

1. Consider the "Data1" data set. Let y be the response variable and $V_1, V_2, V_3, V_4, \dots, V_{10}$ be the predictors. Please answer the following questions. When you use inherently random methods, make sure to use `random_state = 2`. In addition, please standardize your input before modeling if the model is not scale-invariant.
 - (a) Using `train_test_split` to randomly select three-quarters of the observations as the training set and the remainder as the test set. (3%)
 - (b) Use the training set to find the best linear regression model using the best subset selection, forward stepwise selection and backward stepwise selection methods with the BIC criterion, respectively. What are the variables selected by the three methods? Are they the same? (12%)
 - (c) Fit the lasso regression model on the training set and plot the coefficient estimate for the ten predictors as a function of λ . In addition, find the parameter λ that minimizes the 10-fold cross-validation error. Please scan λ from 5×10^4 to 5×10^{-4} with evenly spaced 100 samples on log scale when doing cross-validation and plotting. (8%)
 - (d) Write down the final fitted model for the model selected by best subset selection and the lasso. Use the training set and test set to compute the training MSEs and test MSEs for these models. (8%)
 - (e) Fit a PLS model on the training set, with M chosen by 10-fold cross-validation. Report the training MSEs and test MSEs obtained, along with the value of M selected by cross-validation. (6%)
2. Consider the "Data2" data set. Let y be the response variable and V_1, V_2, \dots, V_{40} be the predictors. Please answer the following questions and use Area Under the Receiver Operating Characteristic Curve (ROC AUC) (You may find `sklearn.metrics` module and the `predict_proba` method in each classification class is useful for calculating the ROC AUC useful for calculating the ROC AUC) from prediction scores to measure the performance. When you use inherently random methods, make sure to use `random_state = 2`. In addition, please standardize your input before modeling if the model is not-scale invariant.
 - (a) Split the dataset into the training set and test set with equal size and use the training set to build the LDA classifier function. Find the training and test ROC AUC score. (9%)
 - (b) Use the training set to build the QDA classifier function. Find the training and test ROC AUC score. (6%)
 - (c) Find ridge logistic regression model with the tuning parameter λ minimizing the LOOCV classification error. Please scan λ from 5×10^4 to 5×10^{-4} with evenly spaced 100 samples on log scale when doing cross-validation. Find the training and test ROC AUC score of the fitted model. (8%)
 - (d) Find lasso logistic regression model with the tuning parameter λ minimizing the LOOCV classification error. Please scan λ from 5×10^4 to 5×10^{-4} with evenly spaced 100 samples on log scale when doing cross-validation. Find the training and test ROC AUC score of the fitted model. (8%)

- (e) Based on the results of (a)-(d), which model will you suggest for prediction? Why? (3%)
- (f) Base on the model you selected in (e), find the threshold that optimize the f1 score which is defined as $f1 = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$. Compare the f1 score of the original model and the model with adjusted threshold on the test set. (8%)
3. Consider the "Data3" data set. Each row in the data set contains the related information of a house. Let **Price** be the response variable and V_1, V_2, \dots, V_7 be the predictors. We are asked to use linear regression to deal with the problem. Please answer the following questions. When you use inherently random methods, make sure to use `random_state = 2` and be sure to handle the factor variable correctly in your model.
- (a) Compute the cross-correlation between variables and plot the correlation matrix as a heatmap. Describe what you have observed. (5%)
- (b) Try to draw the pair plot for the dataset which allows us to visualize the pairwise correlations between the different features in this dataset in one place. Describe what you have observed. (5%)
- (c) Use all the predictors to build a multiple linear regression model. In addition, try to compute the variance inflation factor (VIF) (Notice that in the `outliers_influence.OLSInfluence` module provides the VIF) for each predictor. Examine whether we have a multicollinearity problem in this dataset or not. Justify your answer. (9%)
- (d) Without knowing the meaning of variables, it is relatively hard to address the multicollinearity problem. Let us turn our attention to obtaining a better fit. Firstly, plot the studentized residual plot and identify the outliers with studentized residuals greater than 3 in absolute value. Secondly, remove these outliers and refit the multiple linear regression model. Does it improve the fitting? (6%)
- (e) Try to add the second-degree (quadratic) and third-degree (cubic) polynomials of **V4** to the model in (c). Does it improve the fitting? (6%)