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~~ 10 Apr 2022* (ONLY late submission is acceptable!)

Name: [redacted]

Student ID: [redacted]

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 $X + (100 - X) \cdot 0.2 = 0.8X + 20$, 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.

[8]. [(True or False) In R programming, the symbol NaN can be used to represent missing values of the data for some imperfect dataset.]

Correct Answer: [False]

Your Answer: [True]

Error Reasons: [I misunderstood the function of the NaN symbol in R. I thought that NaN actually means that it is representing the value that doesn't exist. However, it is talking about not a number instead of the value straight up not existing. If the answer were to be true, it would have to say null instead of NaN.]

Knowledge Points: [Knowledge of R symbols]

A Similar Question: [T/F: null is the same as NaN symbol in R programming.]

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: [Just because the derivative is zero does not mean you have found the max or min. It only guarantees the slope of the said point is zero. The function can keep on decreasing or increasing however it wants, depending on the function. The whole function has to be looked at instead of just when the derivative equals zero.]

Knowledge Points: [Derivative meanings of curves]

A Similar Question: [T/F: When a derivative goes from negative to positive while going left to right, it guarantees that you have reached the absolute minima of the function]

False

[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: [Histogram]

Your Answer: [Barplot]

Error Reasons: [I did not understand the question correctly and at the same time I was not familiar enough with the best use case scenario for the different types of graphs mentioned in the possible answers. Since we are showing the *distribution* of the prices, it would be a good idea to use a Histogram]

Knowledge Points: [Different purposes of graphs and their best use case scenarios]

A Similar Question: [What would be the best use for histograms?]
A. Illustrating the relationship between a numeric and a categorical variable.

B. Illustrating frequency distribution of a quantitative variable.

C. Illustrating two variables that pair well together so that you can show their relationship.

D. All of the above

[15]. [(1 correct choice only) Which result does the following code describe?

```
r.n <- function(r,n){
    a <- prod(2:n)/prod(2:(n-r))/prod(2:r)
    return(a)
}
```

]

Correct Answer: [r choose n]

Your Answer: [r permute n]

Error Reasons: [I mixed up the permutation and the choose formula, and in the process, I also ended up converting the formula to R incorrectly, so I had to end up guessing this question. The combination formula is supposed to be $\frac{n!}{r!(n-r)!}$ However, I thought it was the permutation formula which was $\frac{n!}{(n-r)!}$ I missed the r! when interpreting the code.]

Knowledge Points: [Incorrect understanding of probability formulas and converting them to R code.]

A Similar Question: [Which result does the following code describe?]

```
r.n <- function(r,n){
    a <- prod(2:n)/prod(2:(n-r))/prod(2:r)
    return(a)
}
```

A. r perm n

B. r choose n

C. n perm r

D. n choose r

[23]. [Given three vectors a, b and c and two scalars β and γ, find the correct statement(s) in the following: (A, B, C, D)]

$(\beta+\gamma)(a+b)=(\beta+\gamma)a+(\beta+\gamma)b$

$-\beta a-\gamma b=-\gamma b-\beta a$

$\gamma a+\beta(b+c)=(\gamma a+\beta b)+\beta c$

Correct Answer: [$\beta a+\gamma a=(\beta+\gamma)a$]

Your Answer: [$(\beta+\gamma)(a+b)=(\beta+\gamma)a+(\beta+\gamma)b$ $\beta a+\gamma a=(\beta+\gamma)a$]

Error Reasons: [I was starting to run out of time and was starting to rush the questions. I didn't bother to look at the other two more carefully because I thought the two was all there is in terms of being equivalent to each other. $-\beta a-\gamma b=-\gamma b-\beta a$ has to be one of the most obvious ones because the order was just switched, so they are still equal. I feel stupid for not noticing this. For $\gamma a+\beta(b+c)=(\gamma a+\beta b)+\beta c$, since there is no multiplication and division on the other side, switching the order is still fine]

Knowledge Points: [Vector addition]

A Similar Question: [Which one is not always true?]

A. $a+b=b+a$

B. $(a+b)+c=a+(b+c)$

C. $a+0=0+a$

D. $a+\frac{b}{c}=\frac{a}{b}+c$

[11]. [(1 correct choice only) Suppose we are interested in predicting whether a news report concerns a ``vaccine'' topic or not (e.g., to work on COVID-19 related applications). In our prior knowledge, 30% of news reports are about ``vaccine'' while 70% are not. Besides, we know that the probability of observing the word `` Pfizer'' in a ``vaccine'' news report is 60% and that in a ``non-vaccine'' news report is 20%. Now, given a news report containing `` Pfizer '', the probability that the news report is about ``vaccine'' is _____.]

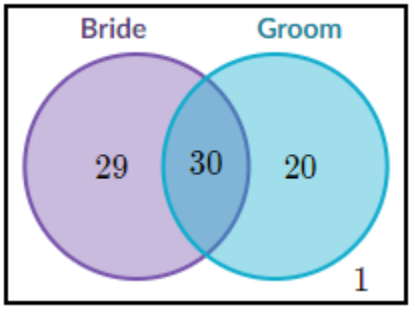
Correct Answer: [Larger than 50%]

Your Answer: [Equal to 50%]

Error Reasons: [I did not utilize the probability formula for permutations correctly. I do not remember what formula I attempted to use during the exam. I did not have a solid enough understanding of application of these formulas for this question. For questions like these, I need to see what is given to me and what I will be able to plug into the formulas from the lectures.]

Knowledge Points: [Conditional probability formulas and applications]

A Similar Question: [The usher at a wedding asked each of the 80 guests whether they were a friend of the bride or of the groom. Here are the results:



Given that a randomly selected guest is a friend of the groom, find the probability they are a friend of the bride.]

A. >50%

B. <50%

C. =50%

[3]. [(True or False) The sample mean approximates the population mean μ for any sample size n.]

Correct Answer: [False]

Your Answer: [True]

Error Reasons: [I thought that they were going to be similar, so I ended up thinking that the mentioned statement in the question was true. However, this is not true because this only applies to samples that are big. For small samples, this tends to be unreliable so you shouldn't rely on it to tell. I forgot to take into the account of the Law of Large numbers also.]

Knowledge Points: [Law of Large Numbers]

A Similar Question: [T/F: A 6-sided dice roll will become closer and closer to 3.5 the more times you roll it.]

True

[7]. [(True or False) In logistic regression, the sigmoid function only works for binary classification.]

Correct Answer: [True]

Your Answer: [False]

Error Reasons: [I thought the functions would work for other use cases and I feel dumb for forgetting that it will only work for binary classifications because values further away from, the cluster of values will end up]

Knowledge Points: [Sigmoid/Logistic function]

A Similar Question: [T/F: $\sigma'(x) = \sigma(x)(1 - \sigma(x))$]

True

[18]. [(1 correct choice only) Given the following short movie reviews, each labeled with a genre, either comedy or action (the genre name is in [boldface] and the word in the reviews are in italic):

• fun, couple, love, love [comedy]

• fast, furious, shoot [action]

• couple, fly, fast, fun, fun [comedy]

• furious, shoot, shoot, fun [action]

• fly, fast, shoot, love [action]

Given a new document D: fast, couple, shoot, fly, we should assign D to the class of ____ measured by a Naive Bayes classifier with add-1 smoothing. The likelihood of observing the words in D conditioned on that class is _____.]

Correct Answer: [action, $2.858 * 10^{-4}$]

Your Answer: [comedy, $2.858 * 10^{-4}$]

Error Reasons: [Even though I think I used the right formula for this question, I was rushing because I didn't proportionally allocate my times properly. I thought this question was very challenging and ended up rushing through this question towards the end and I believe I accidently swapped variables somewhere along my lines of work. In addition, I think my incompetence with probability formulas also contributed to me somehow swapping variables during the process of solving.]

Knowledge Points: [Probability formulas with Naïve Bayes application]

A Similar Question: [Let's say:

- the probability of dangerous fires are rare (1%)
- but smoke is fairly common (10%) due to barbecues,
- and 90% of dangerous fires make smoke

Can you find the probability of dangerous Fire when there is Smoke?]

A. Unsolvable

B. 10%

C. 50%

D. 9%

[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 _____.]

Correct Answer: [x<10.96%]

Your Answer: [x>10.96%]

Error Reasons: [While doing this question, I think I used the right formula, but then towards the end when I was submitting answer, I was not sure if I remember I either clicked the wrong answer because they looked similar, or I swapped some numbers during the calculation and ended up getting an inverted answer.]

Knowledge Points: [Probability applications, formulas, and hypothesis testing]

A Similar Question: [

Carl the farmer has started using organic fertilizer this year. The average weight of his tomatoes last year was 10.3 ounces. A sample of 40 tomatoes from this year's crop has a mean weight of 9.8 ounces with a standard deviation of 1.7 ounces. Test the claim that the mean weight of all of this year's tomatoes is different from last year's mean. Test this claim at the 0.05 significance level.

Find the test statistic conclusion about the null hypothesis and why]

A. Insufficient

B. <10%

C. <5%

[29]. [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: [ABD]

Error Reasons: [I forgot that since this is linear algebra that was being used during the question, I did not take into consideration the fact that it was not possible to do so in linear algebra. Since it is **multiplication** of two matrixes, it will not be doable with just linear algebra]

Knowledge Points: [matrixes, linear algebra]

A Similar Question: [Multiply the following two matrixes of the same size using only linear algebra

$$[1\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ ][1\ 0\ 20\ 0\ 1\ 0\ 5\ 0\ 1\ ]$$

]

A. $[1\ 0\ 20\ 4\ 1\ 0\ 5\ 0\ 1\ ]$

B. $[1\ 0\ 20\ 0\ 1\ 0\ 5\ 5\ 1\ ]$

C. Not solvable

D. $[1\ 0\ 20\ 0\ 1\ 11\ 5\ 0\ 1\ ]$

[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: [p(AC)]

Your Answer: [p(A), p(B), p(AC)]

Error Reasons: [I forgot to take into consideration the fact that they are guaranteed to be independent to each other or not. In order for the other answers to be true, there has to be some form of independence assumption among them with each other. I did not consider that I had to choose the ones that involved B because those were the ones that could not be done without any assumption of independence]

Knowledge Points: [Probability]

A Similar Question: [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, assuming they are independent to each other]

A. $P(A)$

B. $P(B)$

C. $P(ABC)$

D. $P(AC)$

[28]. [Which of the following statement(s) about the definite integral $\int_{-\infty}^{\infty} e^{-\frac{1}{2}(2x-3)^2} dx$]

Correct Answer: [ACD]

Your Answer: [CD]

Error Reasons: [For this question, I was careless and did not read the possible range answer properly and ended up not taking it into consideration. When a question like this comes up, I need to think about how the end answer will be when I substitute more simple terms in the place of other more complicated terms. For question, I just did not read the possible range answer properly]

Knowledge Points: [Calculus]

A Similar Question: [$\int_2^4 \left(\frac{6+x^2}{x^3}\right) dx$]

A. $\frac{9}{16}$

B. $ln(2) + \frac{9}{16}$

C. $-\ ln(2) + \frac{9}{16}$

D. None of the above