BIOST 546
WINTER QUARTER 2021

Homework # 3
Due Via Online Submission to Canvas: Monday, Feb 22 at 10 AM

Instructions: You may discuss the homework problems in small groups, but you must write up the final solutions and code yourself. Please turn in your code for the problems that involve coding. However, code without written answers will receive no credit. To receive credit, you must explain your answers and show your work. All plots should be appropriately labeled and legible, with axis labels, legends, etc., as needed.

1. In this exercise, you will generate simulated data, and will use this data to perform the **lasso regression**. Make sure you set a random seed before you begin.

   (a) Use the `rnorm()` function to generate a predictor X of length $n = 30$, and a noise vector ϵ of length $n = 30$.

   (b) Generate a response vector Y of length $n = 30$ according to the model

   $$Y = 3 - 2X + 3 * X^2 + \epsilon.$$

   (c) Fit a lasso model to the data, using $X, X^2, \ldots, X^7$ in order to predict Y.

      i. Make a plot that displays the value of each coefficient, as a function of λ. You can display λ on the x-axis and "Coefficient Value" on the y-axis. Your plot should look something like the right-hand-side of Figure 6.13 of the textbook, but with λ on the x-axis. Make sure that you display each coefficient in a different color, and use a caption or legend that clearly indicates which coefficient is which.

      ii. Use cross-validation to select the tuning parameter. What tuning parameter value do you choose? Make a plot to justify your choice. Your plot could display "Estimated Test Error" on the y-axis and λ on the x-axis (or it could display other quantities of your choice).

      iii. Fit a lasso model to all n observations, using the tuning parameter value selected in the previous sub-problem. Write out the fitted model. Comment on the fitted model.

   (d) Now generate 1000 new observations generated according to 1(a) and 1(b). Apply the final fitted model from 1c(iii) to these new observations. What is the mean squared error?

In the following exercises, you will work on a real dataset. As in HW2, you will perform binary classification on the Breast Cancer Wisconsin (Diagnostic) Data Set in the csv file `wdbc.data`. The dataset describes characteristics of the cell nuclei present in n (sample size) images. Each image has multiple attributes, which are described in detail in `wdbc.names`. This time, however, you will predict the attribute in column 2, which we denote by Y, given the columns $\{3, 4, \ldots, 32\}$, which we denote by $X_1, \ldots, X_{30}$. The variable Y represents the diagnosis (M = malignant, B = benign).

2. Data exploration + **Simple Logistic Regression**

   (a) Describe the data: sample size n, number of predictors p, and number of observations in each class.

   (b) Divide the data into a training set of 400 observations, and a test set; Set the seed with `set.seed(2)` before you sample the training set.

   (c) Normalize your **predictors**, i.e. for each variable X_j remove the mean and make each variable's standard deviation 1. You should perform this step separately in the training set and test set. Why?

   (d) Compute the correlation matrix of your training predictors (command `cor`) and plot it (e.g. command `ggcorrplot` in the library `ggcorrplot`). Inspect the correlation matrix and explain what type of challenges this dataset may present?

   (e) Fit a (simple) logistic regression model to predict Y given $X_1, \ldots, X_{30}$. Inspect and report the correlation between the variables X_1, X_3 and the magnitude of their coefficient estimates $\hat{\beta}_1, \hat{\beta}_3$ wrt to the other coefficients of the model. Comment on their values and relate this to what we have seen in class.

   (f) Use the glm previously fitted and the Bayes rule to compute the predicted outcome $\hat{Y}$ from the associated probability estimates (computed with `predict`) both on the training and the test set. Then compute the confusion table and prediction accuracy (rate of correctly classified observations) both on the training and test set. Comment on the results.

3. **Ridge Logistic Regression**

   (a) From the normalized training set and validation set, construct a data matrix X (`numeric`) and an outcome vector y (`factor`) .

   (b) On the training set, run a ridge logistic regression model to predict Y given $X_1, \ldots, X_{30}$. Use the following grid of values for lambda: `10^seq(5,-18,length=100)`.
   In R: Use the function `glmnet` *as in the regression setting, but with the additional argument* `family = "binomial"`.

   (c) Plot the values of the coefficients β_1, β_3 (y-axis) in function of `log(lambda)` (x-axis). Comment on the result.

(d) Apply 10-fold cross-validation with the previously defined grid of values for lambda. Report the value of lambda that minimizes the CV misclassification error. We will refer to it as the optimal lambda. Plot the misclassification error (y-axis) in function of log(lambda) (x-axis).

In R: Use `cv.glmnet` *as in the regression setting, but with the additional arguments* `family = "binomial"` *and* `type.measure = "class"`. *Also note that 10-fold cross-validation is the default option in* `cv.glmnet`.

(e) Report the number of coefficients β_j that are different from 0 for the ridge model with the optimal lambda. Comment on the results.

(f) Use the regularized glm previously fitted (with the optimal lambda) – and the Bayes rule – to compute the predicted outcome $\hat{Y}$ from the associated probability estimates; both on the training and the test set. Then compute the confusion table and prediction accuracy both on the training and test set. Comment on the results.

In R: Use the command `predict` *as in the regression framework, but with the additional argument* `type = "response"`*, which indicates the you want in output the predicted probabilities (of the tumor being malignant, in this case).*

(g) Plot the ROC curve, computed on the test set, for a dense grid of possible cutoffs (e.g. 20 intervals).

(h) Compute an estimate of the Area under the ROC Curve (AUC).

4. **Lasso Logistic Regression:** Repeat the sub-problems 3(b) to 3(h) using a lasso regression model.

5. **Discuss the performances** of the simple glm, ridge glm, and lasso glm on the Breast Cancer Wisconsin Data Set in terms of prediction accuracy (on the training and test set) and model interpretability.