COMP1003/1433 Midterm Test Questions

A total is 100', 3' for each True or False question and multichoice questions with only one correct choice. 4' for each multichoice question with one or multiple correct choices.

1. (True or False) Data analytics is also known as data mining.

   False. Data analytics include both data mining and communication and concerns more with the entire methodology while data mining may focus on an individual analysis step. P57, Lecture 1.

2. (True or False) Naïve Bayes is only able to tackle binary classification while logistic regression allows the classification with multiple classes.

   False. Both Naïve Bayes and logistic regression are text classifiers, and the number of classes can be two or more. P36-37, Lecture 2.

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

   False. The sample mean approximates the population mean for a very large sample size n. P17, Lecture 3.

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

   True. The cosine of the angle is the cosine similarity measure. P21, Lecture 4. Larger angle indicates farther distance (and smaller similarity).

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).

   False. A variable resulting in the derivative of 0 might not an optimal solution. P23-24, Lecture 5.

6. (True or False) The differentiation process for a function f(x) allows the measurement of the instantaneous rate of change at (x, f(x)).

   True. P10, Lecture 5.

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

   True. As can be seen from the function graph, the outlier values squash towards two sides 0 or 1. P33, Lecture 5.

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

   False. The symbol NA is used to represent missing values. P45, Lecture 6.

9. (True or False) The R code `x=seq(-4,4,0.01); plot(x, pnorm(x, 0, 1), col = "red");' draws the density function diagram of normal distribution.

False. The R function `pnorm(q)’ for the definition of cumulative probability function instead of the density function. P56, Lecture 6.

10. (True or False) If two discrete random variables X and Y are independent, then we can have E(XY)=E(X)E(Y).

    True. If X and Y are independent, we can have P(X=x, Y=y)=P(X=x)P(Y=y) for any given x and y. Then, with the definition of expected values, we can have E(XY)=E(X)E(Y).

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 ______.
    A. Larger than 50%
    B. Smaller than 50%
    C. Equal to 50%

    (A) V: the news report is about ``vaccine’’; NV: the news report is not about ``vaccine’’; P: the news report contains the word ``Pfizer''. Then $P(V|P) = P(P|V) * P(V)/(P(P|V) * P(V) + P(P|NV) * P(NV)) = 0.6 * 0.3/(0.6 * 0.3 + 0.2 * 0.7) = 0.5625 > 50\%.$

12. (1 correct choice only) There are five boxes, where one carries a paper cheque with $1M while the other four each carry plain paper. You will never know which box carries the cheque till one draws the paper out of the box and release the results. Your friend first selects a random box and announced that the paper drawn is plain paper. Now it is your turn to do the lucky draw. The probability for you to draw the cheque becomes ____ compared to the moment before your friend’s drawing result is announced.
    A. Larger
    B. Smaller
    C. Unchanged

    (A) The sample space becomes smaller because your friend helps you exclude a box with plain paper.

13. (1 correct choice only) Given two vectors a=(2,2, 2, 2) and b=(0, 4, 0, 3), the cosine similarity of a and b is: (C)
    A. 0.3
    B. 0.4
    C. 0.7
    D. 0.8

    (C) $\frac{0+8+0+6}{\sqrt{(16*25)}} = 14/20 = 0.7$

14. (1 correct choice only) For x in the range of [0, 1], the area above x-axis and under the curve $\mathrm{f(x)} = \mathrm{x}^3 + e^{2x}$ is in the range of _____.
    A. [0,1]
    B. [1,2]
    C. [2,3]
    D. [3,4]

(D) $\int_0^1 (x^3 + e^{2x})dx = \left(\frac{x^4}{4} + \frac{e^{2x}}{2}\right)|_0^1 = \frac{1}{4} + \frac{e^2}{2} - \frac{1}{2} = 3.445$ in the range of [3,4].

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

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

    A. n choose r
    B. n permute r
    C. r choose n
    D. r permute n

(C) The code is to calculate $\frac{r!}{(r-n)!n!}$. So it is r choose n.

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

(B) We are able to collect the meal prices from the restaurants and can further calculate the average price per a restaurant. So, there would be a vector with 1,000 price numbers (quantitative variables). It would be good to use histograms following the description in page 15, Lecture 7.

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

(C) From the question, we can have $\mathrm{E(X)} = \lim_{n\to+\infty} \sum_{i=1}^{n} \left(\frac{1}{2}\right)^{n} \cdot n$. Let $\mathrm{S}_n = \sum_{i=1}^{n} \left(\frac{1}{2}\right)^{n} \cdot n$ and hence $2\mathrm{S}_n = \sum_{i=1}^{n} \left(\frac{1}{2}\right)^{n-1} \cdot n$. So, we'll have $2\mathrm{S}_n - S_n = \mathrm{S_n} = \sum_{i=1}^{n} \left(\frac{1}{2}\right)^{n-1} = \frac{1-\left(\frac{1}{2}\right)^{n}}{1-\frac{1}{2}} = 2\left(1-\left(\frac{1}{2}\right)^{n}\right)$. Therefore, $\mathrm{E(X)} = \lim_{n\to+\infty} S_n = 2$.

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 ______.

A. comedy, 1.714×10^{-4}
B. action, $2.858 \cdot 10^{-4}$
C. comedy, $2.858 \cdot 10^{-4}$
D. action, $1.714 \cdot 10^{-4}$

(B) The vocabulary $\mathrm{V} = \{\mathrm{fun, couple, love, fast, furious, shoot, fly}\}$. So its size $|\mathrm{V}| = 7$

Let C denotes comedy genre and A denotes action. So, the prior of the two class labels are:

$\mathrm{P(C)} = \frac{|\mathrm{C}|}{|\mathrm{C}|+|\mathrm{A}|} = \frac{2}{5}$ $\mathrm{P(A)} = \frac{|\mathrm{C}|}{|\mathrm{C}|+|A|} = \frac{3}{5}$

For likelihoods of observing different words are calculated as following:

$$\mathrm{P(fast|C)} = \frac{\mathrm{count(fast, C)} + 1}{\mathrm{count(C)} + |\mathrm{V}|} = \frac{1+1}{9+7} = \frac{1}{8}$$

$$\mathrm{P(fast|A)} = \frac{\mathrm{count(fast, A)} + 1}{\mathrm{count(A)} + |\mathrm{V}|} = \frac{2+1}{11+7} = \frac{1}{6}$$

$$\mathrm{P(couple|C)} = \frac{\mathrm{count(couple, C)} + 1}{\mathrm{count(C)} + |\mathrm{V}|} = \frac{2+1}{9+7} = \frac{3}{16}$$

$$\mathrm{P(couple|A)} = \frac{\mathrm{count(couple, A)} + 1}{\mathrm{count(A)} + |\mathrm{V}|} = \frac{0+1}{11+7} = \frac{1}{18}$$

$$\mathrm{P(shoot|C)} = \frac{\mathrm{count(shoot, C)} + 1}{\mathrm{count(C)} + |\mathrm{V}|} = \frac{0+1}{9+7} = \frac{1}{16}$$

$$\mathrm{P(shoot|A)} = \frac{\mathrm{count(shoot, A)} + 1}{\mathrm{count(A)} + |\mathrm{V}|} = \frac{4+1}{11+7} = \frac{5}{18}$$

$$\mathrm{P(fly|C)} = \frac{\mathrm{count(fly, C)} + 1}{\mathrm{count(C)} + |\mathrm{V}|} = \frac{1+1}{9+7} = \frac{1}{8}$$

$$\mathrm{P(fly|A)} = \frac{\mathrm{count(fly, A)} + 1}{\mathrm{count(A)} + |\mathrm{V}|} = \frac{1+1}{11+7} = \frac{1}{9}$$

Finally, we have:

$\mathrm{P(D|C) \cdot P(C) = P(fast|C) \cdot P(couple|C) \cdot P(shoot|C) \cdot P(fly|C)P(C)} = \frac{1}{8} \cdot \frac{3}{16} \cdot \frac{1}{16} \cdot \frac{1}{8} \cdot \frac{2}{5} = 7.324 \cdot 10^{-5}$

And $\mathrm{P(D|A) \cdot P(A) = P(fast|A) \cdot P(couple|A) \cdot P(shoot|A) \cdot P(fly|A)P(A)} = \frac{1}{6} \cdot \frac{1}{18} \cdot \frac{5}{18} \cdot \frac{1}{9} \cdot \frac{3}{5} = 1.714 \cdot 10^{-4}$

Therefore, we will classify the new document D into action genre ($1.714 \cdot 10^{-4} > 7.324 \cdot 10^{-5}$). The likelihood to observe words in D conditioned on action is $\frac{1}{6} \cdot \frac{1}{18} \cdot \frac{5}{18} \cdot \frac{1}{9} = 2.858 \cdot 10^{-4}$.

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%

    (C) Let $\bar{X}$ denotes the average days for the sampled 64 COVID-19 patients to obtain >30 CT value. The time for all COVID-19 patients to obtain >30 CT value satisfies general normal with the expected value of μ and standard deviation $\sigma = 2.5$ days. Let $\mathrm{Z} = \frac{\bar{X}-\mu}{\sigma/\sqrt{64}}$The p-value is $\mathrm{P}(|\bar{X} - \mu| \geq 0.5) = \mathrm{P}\left(|\mathrm{Z}| \geq \frac{0.5}{\frac{2.5}{\sqrt{64}}}\right) = 2\phi(-1.6) = 2 \cdot 5.48\% = 10.96\%$.

20. (1 correct choice only) Given 3 clusters, the representative (centroid) of Cluster 1, 2, 3, and 4 are (1,3,3), (7,1,4), (0,0,0), and (5,8,1), respectively. For a new data point p=(3,5,2), according to cluster assignment strategy of k-means algorithm (based on Euclidian distance), which cluster should p belong to:
    A. Cluster 1

B. Cluster 2
C. Cluster 3
D. Cluster 4

(A) Let c1, c2, c3, and c4 represent the centroids of Cluster 1, 2, 3, and 4. Then, we have the following: $||\mathrm{p}-\mathrm{c}1|| = \sqrt{2^2+2^2+1^2} = \sqrt{9}$, $||\mathrm{p}-\mathrm{c}2|| = \sqrt{4^2+4^2+2^2} = \sqrt{36}$, $||\mathrm{p}-\mathrm{c}3|| = \sqrt{3^2+5^2+2^2} = \sqrt{38}$, and $||\mathrm{p}-\mathrm{c}4|| = \sqrt{2^2+3^2+1^2} = \sqrt{14}$. So, p is closest to Cluster 1, we should assign it to this cluster.

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”.

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)

    (C) P(AC)=P(A|C)P(C). Others cannot be calculated because there’s no independence assumption.

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)
    A. $-\beta\mathrm{a} - \gamma\mathrm{b} \ = -\gamma\mathrm{b} - \beta\mathrm{a}$
    B. $\gamma\mathrm{a} \ + \beta(\mathrm{b} \ + \ \mathrm{c}) \ = \ (\gamma\mathrm{a} \ + \beta\mathrm{b}) \ + \beta\mathrm{c}$
    C. $(\beta+\gamma)(\mathrm{a} \ + \ \mathrm{b}) = (\beta+\gamma)\mathrm{a} \ + (\beta+\gamma)\ \mathrm{b}$
    D. $\beta\ \mathrm{a} \ + \gamma\ \mathrm{a} = (\beta \ + \gamma)\ \mathrm{a}$

    (ABCD) Page 9 and 11, Lecture 4.

24. Find the correct statement(s) in the following: (B, D)
    A. [f(g(x))]’ = f’(x)g’(x)
    B. [f(g(x))]’ = f’(g(x))g’(x)
    C. [f(x)g(x)]’ = f’(x)g’(x)
    D. [f(x)g(x)]’ = f’(x)g(x) + f(x)g’(x)

    (BD) B is the chain rule while D is the product rule.

25. 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.

(ACD) B is incorrect because Naïve Bayes is a generative model. Other statements are true derived from our discussions in Lecture 2.

26. 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)

(BCD) One may explore that in the R system.

27. 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

(ABD) P45, Lecture 2.

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

A. The result is in the range of [1,2].

B. The exact result is 1.5.

C. Chain rule can help solve the problem.

D. The properties of normal distribution may be helpful.

(ACD) Let f(x) $= e^{-\frac{1}{2}(2x-3)^2}$, $\mathrm{u} = 2\mathrm{x} - 3$, so $\mathrm{du} = 2\mathrm{dx}$. We can then have $\int_{-\infty}^{\infty} f(x)dx =$ $\int_{-\infty}^{+\infty} \frac{1}{2} \mathrm{e}^{-\frac{u^2}{2}} du = \frac{1}{2}\sqrt{2\pi} = 1.2533$

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.

(AD) Equal vectors have the same dimension, so A is true. Similarly, equal matrices have the same size, so D is true. B may not be doable if the row number does not equal to the column number. C may not be doable if the vector is a zero vector (which may correspond to the length of 0 in the denominator of cosine similarity).

30. Given the following data observations: 6, 3, 2, 4, 9, 1, 7, 6, which of the following is (are) correct?
    A. The sample mean of these numbers is 4.75.
    B. The sample median of these numbers is 5.
    C. The sample range of these numbers is 8.
    D. The sample standard deviation of these numbers is in the range of [7,8].

    (ABC) Following the formula in page 11, Lecture 3, we can verify that ABC all correct. The sample variance is 7.357 while the sample standard deviation is 2.712 (not in the range of [7,8]).