEs.

3. Consider the “Data3” data set. Each row in the data set contains the related information of a visit by a visitor in a session (電商平台), which contains the following variables.

- (i) **Administrative, Administrative Duration, Informational, Informational Duration, Product Related and Product Related Duration**: the number of different types of pages (電商平台網頁) visited by a visitor in that session and the total time spent in each of these page categories.
- (ii) **Bounce Rate** : the percentage of a visitor who enter the site from that page and then leave without triggering any other requests to the analytics server during that session.
- (iii) **Exit Rate** : the percentage that were the last in the session.
- (iv) **Page Value** : the average value for a web page that a user visited before completing an e-commerce transaction.
- (v) **Special Day** : the closeness of the site visiting time to a specific special day.
- (vi) **visitor type** : as returning, new visitor or Other.
- (vii) **weekend** : a Boolean value indicating whether the date of the visit is weekend.
- (viii) **Revenue** : a binary response variable (是否消費).

Use the variables in (i)-(vii) as the predictors and the “revenue” in (viii) as the response variable to answer the following questions.

- (a) Convert the categorical predictors into indicator predictors.
- (b) Use the training set to build the LDA classifier function for the “revenue”. Find the training and test classification errors when the threshold (of the probability Revenue =1) is 0.5.
- (c) Use the training set to build the QDA classifier function. Find the training and test classification errors when the threshold is 0.5.
- (d) Find lasso logistic regression model for “Revenue” with the tuning parameter  $\lambda$  minimizing the 5-fold cross validation classification error. What are the recall and precision for the test set when the threshold is 0.5.
- (e) Use the test set to find the sensitivity and specificity for the thresholds  $\{0.01h \mid 1 \leq h \leq 100\}$  for the three models found in (b)-(d), and plot the three ROC curves (sensitivity v.s. specificity) on the same plot.
- (f) Based on the results of (e), which model will you suggest for prediction? Why?