Exercises

September 13, 2017

1 Conceptual

1. Carefully explain the differences between the KNN classifier and KNN regression methods.

2. Suppose we have a data set with five predictors, $X_1 = \text{GPA}$, $X_2 = \text{IQ}$, $X_3 = \text{Gender}$ (1 for Female and 0 for Male), $X_4 = \text{Interaction between GPA and IQ}$, and $X_5 = \text{Interaction between GPA and Gender}$. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10$.

   (a) Which answer is correct, and why?

      i. For a fixed value of IQ and GPA, males earn more on average than females.

      ii. For a fixed value of IQ and GPA, females earn more on average than males.

      iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

      iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

   (b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

   (c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

3. Suppose we collect data for a group of students in a statistics class with variables $X_1 = \text{hours studied}$, $X_2 = \text{undergrad GPA}$, and $Y = \text{receive an A}$. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6, \hat{\beta}_1 = 0.05, \hat{\beta}_2 = 1$.

   (a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

   (b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

4. This problem has to do with odds.

   (a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?

   (b) Suppose that an individual has a 16% chance of defaulting on her credit card payment. What are the odds that she will default?

5. We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain $p + 1$ models, containing $0, 1, 2, ..., p$ predictors. Explain your answers:

   (a) Which of the three models with k predictors has the smallest training RSS?

   (b) Which of the three models with k predictors has the smallest test RSS?

   (c) True or False:

      i. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k+1)-variable model identified by forward stepwise selection.

      ii. The predictors in the k-variable model identified by back- ward stepwise are a subset of the predictors in the (k + 1)- variable model identified by backward stepwise selection.

      iii. The predictors in the k-variable model identified by back- ward stepwise are a subset of the predictors in the (k + 1)- variable model identified by forward stepwise selection.

      iv. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k+1)-variable model identified by backward stepwise selection.

      v. The predictors in the k-variable model identified by best subset are a subset of the predictors in the (k + 1)-variable model identified by best subset selection.

6. Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}) \text{ subject to } \sum_{j=1}^{p}|\beta_j| \leq s$$

   for a particular value of s. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

   (a) As we increase s from 0, the training RSS will:

      i. Increase initially, and then eventually start decreasing in an inverted U shape.

ii. Decrease initially, and then eventually start increasing in a U shape.
iii. Steadily increase.
iv. Steadily decrease.
v. Remain constant.

(b) Repeat (a) for test RSS. (c) Repeat (a) for variance.

(c) Repeat (a) for (squared) bias.

(d) Repeat (a) for the irreducible error.

7. Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n})y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}) + \lambda \sum_{j=1}^{p} \beta_j^2$$

for a particular value of λ. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

(a) As we increase λ from 0, the training RSS will:
i. Increase initially, and then eventually start decreasing in an inverted U shape.
ii. Decrease initially, and then eventually start increasing in a U shape.
iii. Steadily increase.
iv. Steadily decrease.
v. Remain constant.

(b) Repeat (a) for test RSS. (c) Repeat (a) for variance.

(c) Repeat (a) for (squared) bias.

(d) Repeat (a) for the irreducible error.

8. It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas the lasso may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting. Suppose that $n = 2, p = 2, x_{11} = x_{12}, x_{21} = x_{22}$. Furthermore, suppose that $y_1 + y_2 = 0$ and $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$, so that the estimate for the intercept in a least squares, ridge regression, or lasso model is zero: $\hat{\beta}_0 = 0$.

(a) Write out the ridge regression optimization problem in this set- ting.

(b) Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$.

(c) Write out the lasso optimization problem in this setting.

(d) Argue that in this setting, the lasso coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are not unique?in other words, there are many possible solutions to the optimization problem in (c). Describe these solutions.

9. This question relates to the plots in Figure 8.12.

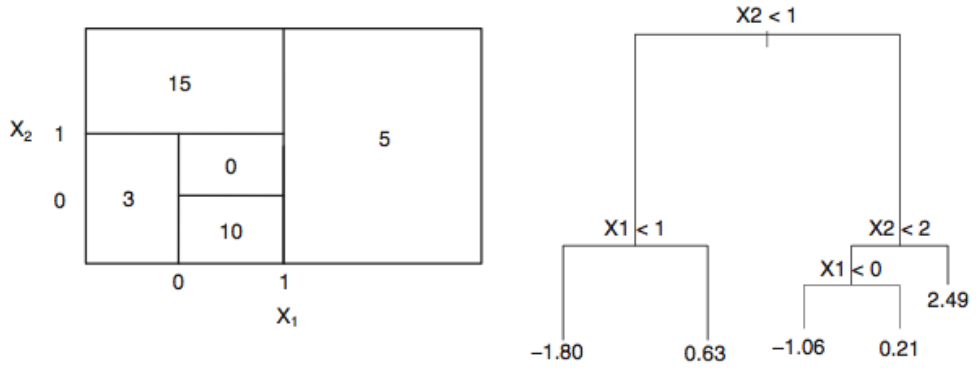

FIGURE 8.12. Left: *A partition of the predictor space corresponding to Exercise 4a.* Right: *A tree corresponding to Exercise 4b.*

(a) Sketch the tree corresponding to the partition of the predictor space illustrated in the left-hand panel of Figure 8.12. The num- bers inside the boxes indicate the mean of Y within each region.

(b) Create a diagram similar to the left-hand panel of Figure 8.12, using the tree illustrated in the right-hand panel of the same figure. You should divide up the predictor space into the correct regions, and indicate the mean for each region.

10. Suppose we produce ten bootstrapped samples from a data set containing red and green classes. We then apply a classification tree to each bootstrapped sample and, for a specific value of X, produce 10 estimates of $P(ClassisRed|X)$:

$$0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, and 0.75.$$

There are two common ways to combine these results together into a single class prediction. One is the majority vote approach discussed in this chapter. The second approach is to classify based on the average probability. In this example, what is the final classification under each of these two approaches?

11. We have seen that in $p = 2$ dimensions, a linear decision boundary takes the form $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$. We now investigate a non-linear decision boundary.

(a) Sketch the curve

$$(1 + X_1)^2 + (2 - X_2)^2 = 4$$

(b) On your sketch, indicate the set of points for which

$$(1 + X_1)^2 + (2?X_2)^2 > 4,$$

as well as the set of points for which

$$(1 + X_1)^2 + (2?X_2)^2 \leq 4.$$

(c) Suppose that a classifier assigns an observation to the blue class if

$$(1 + X_1)^2 + (2?X_2)^2 > 4,$$

and to the red class otherwise. To what class is the observation $(0, 0)$ classified? $(?1, 1)?(2, 2)?(3, 8)?$

(d) Argue that while the decision boundary in (c) is not linear in terms of X1 and X2, it is linear in terms of X1, X12, X2, and X2 .

12. 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. Sketch the observations.

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

(b) Sketch the optimal separating hyperplane, and provide the equa- tion for this hyperplane (of the form (9.1)).

(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 hyper- plane, 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.

13. Suppose that we have four observations, for which we compute a dissimilarity matrix, given by For instance, the dissimilarity between the first and second obser- vations

$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}$$

is 0.3, and the dissimilarity between the second and fourth observations is 0.8.

(a) On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observa- tions using complete linkage. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.

(b) Repeat (a), this time using single linkage clustering.

(c) Suppose that we cut the dendogram obtained in (a) such that two clusters result. Which observations are in each cluster?

(d) Suppose that we cut the dendogram obtained in (b) such that two clusters result. Which observations are in each cluster?

(e) It is mentioned in the chapter that at each fusion in the den- drogram, the position of the two clusters being fused can be swapped without changing the meaning of the dendrogram. Draw a dendrogram that is equivalent to the dendrogram in (a), for which two or more of the leaves are repositioned, but for which the meaning of the dendrogram is the same.

14. In this problem, you will perform K-means clustering manually, with $K = 2$, on a small example with $n = 6$ observations and $p = 2$ features. The observations are as follows.

(a) Plot the observations.

Obs.	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

(b) Randomly assign a cluster label to each observation. You can use the sample() command in R to do this. Report the cluster labels for each observation.

(c) Compute the centroid for each cluster.

(d) Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.

(e) Repeat (c) and (d) until the answers obtained stop changing.

(f) In your plot from (a), color the observations according to the cluster labels obtained.

2 Applied

1. Auto 데이터를 이용하여 다음을 코딩하라.(결측값은 제거할 것)

   - lm() 함수를 이용해 mpg를 반응 변수, horsepower를 예측 변수로 하는 단순 선형 회귀 분석을 수행하고 summary() 함수를 이용해 결과를 나타내라. 또한, horsepower 값이 98일 때, mpg의 95% 신뢰구간과 예측구간을 각각 구하라.
   - 예측변수와 반응변수를 축으로 하는 그래프를 그려라. 그리고 abline() 함수를 이용해 회귀직선을 나타내라.
   - plot() 함수를 이용해 최소 제곱 회귀 분석의 적합성을 판단하는 그래프들을 나타내라.

2. Auto 데이터를 이용하여 다음을 코딩하라.(결측값은 제거할 것)

   - lm() 함수을 이용해 mpg를 반응변수로 하는 다중 선형 회귀 분석을 수행하라. summary() 함수를 이용해 결과를 나타내라. 이 때, 예측 변수는 name 변수를 제외한 다른 모든 변수로 지정한다.
   - plot() 함수를 이용해 선형 회귀 분석의 적합성을 판단하는 그래프들을 나타내라.
   - cylinders와 displacement의 교호작용과 displacement와 weight 의 교호 작용을 동시에 포함하는 선형 회귀 분석을 수행하라.(다른 변수들은 무시할 것)
   - log(X), $\sqrt{æX}$, X^2 등으로 변환한 변수들을 이용해 회귀 분석을 수행해보라.

3. Carseats 데이터를 이용하여 다음을 코딩하라.

   - Sales를 반응변수로 하는 다중 선형 회귀 분석을 수행하라. summary() 함수를 이용해 결과를 나타내라. 이 때, 예측 변수는 Price, Urban 과 US 변수로 지정한다.
   - 첫번째 문제에 근거하여, 유의한 변수들만 예측 변수로 지정하여 다중 선형 회귀 분석을 수행하라. 그리고 첫번째 문제의 모형과 두번째 문제의 모형 중 어느 것이 더 적합한가?
   - Price와 US의 교호작용을 포함하는 다중 선형 회귀 분석을 수행하라.
   - 두번째 문제의 모형에서 각 회귀계수의 95%신뢰구간을 구해 보고, plot() 함수를 이용해 선형 회귀 분석의 적합성을 판단하는 그래프들을 나타내라.

4. cement 데이터를 이용하여 다음을 코딩하라. (게시판 cement.txt 이용)

   - 예측 변수들 간의 상관계수 행렬을 구해보고 패키지(corrplot) 를 이용하여 예측 변수들 간의 관계를 그려라. (corrplot() 함수를 이용)
   - vif() 함수를 이용하여 예측 변수들간의 다중공선성이 존재하는지를 판단해라.
   - 반응변수를y로하는다중선형회귀분석을수행하고 개별적 계수(β_i)에 대한 유의성 t 검정과 다중 회귀 모형에 대한 유의성 F검정을 비교하여라.
   - 예측변수x3과x4를제거하여다중선형회귀분석을 수행하라. 그리고 vif() 함수를 이용하여 다중공선성이 존재하는지를 판단해라.

5. mpg 자료를 이용하여 다음을 코딩하라.(결측값은 제거할 것). (조건: 반응 변수:mpg01, 예측 변수: cylinders, weight, displacement, horsepower로 지정)

   - mpg의 값이 중앙값보다 크면 1로, 아니면 0으로 할당하는 이항 변수 자료를 만들어라. 그리고 Auto 자료에 mpg01 이라는 변수명으로 추가하라.
   - 선형 판별 분석(LDA)을 수행하라. 이때, 자료를 짝수년을 기준으로 훈련 자료와 시험자료로 분리하고 전체 오차율을 구하여라.
   - 이차 판별 분석(QDA)을 수행하라. 그리고 혼동 행렬 (confusion matrix)을 만들고 전체 오차율을 구하여라.
   - 다중 로지스틱 회귀 분석을 수행하라. 그리고 시험자료에서 예측한 값과 실제 값을 비교하고 전체 오차율을 구하여라.
   - 최근접이웃방법(KNN)을 수행하라. 그리고 $K = 1, K = 10, K = 100$일 때, 각 전체 오차율을 구하여라. .

6. Default(ISLR 패키지) 자료를 이용하여 다음을 코딩하라. 단, 반응변수는 default(파산 여부)이다.

   - summary()와 glm()함수를 이용하여, income과 balance를 예측변수로 하는 로지스틱 모형을 적합하고 회귀 계수의 표준오차를 알아내라.

- Default 데이터와 인덱스를 인풋으로 하고 income과 balance의 로지스틱 회귀계수를 아웃풋으로 하는 함수 boot.fn()을 만들어라.
- boot()함수와 위에서 만든 boot.fn()을 이용해 income과 balance의 로지스틱 회귀계수의 표준오차 추정치를 구하라. 결과를 첫번째 문제에서 구한 값과 비교해보자.

7. 다음의 코드로 시뮬레이션 자료를 생성한다.

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

- 다음 모델들에 대해 하나남기기 교차검증 시험오차들을 계산하라.

  (a) $Y = \beta_0 + \beta_1 X + \epsilon$

  (b) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$

  (c) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

  (d) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$

  hint : data.frame(x,y)를 이용해 두 변수를 Dataset으로 묶자.
- 어떤 모델이 가장 작은 하나남기기 교차검증 시험오차를 갖는가? 이 결과는 glm()이 도출하는 회귀계수의 유의성과 같은 결론을 나타내는가?

8. data.txt의 자료를 이용하여 다음을 코딩하라. ($y_i = 3 + 2x_i - 3x_i^2 + 0.3x_i^3 + \epsilon_i$ 모형에서 산출된 자료이다. ϵ_i $N(0,1), x_i$ $N(0,1), i = 1, ..., n$)

- 예측변수들을 $X, X^2, ..., X^10$ 까지 포함하는 모형을 통해 최적부분집합선택을 수행하라. C_p, BIC, adjusted R^2에 의하면 가장 좋은 모형은 무엇인가? (Hint : regsubsets() 함수 이용)
- 동일한 문제를 전진선택법과 후진선택법을 이용했을 때, 가장 좋은 모형은 무엇인가?
- $X, X^2, ..., X^10$ 예측 변수들을 이용하여 라쏘(lasso)모형에 적합하라. 그리고 교차검증 방법을 이용하여 최적 λ를 구하고, λ에 관한 교차 검증 오차 그래프를 그려라.

9. 패키지 MASS의 Boston 데이터를 이용하라. 변수 dis는 보스톤고용센터와의 거리를, 변수 nox는 질소산화물(nitrogen oxides)농도를 나타낸다.

- dis를 예측변수, nox를 반응변수로 하는 3차 다항회귀분석을 시행하고 결과를 나타내라. 또한 그래프에 데이터와 회귀선을 표시하라.

- 같은 변수로 회귀 스플라인을 적합하라. 자유도 df는 4로 지정한다. 매듭은 어떤 값으로 결정되었는가?
- 비소매업(non-retail) 산업의 비율을 나타내는 indus를 예측변수로 추가하여 일반화가법모형을 적합하라. indus는 자유도 4의 자연 스플라인기저로, dis는 자유도 4의 평활스플라인기저로 구성하자.

10. Carseat.txt 자료를 이용하라. 반응변수는 Sales이다.

    - 데이터를 훈련자료와 시험자료로 나누고 회귀나무모형을 적합하라. 나무를 그리고 시험오차를 구하라.
    - 교차검증법을 이용해 나무의 최적 복잡도를 결정하라.(cv.tree 의 옵션 FUN은 prune.tree로 준다.) 가지치기가 시험오차를 향상시키는가?
    - 배깅으로 분석하라. 시험오차는 얼마인가? importance() 함수를 이용해 어떤 변수들이 가장 중요한지 알아보자.
    - 랜덤숲으로 분석하라.시험오차는 얼마인가? importance() 함수를 이용해 어떤 변수들이 가장 중요한지 알아보자.

11. Caravan 데이터를 이용하라.

    - Purchase의 데이터값을 Yes는 1로, No는 0으로 바꾼 후, 처음 1000개의 관측치로 이루어진 훈련자료와 나머지로 이루어진 시험자료를 만들어라.
    - Purchase를 반응변수로 하여 훈련자료에 부스팅 모형을 적합하라.(Hint:반응변수가 이항 변수(1 or 0)인 경우 distribution 옵션을 "bernoulli"로 한다.) 이 때, 1000개의 나무와 조율 파라미터 0.01을 이용하자. 어떤 설명변수가 가장 영향력이 큰가?
    - 적합한 부스팅 모형으로 시험자료를 예측하라. 단, 구입(Yes) 할 확률이 20%이상이면 구입을 하는 것으로 예측한다. 혼동행렬을 구하라. 구입할 것으로 예측한 사람 중 실제 구매한 사람의 비율은 얼마인가? 로지스틱 회귀모형의 결과와 비교해보자.

12. OJ.txt의 자료를 이용하여 다음을 코딩하여라.

    - 훈련 자료를 800개를 만들고 나머지는 시험 자료로 만들어라.
    - 반응 변수를 Purchase, 나머지 변수들을 예측 변수로 하여 cost=0.01인 지지벡터분류기를 적합하고 훈련 오차율과 시험 오차율을 구하여라.(단, 훈련 자료 이용)
    - tune()함수를 이용하여 최적 cost를 구하고, 구한 값을 적용하여 훈련 오차율과 시험 오차율을 구하여라. (cost의 범위는 0.01부터 10까지이다.)
    - 위의 두 문제를 방사 커널을 이용한 지지벡터기계에 적용하여 구하여라.(단, 감마는 모두 기본값으로, 첫번째 문제의 cost도 기본값으로 한다)

13. Ch10Ex11.csv 자료는 40개의 조직샘플에 대한 1000개 유전자의 관측치다. 처음의 20개는 건강한 그룹으로부터, 나머지 20개는 질병 그룹으로부터 얻었다.

- 이 데이터는 각 행이 특정 유전자변수(1000)를, 각 열이 개체 (40)를 나타낸다. 즉, 일반적인 자료구조와 다르다. 자료를 불러들이고 적절하게 변형하라.
- K 평균 군집분석을 수행하여 개체들을 2 그룹으로 나누어라. 건강한 그룹과 질병 그룹으로 잘 나누어 지는가?
- 계층군집분석을 각 연결법(complete, average, single)에 대해 수행하고 덴도그램을 나타내라. 단, 상관계수 기반의 거리 (correlation-based distance)를 이용한다.

14. MASS 패키지의 Boston 자료를 이용하라.

    - Boston 자료의 주성분을 분석하라. 첫번째 주성분과 두번째 주성분을 나타내는 그림(쌍도)를 그려보자.
    - 각 주성분의 변동성으로 설명되는 비율을 계산하여 그래프 (스크리 그림)로 나타내라. 누적비율에 대해서도 그래프를 그려보자.
    - 반응변수를 crim(범죄율)로 하여 주성분회귀분석을 수행하라. 모형에 포함되는 주성분 수가 증가함에 따라 평균제곱오차가 어떻게 변하는지 그래프로 나타내라.

15. personality.csv 자료를 이용하라.

    - 자료를 표준화하고 데이터 프레임 형태로 바꾸어라. 그리고 corrplot 패키지를 이용하여 상관계수 그래프를 만들고 자료를 분석하여라.
    - 회전 방법을 사용하지 않고, 인자를 10개로 하는 인자 분석을 수행하라. 전체 분산에 대한 누적비율을 얼마인가?
    - 또한, 공통성(communality)을 구해보고, 인자 적재값들을 이용하여 인자 1과 인자 2에 대한 그래프를 그려라.
    - 회전 방법이 varimax이고 인자를 10개로 하는 인자 분석을 수행하라. 인자 1과 인자 2에 대한 그래프도 그려보고 회전방법을 사용하지 않은 인자 분석과 비교해보자.