COMP1433 Introduction to Data Analytics

COMP1003 Statistical Tools and Applications

"*He who by reanimating the Old can gain knowledge of the New is fit to be a teacher*.「溫故而知新，可以為師矣。」"

--- Confucius (孔子)

Error Analysis for the Midterm Test

Due: *23:59 7 Apr 2022* (No late submission is acceptable!)

Name: ZHOU, Siyu

Student ID:

Read before you get started.

1) Please use the following template to finish the error analysis. You may copy and paste the template several times to handle multiple questions for error analysis.
2) If one correctly handles one question where they didn't give the correct answers in the midterm test, they would be able to take back 20% of the scores. So, assuming that one obtained X out of 100 in the midterm, the maximal scores possibly obtained after the post-midterm plan is $\mathrm{X} + (100 - \mathrm{X}) \cdot 0.2 \ = \ 0.8\mathrm{X} + 20$, still exhibiting positive and significant correlation with the midterm performance X for the fairness concern, while some minor awards (capped at 5% of the overall grade) are allowed to encourage students to work hard in the post-midterm plan.
3) For 2), "correctly handles one question" means that the error reasons make sense and the knowledge points are correctly summarized, and the similar question indeed covers the knowledge point.
4) It is optional to engage in the post-midterm plan and submit the error anlaysis.

4. (True or False) The angle of two vectors can be used to measure their distance.

Correct Answer: True

Your Answer: False

Error Reasons: The cosine value of the angle between vectors measures the difference between two vectors. The more smaller cosine similarity of two vectors, the shorter the distance is. It measures the proportional similarity of two variables in all directions. I only think of cosine similarity only showing the level of how two vectors nearing each other.

Knowledge Points: cosine similarity and vector distance

A Similar Question: (True or False) Cosine similarity of two vectors shows the distance between two vectors.

Answer: False

5. (True or False) For any function f(x), we will find its maximal or minimal solution via solving the equation of f'(x)=0, where f'(x) means the derivative of f(x).

Correct Answer: False

Your Answer: True

Error Reasons: A derivative f'(x) equal to 0 indicates that the function may have an extreme point. Where there is an extreme value, the f'(x) must be 0; The point where f'(x) is 0 is not necessarily the extreme point. Because there are three cases if f'(x) = 0, stationary point, maximum point, minimum point. I neglected the case - stationary point, then I didn't give out the correct solution.

Knowledge Points: three case of extreme point of a function

A Similar Question: (True or False) For a extreme point, it must be a maximum point, minimum point, and stationary point.

Answer: True.

16. (1 correct choice only) Suppose that we are analyzing average meal price in HK restaurants under the COVID-19 crisis and collected the data of meal prices from around 1,000 restaurants. If the boss would be interested in knowing the distribution of average meal price per restaurant. Which of the following graph should be a good alternative to demonstrate the required data distribution from the ggplot2 R package?

A. Barplot

B. Histogram

C. Scatterplot

D. All of them

Correct Answer: B

Your Answer: C

Error Reasons: Scatterplot graph can just show every restaurant' meal price, and histogram may show the frequency distribution of a quantitative variable (i.e. the different price numbers' frequency distribution). As for barplot, we can create a graph containing individual and interleaved bars for qualitative variable such as gender, and every student. I didn't notice the data distribution, then I didn't give out the correct solution.

Knowledge Points: Difference between Scatterplot, barplot and histogram.

A Similar Question: (1 correct choice only) Suppose that we are analyzing average grades for students' Mid-term exam. If the teaching staff wants to know the distribution of grades for every student. Which graph should be a good alternative to demonstrate the required data distribution from the ggplot2 R package?

A. Barplot

B. Histogram

C. Scatterplot

D. All of them

Answer: A

17. (1 correct choice only) There are 2 types of Happy Meal toys in the McDonald’s. Each of the toy type will be given with equal chances to a customer who buys the Happy Meal. Suppose that there is only one toy type that has the castle and Little Mary wants to get the castle very much. Let X denotes the random variable indicating the number of Happy Meals Little Mary should buy till she gets the castle. Then, the expected value of X should be ______.

A. 1 B. 3/2 C. 2 D. 5/2

Correct Answer: C

Your Answer: A

Error Reasons: During the test, I think key knowledge point for this question would be binomial distribution. And the expected value for binomial distribution is E(x)=np=2*(1/2)=1. Actually, we should set the expected value of the number of successful purchases as E(x) , and the probability of successful purchases as 1/2 . $\lim_{n\to+\infty}\sum_{i=1}^{n}(\frac{1}{2})^{n}$, let $S_n = \sum_{i=1}^{n}(\frac{1}{2})^{n}\times n$, then $2S_n = \sum_{i=1}^{n}(\frac{1}{2})^{n-1}\times n$, then $2S_n - S_n = S_n = \sum_{i=1}^{n}\left(\frac{1}{2}\right)^{n-1} = \frac{1-\left(\frac{1}{2}\right)^{2}}{1-\frac{1}{2}} = 2(1-\left(\frac{1}{2}\right)^{n})$, then $E(x) = \lim_{n\to+\infty}\sum_{i=1}^{n}2(1-\left(\frac{1}{2}\right)^{n}) = 2$

Knowledge Points: expected value for random variable

A Similar Question: (1 correct choice only) There are 12 types of toys in one series of PopMart’s products. Each of the toy type will be given with equal chances to customers when they are buying specific series. Suppose that there is only one toy is in angel costume and Little Mary wants to get that toy. Let X denotes the random variable indicating the number of toys Little Mary should buy till she gets the castle. Then, the expected value of X should be ______.

A. 6 B. 9 C. 12 D. 15

Answer: 12

19. (1 correct choice only) In a new research paper published by University B, it takes 5 days on average for a COVID-19 patient to have > 30 CT value (tested negative). It is known that the time for a COVID-19 patient to have > 30 CT value satisfies general normal with the standard deviation as 2.5 days. University P would be interested in knowing whether they can trust University B's results (the null hypothesis). So, they examined the sample of 64 COVID-19 patients and the time for their CT value to go back to a > 30 status is 5.5 days on average. Given the observations, if University P accepts University B's statement on the level of significance as x, then ______.

A. x<5.48% B. x>5.48% C. x<10.96% D. x>10.96%

Correct Answer: C

Your Answer: B

Error Reasons: During the test, I forget to multiply the probability by 2 because I forget this should be considered as a two-sides probability. And average days of sample 64 COVID-19 patients to obtain >30 CT value, the x should be calculated in this way. $P(|\bar{X} - \mu| \geq 0.5) = P\left(|Z| \geq \frac{0.5}{\frac{2.5}{\sqrt{64}}}\right) =$ $2\,\emptyset\,(-1.6) = 2 \times 5.48\% = 10.96\%.$

Knowledge Points: Hypothesis tests

A Similar Question: (1 correct choice only) A sample size 64 from a normal population with $\sigma = 0.234$, has sample mean $\bar{X} = 4.847$, will we____

A. accept H0 at the level of significance 10%

B. reject H0 at the level of significance 10%

C. all of above

Answer: B

21. (1 or multiple correct choice(s)) Which of the following R commands allow us to create a matrix of $\begin{pmatrix} 2 & 4 \\ 3 & 5 \end{pmatrix}$.

A. matrix(c(2:5), nrow=2)

B. matrix(c(2,3,4,5), nrow=2, ncol=2)

C. matrix(c(2,3,4,5), by row=FALSE, nrow=2, ncol=2)

D. matrix(c(2,4,3,5), by row=TRUE, ncol=2)

(AB) Page 34, Lecture 6. “by row” in C and D should be “byrow”.

Correct Answer: AB

Your Answer: ABCD

Error Reasons: I didn’t notice there’s a space between “by” and “row”. In RGui, it will shows one example: *matrix(data = NA, nrow = 1, ncol = 1, byrow = FALSE, dimnames = NULL)* by entering *help(matrix)*. This means that there’s no space between “by” and “row”, and “by row” in C D should be “byrow”.

Knowledge Points: R command for how to input data.

A Similar Question: (1 or multiple correct choice(s)) Which of the following R command(s) allow(s) us to create a vector with the integers from 2 to 6?

A. c(2:6)

B. c(2,3,4,5,6)

C. 2:6

D. [2:6]

Answer: ABC

22. (1 or multiple correct choice(s)) Suppose we know the probability of event A conditioned on C is p(A|C), the probability of event B conditioned on C is p(B|C), and the probability of C is p(C). Which of the following probabilities can be calculated for sure (there's no independence assumption among A, B, and C):

A. p(A)

B. p(B)

C. p(AC)

D. p(ABC)

Correct Answer: C

Your Answer: CD

Error Reasons: During the test, I did not think clearly about calculating P(ABC), because we can already calculate P(AC) and P(BC). When we don't know the independence assumption among A, B, C, we cannot calculate P(ABC). If A, C are independent, and B, C are independent, we can calculate p(A) and p(B).

Knowledge Points: Conditional Probability

A Similar Question: (1 or multiple correct choice(s)) Suppose we know the probability of event A conditioned on event B is (p(A|B)), and the probability that event B happens is P(B). And event A and B are independent. Assuming P(A) is not equal to 0, Which of the following probabilities can be calculated for sure:

A. P(A)

B. P(B)

C. P(AB)

D. All above

Answer: D

25. (1 or multiple correct choice(s))For naive Bayesian classifier, which of the following statements are correct?

A. It is not sensitive to missing data, and the algorithm is relatively simple, which is often used in text classification

B. Naive Bayes is a discriminant model, which calculates the conditional probability by learning the known samples.

C. It has a solid mathematical foundation and stable classification efficiency.

D. It is relevant to the choice of a priori probability, so there is a certain error rate in classification.

Correct Answer: ACD

Your Answer: ABCD

Error Reasons: Naïve Bayes is a generative model because we can calculate probability for unknown samples. During the exam, when I read the following explanation, this option seemed reasonable because we learned tree diagram for solving those classification problems and I chose it without thinking about it carefully.

Knowledge Points: Properties of Naïve Bayesian classifier

A Similar Question: (1 or multiple correct choice(s)) Naïve Bayes is a(n) ____ classifier.

A. discrete

B. generative

C. linear

D. non-linear

Answer: BC

26. (1 or multiple correct choice(s)) Which of the following statements is/are illegal?

A. `-`(5, 2)

B. +(5, 22)

C. 2d_matrix <- matrix(11:16, nrow=3, ncol=2)

D. x <- c(1, 4, 6.25);x[c(-1,2)] <- c(11, 13)

Correct Answer: BCD

Your Answer: AB

Error Reasons: A is a subtraction calculation, and I forget this expression is also one way to express subtraction calculation. B are illegal because the "+" symbol did not quote by single quotation marks. As for C, "2" is an unexpected symbol for R language, "2" cannot be the beginning characters for a object. There is an indexing problem in the second part of choice D, we cannot have both positive index and negative index.

Knowledge Points: R commands' errors

A Similar Question: (1 or multiple correct choice(s)) Which of the following statements is/are illegal?

A. read.table("./mydata.csv", header=TRUE, sep="/t")

B. +(5,4)

C. mymatrix <- matrix(11:14, nrow=2, ncol=2)

D. prob_positive <- pnorm(1.96) ; probpositive

Answer: D

27. (1 or multiple correct choice(s)) Which of the following is(are) the assumptions of a Naïve Bayes classifier?

A. Position of the words doesn’t matter.

B. The probability to observe words are independent conditioned on the class.

C. The probability of word occurrences in the documents are independent with each other.

D. A document can be represented by the count of words

Correct Answer: ACD

Your Answer: BC

Error Reasons: I didn’t read all of four choice carefully and didn’t check the both assumptions carefully. There’re two assumptions, Bag of Words assumption (Assume position doesn’t matter), Conditional Independence (Assume the feature probabilities are independent given the class). A and C are correct. Also, documents can be represented as feature x1, x2, …, xn, then choice D is correct.

Knowledge Points: Assumptions about Naïve Bayes Classifier

A Similar Question: Which of the following is(are) the assumptions of a Naïve Bayes classifier?

A. Position of the words influence the Naïve Bayes classifier.

B. The probability to observe words are independent conditioned on the class.

C. The probability of word occurrences in the documents are independent with each other.

D. A document can be represented by the count of words

Answer: CD

29. (1 or multiple correct choice(s)) Which of the following operation(s) is (are) FOR SURE doable in the linear algebra:

A. The Euclidean distance of two equal vectors.

B. The multiplication of two equal matrices.

C. The angle of two equal vectors.

D. The addition of two equal matrices.

Correct Answer: AD

Your Answer: BCD

Error Reasons: During the test, I mixed up with vector and matrices in choice B, C. But I'm not sure why I didn't choose A as a correct answer, maybe I was not thinking clearly. A should be doable because equal vectors have same dimension, then Euclidean distance can be calculated. B may not be doable because row number may not equal to the column number. C may not be doable because if the vector's length equals 0, then angle cannot be calculated.

Knowledge Points: Calculation operation of vectors and matrices

A Similar Question: (1 or multiple correct choice(s)) Which of the following operation(s) is (are) FOR SURE doable in the linear algebra:

A. The dot product of two equal vectors.

B. The multiplication of two equal matrices.

C. The Euclidean distance of two equal vectors.

D. The addition of two equal matrices.

Answer: ACD