Test 1 sample

DSO 530: Applied Modern Statistical Learning Methods

2020

This is a test1 from past years. You can take a look at the format. But due to changes in contents, you are not expected to solve all the problems in the test. For example, we swtiched from R to Python. For another example, question 8 about outliers is not covered in the current year.

You have 90 minutes to do the problems. For multiple choice questions (1-16), make sure to read the questions very carefully and circle the best answer. If you circle multiple answers for a question, you will receive zero for that question. For short answer questions (17-20), write concisely and clearly. This test is open notes, but no electronic equipment other than a calculator is allowed. Questions 16, 17, 18, 19 and 20 are worth 2 points each, and the rest are worth 1 point each. The total points are 25.

Name (Last, First):

USC Student ID number:

Sign on the line below to pledge that you did not give or receive assistance on the questions for this exam from anyone else.

---

part a) multiple choices

1. Let A and B be two events. Suppose we know $P(A \text{ and } B) = 0.5$, $P(A) = 0.6$ and $P(B) = 0.6$. Are the events A and B independent?

- A) Yes
- B) No
- C) With 50% of the chance, yes; with 50% of the chance, no
- D) It cannot be determined without additional information

answer: B). Events A and B are independent if $P(A \text{ and } B) = P(A) \cdot P(B)$.

2. Which of the following is/are true about correlation r?

i) It measures all dependent relationships between two numerical variables

ii) It does not have a unit

iii) If $r(X, Y) = 1$, then regress Y on X, we will get slope 1 in the least squares regression line

iv) r is not a robust measure

- A) i), ii), iii), iv).

- B) ii) and iii)
- C) i), ii) and iv).
- D) ii) and iv).
- E) iii) and iv).

answer: D). i) is not correct because r only measure the linear dependence iii). it is not correct because the condition just implies all points on a positively sloeped line.

Questions $3-9$ are based a `Boston` dataset we have seen in class.

```
library(MASS); attach(Boston); summary(Boston); dim(Boston)
```

```
##       crim                zn              indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm              age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20  
##  Mean   :12.65   Mean   :22.53  
##  3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :37.97   Max.   :50.00
```

```
## [1] 506  14
```

3. Based on the above summary, is the variable `crim` a good predictor for `medv`?

- A) Yes.
- B) No.
- C) We need additional information to determine the answer, and the meaning of "good" is too vague.

answer: C)

4. What is `the training sample size − the number of predictors` (− is the minus sign) for this `Boston` data if we were to use all other variables as predictors to predict the median house price `medv`, and the whole dataset as the training set?

- A) 506
- B) 1,000
- C) 14
- D) 492
- E) 493

answer: E). In this example $n = 506$ and $p = 13$, so $n - p = 493$.

5.

```
lm.fit1 = lm(medv ~ rm); summary(lm.fit1)
```

```
##
## Call:
## lm(formula = medv ~ rm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.346  -2.547   0.090   2.986  39.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.671      2.650  -13.08   <2e-16 ***
## rm             9.102      0.419   21.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.616 on 504 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4825
## F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

What percentage of the variation in the response is NOT explained by the predictor `rm`?

- A) 48.35%
- B) 51.65%
- C) 48.25%
- D) 51.75%

answer: B). Note the `NOT`. So we should look at $1 - R^2 = 1 - 0.4835 = 0.5165$.

6. The variable `chas` is a dummy variable that takes values 0 and 1. If we change all 0 values to -1, how does this re-coding affect prediction using `chas` as a predictor?

- A) It does not have any effect.
- B) We should not do such re-coding at all, because this totally messes up the dataset.

answer: A). All the regression coefficient for `chars` will change, this re-coding should not change prediction at all.

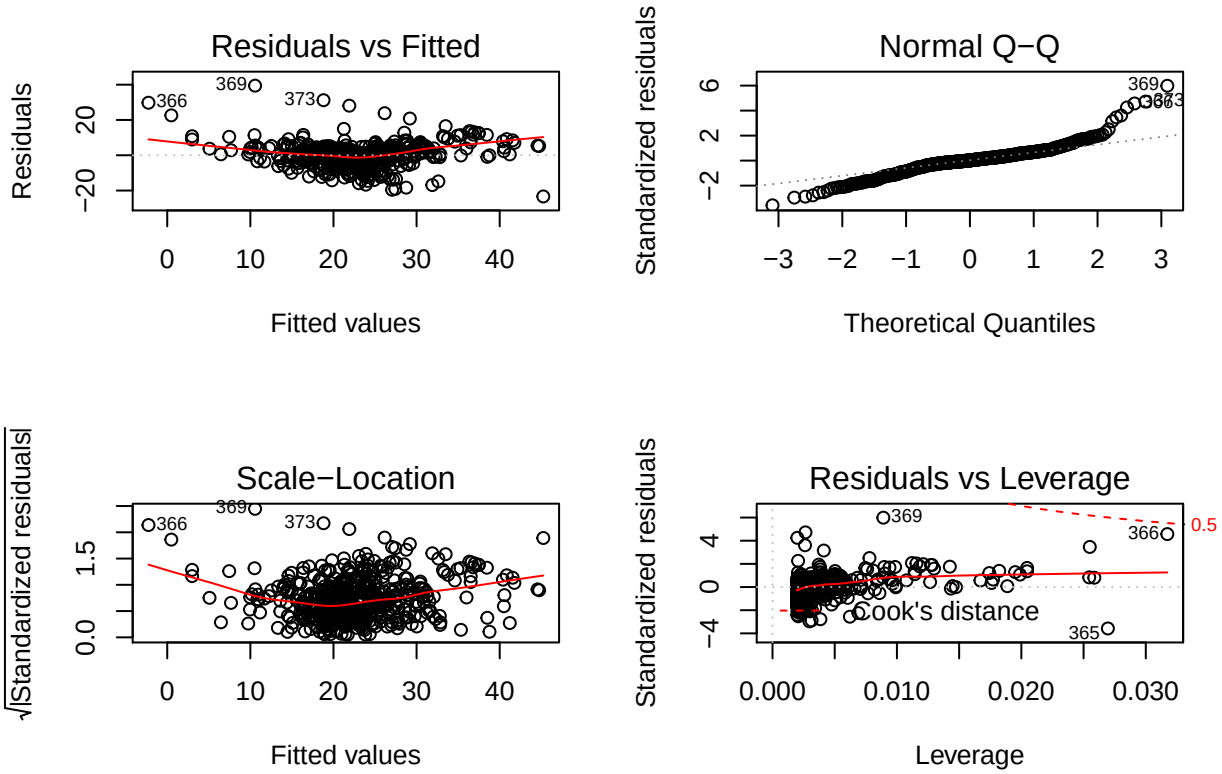
7. Suppose we want to randomly choose 200 rows in dataframe `Boston` as training data, which of the following `R` code gives you this training dataset

- A) Boston[1:200,]
- B) Boston[,1:200]
- C) Boston[sample(1:506, 200),]
- D) Boston[,sample(1:506, 200)]

answer: C). Some people use A), this is not correct because that means selecting the first 200 rows.

8.

```
par(mfrow=c(2,2))
plot(lm.fit1)
```



The above plots indicate

- A) there are possible outliers
- B) there are no outliers at all
- C) The above plots have nothing to do with detecting outliers. We need other kinds of plots for this purpose.

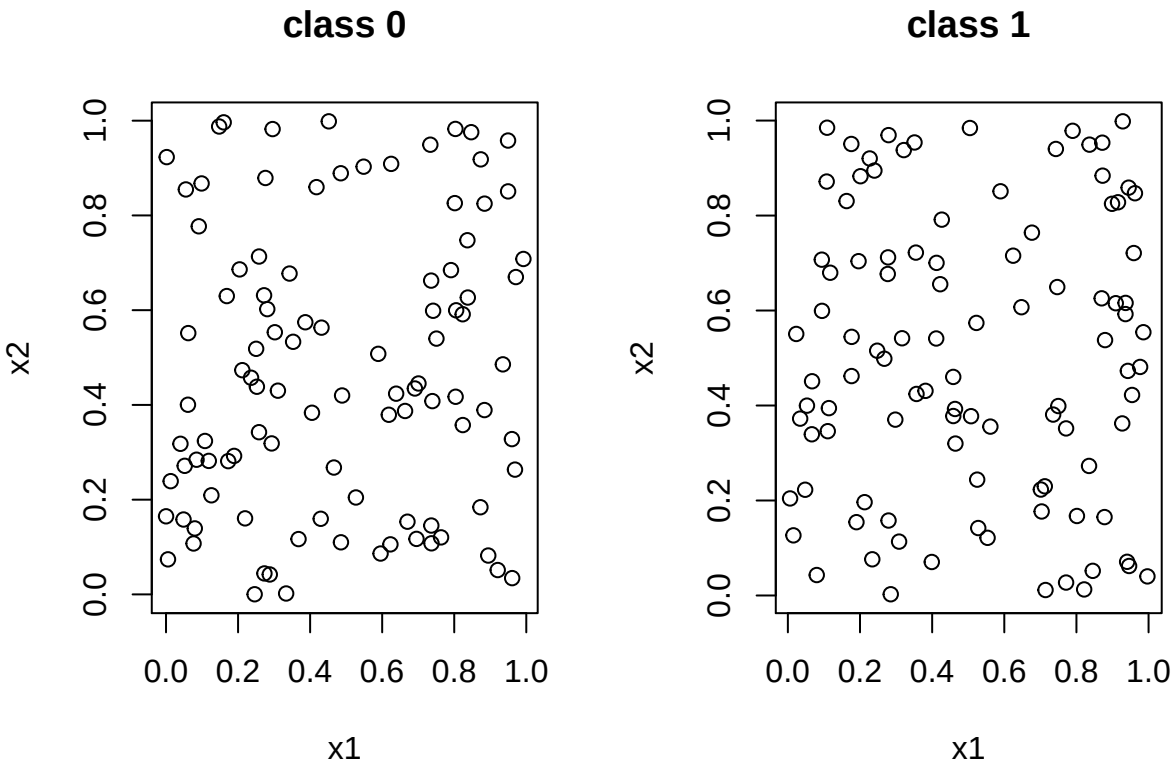
answer: A). Look at the plot on the lower right corner.

9. In the dataset `Boston`, what is the correlation between the variables `medv` and `nox`?

- A) 0.4835
- B) 0.6953
- C) 1
- D) We have not been given enough information so far to answer this question.

answer: D).

10. Given the following scattered plots (on training data) of feature measurements for class 0 and class 1 respectively, which classification method do you expect will perform best on training data before you actually implement the algorithms



- A) LDA
- B) QDA
- C) knn with proper choice of k

answer: C). It is quite clear no simple decision boundary can separate these two classes well.

11. In the same setting as the above question, which classification method do you expect will achieve below 5% classification error on test data before you actually implement the algorithms

- A) LDA
- B) QDA
- C) knn with proper choice of k
- D) None of the above

answer: D). The two classes do not seem to separate much, so on test data, really no classifier should be expected to deliver less than 5% classification error.

12. Divide a dataset (suppose no two observations have the same feature values) so that two-thirds are training set and one-third are test set, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next we use 1-nearest neighbors (i.e., $K = 1$) and get an error rate (over all data) of 12%. Based on these results, which method should we prefer to use for classification of new observations?

- A) Logistic regression
- B) 1-nearest neighbors
- C) the two methods are actually equally good.

answer: A). hint: the training error for KNN with K=1 is 0

13. Which of the following is /are classic example(s) of Simpson's paradox?

i) Berkeley graduate admission discrimination case

ii) People vs. collins case

- A) i)
- B) ii)
- C) i) and ii)
- D) neither i) nor ii)

answer: A). People vs. collins case is usually used to demonstrate one should not assume blindly events are independent.

Questions $14-15$ are based the `Caravan` data that we have seen in class.

```
library(ISLR); attach(Caravan); dim(Caravan)
```

```
## [1] 5822   86
```

```
lr.fit = glm(Purchase ~ ., data= Caravan, family = "binomial")
```

14. From the following confusion matrix,

```
lr.prob = predict(lr.fit, type = "response")
lr.predict = ifelse(lr.prob > 0.2, 'Yes', 'No')
table(lr.predict, Purchase)
```

```
##           Purchase
## lr.predict   No  Yes
##        No  5277  261
##        Yes  197   87
```

From the above confusion matrix, how many instances were incorrectly classified?

- A) 261
- B) 197
- C) 87
- D) 5277
- E) None of the above

answer: E). 261 + 196 instances were misclassified.

15. Also from the above confusion matrix, calculate the type II error (treat `No` as class `0`)

- A) 4.7%
- B) 69.4%
- C) 3.6%
- D) 75%

answer: D). $261/(261 + 87)$

16. Let

```
a  = matrix(c(1:9), 3, 3); b = 1
```

What happens if you compute $a + b$ in R?

- A) an error message because 'a' and 'b' are of different dimensions
- B) The outcome is a scalar 1
- C) The outcome is a 3×3 matrix, with the $(1, 3)$ entry equal 8
- D) The outcome is a 3×3 matrix, with the $(1, 3)$ entry equal to 4
- E) None of the above is correct

answer: C). This problem tests the vectorization concept that R turoial 2 covers.

part b) short answer questions (write concisely)

17. Explain how `KNN` classification method works for $k = 3$.

When given a instance to classify, 3-NN algorithm picks 3 points nearest to the given instance from the training set (1 point). The most frequent class among these 3 points will be assigned to the new instance (1 point).

18. We have a box that contains three balls: red, blue and black. First we draw three balls from the box without replacement, what is the outcome? What if we draw three balls with replacement from the box, how many possible outcomes are there? (We consider $\{red, red, black\}$ and $\{red, black, red\}$ as the same outcome.)

Without replacement: 1 outcome $\{red, blue, black\}$ (1 point)

With replacement: 10 outcomes $\{red, red, red\}$, $\{red, red, blue\}$, $\{red, red, black\}$, $\{red, blue, blue\}$, $\{red, blue, black\}$, $\{red, black, black\}$, $\{blue, blue, blue\}$, $\{blue, blue, black\}$, $\{blue, black, black\}$, $\{black, black, black\}$ (1 point)

19. Your team ran a linear regression with two independent variables X_1 and X_2 and outcome variable Y. From the R summary output, the coefficient estimates for X_1 is bigger than that of X_2. Your teammates tend to conclude that X_1 is more important than X_2 in predicting the Y variable. Do you agree, and why?

No. Coefficients are associated with the scale of the predictor. So smaller coefficient does not necessarily mean the predictor is less important, maybe it's just because the scale of predictor is very large. (There are many reasons to not agree. Grading is handled on case-by-case basis. 1 point for not agree. 1 point for a reasonable answer)

20. In statistical learning, sometimes we care about "prediction" more, and sometimes we care about "inference" more. i). If we were to build an email spam detector ourselves, which of the two goals should we focus more, and why? ii). If type I error means placing a normal email into the spam folder and type II error means placing a spam to the inbox. From a user perspective, which of the two errors is more severe, and why?

i). Prediction. We don't care much about why an email is a spam since our goal is a detector, not to study features of spam email. So prediction is our focus. (1 point)

ii). Type I error. For users, missing an email which was wrongly put into spam can lead to pontential loss of important information, while finding a spam in inbox is only something annoying. (1 point)