

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 a methods that are inherently random, make sure to use `random_state = 1`.
  - (a) Split the data into the training set and testing set with equal size using the function `train_test_split`. (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 AIC criterion, respectively. What are the variables selected by the three methods? Are they the same? (13%)
  - (c) Use the training set and test set to compute the training MSEs and test MSEs for the best subset selection models with predictor number  $= 1, 2, \dots, 10$ . Plot the training MSEs v.s. the predictor number and the test MSEs v.s. the predictor number. (6%)
  - (d) Fit the ridge regression model and choose the parameter  $\alpha$  or  $\lambda$  that minimizes the 5-fold cross-validation error. What are the training and test mean square errors? Please scan  $\alpha$  from  $10^3$  to  $10^{-3}$  with evenly spaced 50 samples when doing cross-validation. In addition, please standardize your input before modeling. (9%)
  - (e) Fit the lasso regression model and choose the parameter  $\alpha$  or  $\lambda$  that minimizes the 5-fold cross-validation error. What are the training and test mean square errors? Please scan  $\alpha$  from  $10^3$  to  $10^{-3}$  with evenly spaced 50 samples when doing cross-validation. In addition, please standardize your input before modeling. (9%)
2. Consider the “Data2” data set. Let  $y$  be the response variable and  $V_1, V_2, \dots, V_5$  be the predictors. Use the f1-score which is defined as  $\text{f1} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$  to measure the performance. When you use a methods that are inherently random, make sure to use `random_state = 1`.
  - (a) Use the training set to build the LDA classifier function. Find the training and test f1-score. (6%)
  - (b) Use the training set to build the QDA classifier function. Find the training and test f1-score. (6%)
  - (c) Build the KNN classifier based on the training data set. Perform 5-fold cross validation to select the number of  $K$  (Please scan  $K$  from 1 to 20 with step size 1). Which value of  $K$  attains the largest accuracy ( $\text{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$ )? Find the training and test f1-score for the classifier with the selected number of  $K$ . Please standardize your input before modeling. (9%)
  - (d) Use the training set to build the naive Bayes classifier function (Assumed that each predictor is drawn from Gaussian distribution). Find the training and test f1-score. (6%)
  - (e) Based on the results of (a)-(d), which model will you suggest for prediction? Why? (3%)

3. Consider the “Data3” data set. Each row in the data set contains the related information of a house, which contains the following variables.

- (i) **SqFtTotLiving**: the size of the main house in the unit of square footage.
- (ii) **SqFtLot**: the size of the lot in the unit of square footage.
- (iii) **Bathrooms**: Number of bathrooms per bedroom.
- (iv) **Bedrooms**: Number of Bedrooms per house.
- (v) **PropertyType**: A factor variable that describes the type of the house.
- (vi) **ZipArea**: A factor variable that describes the region where the house belongs according to the Zip code. (Remember to transform this variable to categorical variable via `df['ZipArea'] = df['ZipArea'].astype('category')` before modeling)
- (vii) **YrtBuilt**: The year in which the house was built.
- (viii) **SalePrice**: The estimated sale price of the house which is the response variable.

Use the variables in (i)-(vii) as the predictors and the “SalePrice” in (viii) as the response variable to answer the following questions. Be sure to handle the factor variable correctly in your model.

- (a) Plot the histogram of **YrtBuilt** after the year 1950. Describe what you have observed. (5%)
- (b) Plot the histogram of **SqFtToLiving** for the house that was built after 2000 and before 2010. Compute a kernel density estimate to smooth the distribution and show it on the plot (You can use the function `histplot` or `kdeplot` in the `seaborn` module). Describe what you have observed. (6%)
- (c) Plot the boxplot of **SqFtToLiving** for the years 1950, 1975, 2000 and 2015 (You may find the command `pandas.DataFrame.query` useful). Describe what you have observed. (5%)
- (d) Use all the predictors to build a multiple linear regression model. Examine the coefficient of the predictor **Bedrooms** and explain the meaning of this coefficient. Does the estimate match your intuition and why? (6%)
- (e) It is natural to presume that the relationship between house size and the sale price depends on location. Add the interaction term between **SqFtToLiving** and **ZipArea** to the above model and refit the multiple linear regression. Does this interaction term improve the fitting? (6%)
- (f) Extract the subset whose **ZipArea** is 1 from the original data. Fit a multiple linear regression model for this subset using all the predictors. Plot leverage ( $h_i$ ) vs studentized residuals ( $t_i$ ). In addition, based on the Cook's distance  $= \frac{1}{p} t_i^2 \frac{h_i}{1-h_i}$ , which point is the most influential point ( $p$  is the number of predictors)? (Note that `outliers_influence.OLSInfluence` module provides the Cook's distance). Compare the coefficients between the original model and the refitted model without this data point; what do you observe? (12%)