

ECE 625
Instructor: Di Niu
Assignment 2

Due date: see the course website

- You must use either Word or Latex to write up your assignment. Hand written assignments are not accepted.
- **Include all necessary R outputs in your answers.**
- In addition, email me and the marking TA all your R code in one zip file named ECE625_Assign2_FirstName_LastName.zip. The subject of the email should be “ECE 625 Assignment 2 R code.”
- In the zip file, you need to have an individual file named Assign2_Q#.R for each question that needs R code, e.g., Assign2_Q9.R.

You may lose marks if you do not follow the instructions above or the requested format.

Q1. #-----##

This question should be answered using the **Weekly** data set, which is part of the **ISLR** package. This data is similar in nature to the **Smarket** data from this chapter’s lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

- (a) Produce some numerical and graphical summaries of the **Weekly** data. Do there appear to be any patterns?
- (b) Use the full data set to perform a logistic regression with **Direction** as the response and the five lag variables plus **Volume** as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?
- (c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

- (d) Now fit the logistic regression model using a training data period from 1990 to 2008, with `Lag2` as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).
- (e) Repeat (d) using LDA.
- (f) Repeat (d) using QDA.
- (g) Which of these methods appears to provide the best results on this data?
- (h) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data.

Q2.

In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the `Auto` data set.

- (a) Create a binary variable, `mpg01`, that contains a 1 if `mpg` contains a value above its median, and a 0 if `mpg` contains a value below its median. You can compute the median using the `median()` function. Note you may find it helpful to use the `data.frame()` function to create a single data set containing both `mpg01` and the other `Auto` variables.
- (b) Explore the data graphically in order to investigate the association between `mpg01` and the other features. Which of the other features seem most likely to be useful in predicting `mpg01`? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.
- (c) Split the data into a training set and a test set.
- (d) Perform LDA on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained?
- (e) Perform QDA on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained?
- (f) Perform logistic regression on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained?

Q3.

Here we explore the maximal margin classifier on a toy data set.

- (a) We are given $n = 7$ observations in $p = 2$ dimensions. For each observation, there is an associated class label.

Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

Sketch the observations.

- (b) Sketch the optimal separating hyperplane, and provide the equation for this hyperplane
- (c) Describe the classification rule for the maximal margin classifier. It should be something along the lines of “Classify to Red if $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$, and classify to Blue otherwise.” Provide the values for β_0 , β_1 , and β_2 .
- (d) On your sketch, indicate the margin for the maximal margin hyperplane.
- (e) Indicate the support vectors for the maximal margin classifier.
- (f) Argue that a slight movement of the seventh observation would not affect the maximal margin hyperplane.
- (g) Sketch a hyperplane that is *not* the optimal separating hyperplane, and provide the equation for this hyperplane.
- (h) Draw an additional observation on the plot so that the two classes are no longer separable by a hyperplane.

Q4. #-----##

This problem involves the **OJ** data set which is part of the **ISLR** package.

- (a) Create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.
- (b) Fit a support vector classifier to the training data using `cost=0.01`, with `Purchase` as the response and the other variables as predictors. Use the `summary()` function to produce summary statistics, and describe the results obtained.
- (c) What are the training and test error rates?
- (d) Use the `tune()` function to select an optimal `cost`. Consider values in the range 0.01 to 10.
- (e) Compute the training and test error rates using this new value for `cost`.
- (f) Repeat parts (b) through (e) using a support vector machine with a radial kernel. Use the default value for `gamma`.
- (g) Repeat parts (b) through (e) using a support vector machine with a polynomial kernel. Set `degree=2`.
- (h) Overall, which approach seems to give the best results on this data?

Q5. #-----##

It was mentioned in the chapter that a cubic regression spline with one knot at ξ can be obtained using a basis of the form $x, x^2, x^3, (x - \xi)_+^3$, where $(x - \xi)_+^3 = (x - \xi)^3$ if $x > \xi$ and equals 0 otherwise. We will now show that a function of the form

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - \xi)_+^3$$

is indeed a cubic regression spline, regardless of the values of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$.

- (a) Find a cubic polynomial

$$f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3$$

such that $f(x) = f_1(x)$ for all $x \leq \xi$. Express a_1, b_1, c_1, d_1 in terms of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$.

- (b) Find a cubic polynomial

$$f_2(x) = a_2 + b_2x + c_2x^2 + d_2x^3$$

such that $f(x) = f_2(x)$ for all $x > \xi$. Express a_2, b_2, c_2, d_2 in terms of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$. We have now established that $f(x)$ is a piecewise polynomial.

- (c) Show that $f_1(\xi) = f_2(\xi)$. That is, $f(x)$ is continuous at ξ .
- (d) Show that $f'_1(\xi) = f'_2(\xi)$. That is, $f'(x)$ is continuous at ξ .
- (e) Show that $f''_1(\xi) = f''_2(\xi)$. That is, $f''(x)$ is continuous at ξ .

Therefore, $f(x)$ is indeed a cubic spline.

Hint: Parts (d) and (e) of this problem require knowledge of single-variable calculus. As a reminder, given a cubic polynomial

$$f_1(x) = a_1 + b_1x + c_1x^2 + d_1x^3,$$

the first derivative takes the form

$$f'_1(x) = b_1 + 2c_1x + 3d_1x^2$$

and the second derivative takes the form

$$f''_1(x) = 2c_1 + 6d_1x.$$

Q6. #-----##

This question uses the variables **dis** (the weighted mean of distances to five Boston employment centers) and **nox** (nitrogen oxides concentration in parts per 10 million) from the **Boston** data. We will treat **dis** as the predictor and **nox** as the response.

- (a) Use the **poly()** function to fit a cubic polynomial regression to predict **nox** using **dis**. Report the regression output, and plot the resulting data and polynomial fits.
- (b) Plot the polynomial fits for a range of different polynomial degrees (say, from 1 to 10), and report the associated residual sum of squares.
- (c) Perform cross-validation or another approach to select the optimal degree for the polynomial, and explain your results.

- (d) Use the `bs()` function to fit a regression spline to predict `nox` using `dis`. Report the output for the fit using four degrees of freedom. How did you choose the knots? Plot the resulting fit.
- (e) Now fit a regression spline for a range of degrees of freedom, and plot the resulting fits and report the resulting RSS. Describe the results obtained.
- (f) Perform cross-validation or another approach in order to select the best degrees of freedom for a regression spline on this data. Describe your results.

Q7. #-----##

This question uses the variables `dis` (the weighted mean of distances to five Boston employment centers) and `nox` (nitrogen oxides concentration in parts per 10 million) from the `Boston` data. We will treat `dis` as the predictor and `nox` as the response.

- (a) Plot the kernel smoothing fits using `the box kernel` for a range of bandwidths (say, from 0.1 to 5), and report the associated residual sum of squares.
- (b) Perform cross-validation or another approach to select the optimal bandwidth for the polynomial, and explain your results.
- (c) Plot the kernel smoothing fits using `the Gaussian kernel` for a range of bandwidths (say, from 0.11 to 5), and report the associated residual sum of squares.
- (d) Perform cross-validation or another approach to select the optimal bandwidth for the polynomial, and explain your results.
- (e) Plot the `loess` fits for a range of spans (say, from 0.1 to 5), and report the associated residual sum of squares.
- (f) Perform cross-validation or another approach to select the optimal span for the polynomial, and explain your results.

Q8. #-----##

We now review *k*-fold cross-validation.

- (a) Explain how *k*-fold cross-validation is implemented.
- (b) What are the advantages and disadvantages of *k*-fold cross-validation relative to:
 - i. The validation set approach?
 - ii. LOOCV?

Q9. #-----##

We will now perform cross-validation on a simulated data set.

- (a) Generate a simulated data set as follows:

```
> set.seed(1)
> y=rnorm(100)
> x=rnorm(100)
> y=x-2*x^2+rnorm(100)
```

In this data set, what is n and what is p ? Write out the model used to generate the data in equation form.

- (b) Create a scatterplot of X against Y . Comment on what you find.
(c) Set a random seed, and then compute the LOOCV errors that result from fitting the following four models using least squares:

- i. $Y = \beta_0 + \beta_1 X + \epsilon$
- ii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$
- iii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$
- iv. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$.

Note you may find it helpful to use the `data.frame()` function to create a single data set containing both X and Y .

- (d) Repeat (c) using another random seed, and report your results. Are your results the same as what you got in (c)? Why?
(e) Which of the models in (c) had the smallest LOOCV error? Is this what you expected? Explain your answer.
(f) Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (c) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?

Q10. #-----##

We will now consider the `Boston` housing data set, from the `MASS` library.

- (a) Based on this data set, provide an estimate for the population mean of `medv`. Call this estimate $\hat{\mu}$.
(b) Provide an estimate of the standard error of $\hat{\mu}$. Interpret this result.

Hint: We can compute the standard error of the sample mean by dividing the sample standard deviation by the square root of the number of observations.

- (c) Now estimate the standard error of $\hat{\mu}$ using the bootstrap. How does this compare to your answer from (b)?
 - (d) Based on your bootstrap estimate from (c), provide a 95 % confidence interval for the mean of `medv`. Compare it to the results obtained using `t.test(Boston$medv)`.
- Hint: You can approximate a 95 % confidence interval using the formula $[\hat{\mu} - 2SE(\hat{\mu}), \hat{\mu} + 2SE(\hat{\mu})]$.*
- (e) Based on this data set, provide an estimate, $\hat{\mu}_{med}$, for the median value of `medv` in the population.
 - (f) We now would like to estimate the standard error of $\hat{\mu}_{med}$. Unfortunately, there is no simple formula for computing the standard error of the median. Instead, estimate the standard error of the median using the bootstrap. Comment on your findings.
 - (g) Based on this data set, provide an estimate for the tenth percentile of `medv` in Boston suburbs. Call this quantity $\hat{\mu}_{0.1}$. (You can use the `quantile()` function.)
 - (h) Use the bootstrap to estimate the standard error of $\hat{\mu}_{0.1}$. Comment on your findings.