COMP1003/1433 Exercise Questions (Week 1 – Week 12)

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) 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.

3. (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.

4. (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.

5. (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.

6. (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.

7. (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).

8. (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%.

9. (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.

10. (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) 0+8+0+6/sqrt(16*25)=14/20=0.7

11. (1 correct choice only) For x in the range of [0, 1], the area above x-axis and under the curve $f(x) = x^3 + e^{2x}$ 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].

12. (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 $E(X) = \lim_{n\to+\infty} \sum_{i=1}^{n} \left(\frac{1}{2}\right)^n \cdot n$. Let $S_n = \sum_{i=1}^{n} \left(\frac{1}{2}\right)^n \cdot n$ and hence $2S_n = \sum_{i=1}^{n} \left(\frac{1}{2}\right)^{n-1} \cdot n$. So, we'll have $2S_n - S_n = S_n = \sum_{i=1}^{n} \left(\frac{1}{2}\right)^{n-1} = \frac{1-\left(\frac{1}{2}\right)^n}{1-\frac{1}{2}} = 2\left(1 - \left(\frac{1}{2}\right)^n\right)$.

Therefore, $E(X) = \lim_{n\to+\infty} S_n = 2$.

13. (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, $1.714 \text{ x } 10^{-4}$
B. action, $2.858 \cdot 10^{-4}$
C. comedy, $2.858 \cdot 10^{-4}$
D. action, $1.714 \cdot 10^{-4}$

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

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

$$P(C) = \frac{\text{count}(C)}{\text{count}(C)+\text{count}(A)} = \frac{2}{5} \quad P(A) = \frac{\text{count}(A)}{\text{count}(C)+\text{count}(A)} = \frac{3}{5}$$

For likelihoods of observing different words are calculated as following:

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

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

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

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

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

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

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

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

Finally, we have:

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

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

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 $\frac{1}{6} \cdot \frac{1}{18} \cdot \frac{5}{18} \cdot \frac{1}{9} = 2.858 \cdot 10^{-4}$.

14. (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 $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{64}}$The p-value is

$$P(|\bar{X} - \mu| \geq 0.5) = P\left(|Z| \geq \frac{0.5}{\frac{2.5}{\sqrt{64}}}\right) = 2\phi(-1.6) = 2 \cdot 5.48\% = 10.96\%.$$

15. (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. $||p - c1|| = \sqrt{2^2 + 2^2 + 1^2} = \sqrt{9}$, $||p - c2|| = \sqrt{4^2 + 4^2 + 2^2} = \sqrt{36}$, $||p - c3|| = \sqrt{3^2 + 5^2 + 2^2} = \sqrt{38}$, and $||p - c4|| = \sqrt{2^2 + 3^2 + 1^2} = \sqrt{14}$. So, p is closest to Cluster 1, we should assign it to this cluster.

16. (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): (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.

17. (1 or multiple correct choice(s)) 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. $-\beta a - \gamma b = -\gamma b - \beta a$
B. $\gamma a + \beta(b + c) = (\gamma a + \beta b) + \beta c$
C. $(\beta + \gamma)(a + b) = (\beta + \gamma)a + (\beta + \gamma)b$
D. $\beta a + \gamma a = (\beta + \gamma)a$

(ABCD) Page 9 and 11, Lecture 4.

18. (1 or multiple correct choice(s)) 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.

19. (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.

    (ACD) B is incorrect because Naïve Bayes is a generative model. Other statements are true derived from our discussions in Lecture 2.
20. (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

    (ABD) P34, Lecture 2.
21. (1 or multiple correct choice(s)) Which of the following statement about the definite integral $\int_{-\infty}^{\infty} e^{-\frac{1}{2}(2x-3)^2} dx$ is 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 $\mathrm{f(x)} = e^{-\frac{1}{2}(2x-3)^2}$, $\mathrm{u} = 2\mathrm{x} - 3$, so $\mathrm{du} = 2\mathrm{dx}$. We can then have $\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{+\infty} \frac{1}{2} \mathrm{e}^{-\frac{u^2}{2}} du = \frac{1}{2}\sqrt{2\pi} = 1.2533$

22. (1 or multiple correct choice(s)) Which of the following operations 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).

23. (1 or multiple correct choice(s)) Given the following data observations: 6, 3, 2, 4, 9, 1, 7, 6, which of the following is 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]).

24. (True or False) The research of big data focuses on the challenging problems of data analytics in large volume.

False. In addition to data volume, it also concerns data in velocity, variety, and veracity. P50, Lecture 1.

25. (True or False) Naive Bayes classifier is one of the most effective methods in text classification, which usually exhibits high accuracy.

False. Naive Bayes ignores the effects of word orders and assumes that features (e.g., words) are independent of each other. These assumptions are often not tenable in practical applications. P34, Lecture 2.

26. (True or False) In a hypothesis test, we reject a null hypothesis (H0) at the 5% level of significance, then we will for sure reject H0 at the 10% level of significance.

True. We reject H0 at 5%, meaning that the p-value of H0 should be smaller than 5%. Then the p-value of H0 is smaller than 10% and we will reject it at the 10% level of significance. P33-34 Lecture 3.

27. (True or False) In linear algebra, a vector is a list of numbers without orders.

False. A vector is an ordered list of numbers. P6 Lecture 4.

28. (True or False) The gradient descent algorithm always converges to the global minimum of the loss function.

False. It may converge to a local minimum (the valley) if the function has multiple valleys (non-convex). P23-24, Lecture 5.

29. (True or False) In most supervised machine learning, the training process is to maximize the decision function, which predicts the labels (y) for any data input (x).

False. The training process is to minimize the loss function, which measures the distance between the predicted labels (y) and their ground truth annotations (y*). P35, Lecture 5.

30. (True or False) In the application of a Naïve Bayes classifier, when it meets the words absent in the training data or a priorly given vocabulary, it is safe to let the classifier simply ignore these words.

True. We'll ignore them because they are unknown words not used for training and knowing which class exhibits more unknown words is generally not a useful thing to know. P43, Lecture 3.

31. (True or False) The clustering results of K-means are very sensitive to how we initialize the clusters.

True. There is no guarantee to minimize the clustering objective of K-means. It simply goes down in each step and the initialization (how we start) is crucial to the clusters we may obtain at the end (how we end). P28, Lecture 4.

32. (1 correct choice only) Bag A contains 4 white and 6 red balls, and bag B contains 6 white and 8 red balls. We randomly select a bag with equal chances and draw a ball from it, which is found to be red. What is the probability that it was drawn from the bag A.

A. 5/12
B. 3/7
C. 20/41
D. 21/41

(D) Let E=the drawn ball is red, F=the drawn ball is from bag A.

$$P(F|E) = \frac{P(E|F)P(F)}{P(E)} = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|\bar{F})P(\bar{F})} = \frac{\frac{6}{10}}{\frac{6}{10} + \frac{8}{14}} = \frac{21}{41}$$

33. (1 correct choice only) Suppose we know the probability of event A conditioned on event B is 0.5 (p(A|B)=0.5), and the probability that event B happens is 0.8. The probability of A and B happening together is:

A. 0.3
B. 0.4
C. 0.5
D. 0.8

(B) P(A,B)=P(A|B)*P(B)=0.8*0.5=0.4

34. (1 correct choice only) Given two vectors a=(1.2 3.3 5.1 2.2) and b=(0.2 1.3 1.1 0.2), the Euclidean distance of a and b is:

A. 3
B. 4
C. 5
D. 6

(C) sqrt(1^2+2^2+4^2+2^2)=sqrt(1+4+16+4)=5

35. (1 correct choice only) Dr. Ling submitted two papers A and B to a conference with an acceptance rate of 25%. On the date of acceptance notification, she received two emails about the results of A and B, respectively. She read the first email and was excited to know that A was accepted to appear at the conference. Conditioned on what she observed so far, what is the probability Ling got both A and B accepted.

A. 1/16
B. 1/4
C. 1/2
D. 1

(B) Assume that event A means paper A accepted while event B means paper B accepts. Then P(AB|A)=0.25*0.25/0.25=0.25

36. (1 correct choice only) Given a function f(x)=K (for any x), where K is a constant. The derivative for f(x) is:

A. K

B. 1
C. 0
D. x

(C) P12, Lecture 5.

37. (1 correct choice only) Given the function $f(x) = x^3 \cdot e^{(x^4+2)}$ and we want to calculate its indefinite integral with the chain rule. Which of the following is a good alternative to construct the composite function $f(x) = f(g(x))$?
A. $g(x) = x^4 + 2$
B. $g(x) = x^3$
C. $g(x) = e^{x^4+2}$
D. $g(x) = e^x$

(A) If $g(x) = x^4 + 2$, then we can have $dg(x) = 4x^3 dx$. Then $f(x) = \frac{1}{4} e^{g(x)} dg(x)$, which allows easy integration with the exponential rule.

38. (1 correct choice only) If x and y are both word count vectors derived from two sentences, which of the following describes the most precise range of the angle between them?
A. $[0, \frac{\pi}{2}]$
B. $[0, \frac{\pi}{4}]$
C. $[0, \pi]$
D. $[0, 2\pi]$

(A) Because both x and y are vectors where all entries are non-negative, their cosine similarity will be in the range of [0,1]. So the angle between them should be in the range of [0,π/2].

39. (1 or multiple correct choice(s)) Naïve Bayes is a(n) ____ classifier.
A. discriminative
B. generative
C. linear
D. non-linear

(BC) It is a generative classifier because it builds the model for each class (measured with the posterior P(c|d) P30, Lecture 2). It is a linear classifier because the model just maximizes the sum of weights (P38, Lecture 2).

40. (1 or multiple correct choice(s)) Which of the following is a factor allowing data analytics to become popular in the last decade.
A. Better models.
B. More power machines.
C. The availability of large-scale data.

(ABC) P47 Lecture 1.

41. (1 or multiple correct choice(s)) Given a discrete random variable X, find the correct statement(s):
A. E(aX) = aE(X)
B. E(aX+b) = aE(X)+b
C. Var(X) = E((X-E(X))^2)
D. Var(X) = E(X^2) – E(X)^2

(ABCD) Page 5 and 7, Lecture 3.

42. (1 or multiple correct choice(s)) Which of the following are supervised machine learning algorithms?
    A. Linear Regression
    B. Logistic Regression
    C. Naïve Bayes
    D. K-means

    (ABC) The algorithms in ABC are all supervised learning methods, which aim to learn the map between data (x) and labels (y) (P30, Lecture 2). K-means is unsupervised learning, where only the data is given without labels (P27, Lecture 4).

43. (1 or multiple correct choice(s)) Which of the following statements must be wrong for any given events A and B?
    A. P(B|A)<P(AB)
    B. P(B)=P(B|A)
    C. P(AB)=P(A)P(B)
    D. P(A|A)=0

    (AD) For A, P(B|A)=P(AB)/P(A) >= P(AB). B and C might be correct if the two events are independent. For D, P(A|A)=P(A)/P(A)=1.

44. (1 or multiple correct choice(s)) Given two vectors x, y and a scalar a, find the correct statement(s) in the following:
    A. ||ax|| = |a| ||x||
    B. ||x+ y|| =||x|| + ||y||
    C. ||x|| = 0 only If x = 0
    D. It is possible for ||x||<0.

(AC) P18, Lecture 4.

45. (1 or multiple correct choice(s)) Given a function $f(x,y,z) = -\frac{1}{\sqrt{x^2+y^2+z^2+xyz}}$, which of the following is an entry in its gradient.
    A. $-\frac{1}{2}(yz+2x)(x^2+y^2+z^2+xyz)^{-\frac{3}{2}}$
    B. $\frac{1}{2}(xy+2z)(x^2+y^2+z^2+xyz)^{-\frac{3}{2}}$
    C. $-(x^2+y^2+z^2+xyz)^{-\frac{1}{2}}$
    D. $-\frac{1}{2}(xz+2y)(x^2+y^2+z^2+xyz)^{-\frac{1}{2}}$

    (B) The three entries of the gradient are:

$$\frac{\partial f(x,y,z)}{x} = \frac{1}{2}(yz+2x)(x^2+y^2+z^2+xyz)^{-\frac{3}{2}}$$

$$\frac{\partial f(x,y,z)}{y} = \frac{1}{2}(xz+2y)(x^2+y^2+z^2+xyz)^{-\frac{3}{2}}$$

$$\frac{\partial f(x,y,z)}{z} = \frac{1}{2}(xy+2z)(x^2+y^2+z^2+xyz)^{-\frac{3}{2}}$$

46. (1 or multiple correct choice(s)) Prof. K was concerned that over 10% of people in HK caught COVID-19 (the null hypothesis H0). So, he invited 400 people in HK to do a COVID-19 test, where the results from 38 of them were positive. Suppose that the accuracy of this COVID-19 test is

100% and it is known that the infection rate of COVID-19 satisfies the normal distribution with the standard deviation of 0.1. Then, we will _____.

A. reject H0 at the significance level of 10%

B. reject H0 at the significance level of 5%

C. accept H0 at the significance level of 10%

D. accept H0 at the significance level of 5%

(CD) Suppose the infection rate at the sample test $\overline{X} = \frac{38}{400} = 0.095$ and the infection rate at the population satisfies $N(\mu, \sigma^2)$, where $\sigma = 0.1$. The p-value is $P(\overline{X} \leq 0.095) = P\left(\frac{\overline{X}-\mu}{\sigma} \leq \frac{0.095-0.1}{\frac{0.1}{\sqrt{400}}}\right) = \phi(-1) = 0.1587 > 0.1 > 0.05.$

47. (True or False) The matrix in R programming can be understood as a two-dimensional array. Each element must have the same data type and be created using the command *matrix*.

True. P34-36, Lecture 6.

48. (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.

49. (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.

50. (True or False) Given the following R code,

```r
patientID<-c(1,2,3,4);
age<-c(25,34,28,52);
diabetes<-c("Type1","Type2","Type1","Type1");
status<-c("Poor","Improved","Excellent","Poor");
patientdata<-data.frame(patientID,age,diabetes,status);
```

The command of `patientdata[1:2][2,2]` queries the age of the patient with the ID 2.

True. The system will return the second row (corresponding to patient ID 2) and second column (corresponding to the age attribute) of the `patientdata' dataframe. P37-38, Lecture 6.

51. (1 correct choice only) If the running result of the following R code is 65535, the n value at line `` x.n(x=2,n=?)'' should be _____.

```r
x.n <- function(x,n){
  h <- 0
  for(i in 0:n){
    h <- h+x^i
  }
  return(h)
}
x.n(x=2, n=?)
```

A. 13

B. 14

C. 15

D.16

(C) (x.n=1+2+4+..+2^{n}=2^{n+1}-1=65535\\n=$\log_2$65536 -1=15)

52. (1 correct choice only) Which result does the following code describe? (A)

```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.

53. (True or False) The knowledge of mathematics and computer science would allow the design of data analytical methods to tackle all the tasks.

False. While knomathematics and computer science knowledge essential for designing data analytical methods, it does not guarantee that these methods can tackle all tasks. Some problems may be too complex, require domain-specific expertise, or involve data limitations that make it challenging to develop a universal solution.

54. (True or False) The availability of effective computational methods, good hardware support, and data with higher quantity together result in the popularity of data science these years.

True. The availability of effective computational methods, good hardware support, and an increasing quantity of data have contributed significantly to the popularity of data science in recent years. Advances in algorithms, the development of powerful hardware like GPUs, and the exponential growth of data generated by various sources have enabled data scientists to uncover insights and create value across numerous industries and applications.

55. (True or False) In probability, an event covers all the possible outcomes of an experiment.

False. In probability, an event is a subset of possible outcomes from an experiment, not all of them. The collection of all possible outcomes is called the sample space. An event could consist of a single outcome or multiple outcomes, depending on the context of the experiment.

56. (True or False) If E and F are two events and P(E|F)=P(E), then we can draw a conclusion that E and F are independent.

True. If E and F are two events and P(E|F) = P(E), then it means that the probability of event E occurring is not affected by the occurrence of event F. This is the definition of independence between two events. Therefore, we can conclude that E and F are independent.

57. (True or False) In a hypothesis testing, if the p-value of a null hypothesis is given, then a smaller level of significance will result in a higher chance to accept the null hypothesis.

True. In hypothesis testing, the p-value represents the probability of obtaining a test statistic at least as extreme as the one observed, assuming the null hypothesis is true. The level of significance is the threshold below which we reject the null hypothesis. If the level of significance is smaller, it requires stronger evidence (smaller p-value) to reject the null hypothesis. Therefore, a smaller significance level will result in a higher chance of accepting the null hypothesis, assuming the p-value remains the same.

58. (True or False) In K-means, no matter how we initialize the clusters, we will always observe the same clustering results at the end when the algorithm converges.

False. The initialization of cluster centroids in the K-means algorithm can significantly affect the final clustering results. If the initial centroids are chosen differently, the algorithm may converge to different local optima. This is because K-means is sensitive to the initial placement of centroids and can get stuck in local optima depending on the starting point. Multiple runs with different initializations are often used to address this issue, and the best result (e.g., the one with the lowest within-cluster sum of squares) is chosen.

59. (True or False) Logistic regression is a popular linear regression model.

False. Logistic regression is not a linear regression model; it is a popular classification algorithm used for binary classification tasks. While both linear regression and logistic regression are based on linear

functions, the key difference lies in the response variable. Linear regression predicts continuous values, whereas logistic regression predicts the probability of a binary outcome. In logistic regression, the linear function is transformed using the logistic function (sigmoid function) to generate probabilities that can be interpreted as class membership probabilities.

60. (True or False) In R programing, it is usually more efficient to implement a function from scratch compared to the use of built-in functions.

False. In R programming, built-in functions are usually more efficient than implementing a function from scratch. Built-in functions are optimized and often written in lower-level languages, like C or Fortran, to improve performance. While it is possible to write custom functions in R, using built-in functions is generally recommended for better performance and code readability.

61. (True or False) The implementation of Monte-Carlo simulation is based on the real random numbers generated by the computer systems.

False. The implementation of Monte-Carlo simulation is based on pseudo-random numbers generated by computer systems, not real random numbers. Pseudo-random numbers are generated using deterministic algorithms that produce number sequences that appear to be random. Although these numbers are not truly random, they exhibit sufficient randomness for most practical purposes, including Monte-Carlo simulations. True random number generation would require a non-deterministic process, such as measuring physical phenomena like radioactive decay or atmospheric noise, which is not feasible for most computer systems.

62. (True or False) Linear regression refers to the model which employs the line in two or higher dimensional space to fit the data.

Linear regression does use a linear function to fit the data. However, the term "line" might not be accurate for all cases, especially when dealing with multiple predictor variables. In such cases, the model can be represented by a plane (for two predictor variables) or a hyperplane (for more than two predictor variables) in higher-dimensional spaces. The primary idea is that the relationship between the predictor and response variables follows a linear pattern.

63. (1 correct choice only) Which of the following statement is true.
    A. If we flip a coin for 1000 times and observe 500 heads, then the probability of observing head in a random coin flipping is 0.5.
    B. Monte Carlo simulation will never possible to guarantee perfect accuracy in practice.
    C. In Monte Carlo simulation, if the sample size is large, then we can for sure be confident about the estimation results.
    D. We usually employ Poisson distribution to model the waiting time between two events, where the events occur independently with a known average rate of the time since the last event.

The correct answer is B. Monte Carlo simulation is a technique that relies on random sampling to approximate a solution to a problem. While increasing the number of samples can improve the accuracy of the simulation, it can never guarantee perfect accuracy due to the inherent randomness involved. However, Monte Carlo simulations can often provide reasonably accurate estimates for many practical applications.

64. (1 correct choice only) Which of the following correctly describes the range of the return value (output) of the R function “rnorm(1)”?
    A. (0,1)

B. (-1,1)
C. (-0.5,0.5)
D. (-∞,+∞)

The correct answer is D. The rnorm(1) function in R generates a single random number from a normal distribution with a default mean of 0 and a standard deviation of 1. The range of values that can be generated from a normal distribution is theoretically from negative infinity to positive infinity.

65. (1 correct choice only) K-means algorithm is used to cluster N points into K groups (suppose that K=10). If it is needed to visualize the distribution (frequency) of points assigned to varying groups, which of the following graph can best employed to tackle the task?
    A. Barplot
    B. Histogram
    C. Scatterplot
    D. Boxplot

The correct answer is A. A barplot is the most suitable choice for visualizing the distribution of points assigned to different clusters or groups. It can clearly display the frequency of points in each group as separate bars, making it easy to compare the sizes of the groups.

66. (1 correct choice only) Which of the following is the legal form of a comment in R programing?
    A. ##This is a comment##
    B. *This is a comment*
    C. //This is a comment//
    D. **This is a comment*

The correct answer is A. In R programming, comments are denoted using the hash symbol (#). You can use a single hash (#) or double hash (##) to start a comment. The text following the hash symbol(s) on the same line is treated as a comment and not executed by the R interpreter.

67. (1 correct choice only) Suppose that you are working with your groupmate to analyze the midterm and final test scores from 3 Hogwarts students --- Harry, Ron, and Hermione via R programing. Your groupmate created a dataframe with the following code to host the midterm and final test scores of the three students. Following the common programing practice, which of the following is most likely to be the final test score of Ron.

```
student_scores <- data.frame(name = c("Harry", "Hermione", "Ron"), midterm = c(80, 100, 70), final = c(75, 95, 60))
```

    A. 60
    B. 70
    C. 75
    D. 95

The correct answer is A. To find the final test score of Ron, you can access the dataframe with the correct row and column:

```
ron_final_score <- student_scores[student_scores$name == "Ron", "final"]
```

So, the final test score for Ron is 60.

68. (1 correct choice only) While running the K-means algorithm, if a cluster has 4 points (1,3,4), (2,5,6), (3,1,2), (6,3,8), then the Euclidian norm (length) of the cluster centroid (representative) is ____.

A. √43
B. 3√3
C. 2√43
D. 3

The correct answer is A. To find the centroid of the cluster, we first need to calculate the mean of the points' coordinates: (1,3,4), (2,5,6), (3,1,2), (6,3,8)

Mean of x coordinates: (1+2+3+6)/4 = 12/4 = 3

Mean of y coordinates: (3+5+1+3)/4 = 12/4 = 3

Mean of z coordinates: (4+6+2+8)/4 = 20/4 = 5

So, the centroid is (3, 3, 5). Now, we need to calculate the Euclidean norm (length) of this centroid: Euclidean norm = √((x-0)^2 + (y-0)^2 + (z-0)^2) = √((3-0)^2 + (3-0)^2 + (5-0)^2) = √(9+9+25) = √43

69. (1 correct choice only) Suppose that two twin girls in a park are playing Rock Paper Scissors (https://en.wikipedia.org/wiki/Rock_paper_scissors) and the three genres --- rock, paper, and scissors --- has equal chance (1/3) to be given by one girl in a round. One passerby told you that at least one girl gave the rock in a specific round R*, while it is unknown which girl gave the rock in R* (because the twins look so much alike). Conditioned on what the passerby observed, what is the probability that round R* is a draw (the two girls gave the same genre)?

A. 1/3
B. ¼
C. 1/5
D. 1/6

The correct answer is C. Given that at least one girl gave the rock in round R*, there are three possible outcomes:

- Girl 1 plays rock, and Girl 2 plays rock (RR).
- Girl 1 plays rock, and Girl 2 plays something other than rock (RP or RS).
- Girl 1 plays something other than rock, and Girl 2 plays rock (PR or SR).

We want to find the probability that the round is a draw (both girls gave the same genre) given that at least one girl played rock. The only draw scenario in these cases is the first one (RR). Now, let's calculate the probability of each outcome:

- P(RR) = P(Girl 1 plays rock) * P(Girl 2 plays rock) = (1/3) * (1/3) = 1/9
- P(RP or RS) = P(Girl 1 plays rock) * P(Girl 2 plays not rock) = (1/3) * (2/3) = 2/9
- P(PR or SR) = P(Girl 1 plays not rock) * P(Girl 2 plays rock) = (2/3) * (1/3) = 2/9

The probability of at least one girl playing rock is the sum of these probabilities:

P(at least one rock) = P(RR) + P(RP or RS) + P(PR or SR) = 1/9 + 2/9 + 2/9 = 5/9

Now, we can find the conditional probability of a draw given at least one girl played rock: P(draw | at least one rock) = P(RR) / P(at least one rock) = (1/9) / (5/9) = 1/5

70. (1 or multiple correct choice(s)) Which of the following is (are) the assumption(s) made by multinomial Naïve Bayes.
    A. The order of the words in a document are not important.
    B. Word occurrences are independent with each other conditioned on the class of the document.
    C. There are only two classes involved in the classification task.
    D. The computing environments have been well set up to avoid the floating-point underflow problem.

The correct answer is AB. The assumptions made by the multinomial Naïve Bayes classifier are:

A. The order of the words in a document are not important. This assumption is true because Naïve Bayes models treat documents as "bags of words" without considering the order of the words.
B. Word occurrences are independent of each other, conditioned on the class of the document. This assumption is true because Naïve Bayes classifiers assume that features (in this case, word occurrences) are conditionally independent given the class.
C. There are only two classes involved in the classification task. This assumption is false because the multinomial Naïve Bayes classifier can handle multiple classes, not just two.
D. The computing environments have been well set up to avoid the floating-point underflow problem. This assumption is not directly related to the multinomial Naïve Bayes model itself. Floating-point underflow is a numerical issue that can be mitigated through techniques like log-probabilities, but it's not an assumption made by the model.

71. (1 or multiple correct choice(s)) Which of the following statement about linear regression is (are) correct?
    A. We assume the data to exhibits heteroscedasticity to allow the data fitting in a line.
    B. Given a random dataset, parabola allows better generalization capability (i.e., to better fit new data) for the data fitting compared to a line because parabola is a polynomial with the higher degree.
    C. In the evaluation of linear regression, the goodness of fit (R^2) takes both the task complexity and the errors in the fit into consideration.
    D. A method of linear regression might be a good tool in the application of time series analysis.

The correct answer is CD.

A. We assume the data to exhibits heteroscedasticity to allow the data fitting in a line. This statement is incorrect. In linear regression, we assume homoscedasticity, which means that the variance of the errors is constant across all levels of the independent variable(s).
B. Given a random dataset, parabola allows better generalization capability (i.e., to better fit new data) for the data fitting compared to a line because parabola is a polynomial with a higher degree. This statement is not necessarily true. While a parabola (a quadratic function) has more flexibility than a straight line, this does not automatically mean that it will have better generalization capability. Overfitting can occur when using a higher-degree polynomial, which may result in poor performance on new, unseen data.
C. In the evaluation of linear regression, the goodness of fit (R^2) takes both the task complexity and the errors in the fit into consideration. This statement is correct. The R^2 value, also known as the coefficient of determination, measures how well the regression model fits the observed data. It considers both the complexity of the model (through the number of independent variables) and the errors in the fit (through the residual sum of squares).

D. A method of linear regression might be a good tool in the application of time series analysis. This statement is correct. Linear regression can be used in time series analysis, particularly when there is a linear trend present in the data. However, it may not be the best choice for time series with complex patterns or seasonality, where other methods like ARIMA or exponential smoothing might be more appropriate.

72. (1 or multiple correct choice(s)) Suppose that the data is represented by vectors, which of the following allow(s) the calculation of data similarity and/or dissimilarity?
    A. Euclidian distance
    B. Euclidian norm
    C. Angles of vectors
    D. Vector addition

The correct answer is AC.

A. Euclidian distance: This is correct. Euclidean distance is a common measure used to calculate the dissimilarity between two data points represented by vectors. The greater the distance, the less similar the data points are.
B. Euclidian norm: This is incorrect. The Euclidean norm, also known as the magnitude or length of a vector, is used to determine the size of a vector but does not directly compare the similarity or dissimilarity between two vectors.
C. Angles of vectors: This is correct. The angle between two vectors can be used to measure their similarity. The cosine similarity, which is based on the cosine of the angle between two vectors, is a popular measure for calculating the similarity between vectors.
D. Vector addition: This is incorrect. Vector addition is an operation that combines two vectors to produce a third vector. It does not directly measure similarity or dissimilarity between data points.

73. (1 or multiple correct choice(s)) Which of the following about machine learning is (are) correct?
    A. K-means and Naïve Bayes are both machine learning methods.
    B. Sigmoid function is the loss function of logistic regression, which will be optimized during the training process.
    C. Chain rule is usually used to the calculate gradient in the model training.
    D. Parameters of the machine learning models should not be randomly initialized because we will have the difficulty to reproduce the results.

The correct answer is AC.

A. K-means and Naïve Bayes are both machine learning methods. This is correct. K-means is an unsupervised machine learning method used for clustering, while Naïve Bayes is a supervised machine learning method used for classification.
B. Sigmoid function is the loss function of logistic regression, which will be optimized during the training process. This is incorrect. The sigmoid function is the activation function used in logistic regression to convert the linear output to a probability. The loss function that is optimized during the training process is typically the log loss (also called cross-entropy loss).
C. Chain rule is usually used to calculate the gradient in the model training. This is correct. The chain rule is a fundamental concept in calculus and is often used in machine learning to compute gradients, especially in the backpropagation algorithm for training neural networks.
D. Parameters of the machine learning models should not be randomly initialized because we will have the difficulty to reproduce the results. This is incorrect. In many machine learning algorithms, especially neural networks, it is common to initialize parameters randomly to

break symmetry and ensure that the model learns diverse features. To ensure reproducibility, you can set a random seed, which will ensure that the same set of random numbers is generated each time the algorithm is run.

74. (1 or multiple correct choice(s)) In a hypothesis testing, suppose that it is known we reject the null hypothesis at the level of significance 10%. Which of the following is(are) for sure to be incorrect?
    A. We may accept the null hypothesis at the level of significance 1%.
    B. We can infer that the p-value of the null hypothesis is larger than 0.1.
    C. The p-value of the null hypothesis might be 0.09.
    D. We might also reject the null hypothesis at the level of significance 5%.

The correct answer includes only option B.

A. We may accept the null hypothesis at the level of significance 1%. This statement is correct. If we reject the null hypothesis at the 10% level of significance, the p-value is less than or equal to 0.1. However, it is still possible that the p-value is larger than 0.01, which means we may not reject the null hypothesis at the 1% level of significance.

B. We can infer that the p-value of the null hypothesis is larger than 0.1. This statement is incorrect. If we reject the null hypothesis at the 10% level of significance, it means the p-value is less than or equal to 0.1, not larger than 0.1.

C. The p-value of the null hypothesis might be 0.09. This statement is correct. If we reject the null hypothesis at the 10% level of significance, it means the p-value is less than or equal to 0.1. A p-value of 0.09 would satisfy this condition.

D. We might also reject the null hypothesis at the level of significance 5%. This statement is correct. If we reject the null hypothesis at the 10% level of significance, the p-value might be less than or equal to 0.1. If the p-value is also less than or equal to 0.05, we would reject the null hypothesis at the 5% level of significance as well.