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) 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.

2. (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. P38 and P45, Lecture 2.

3. (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.

4. (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.

5. (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.

6. (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.

7. (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.

8. (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.

9. (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. P46, Lecture 2.

10. (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.

11. (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.

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

12.(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

13. (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) $\sqrt{1^2 + 2^2 + 4^2 + 2^2} = \sqrt{1 + 4 + 16 + 4} = 5$

14. (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

15. (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.

16. (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 $dg(x) = 4x^3dx$. Then $f(x) = \frac{1}{4}e^{g(x)}dg(x)$, which allows easy integration with the exponential rule.

17. (1 correct choice only) In R programming, which command can convert an image to a PDF file?
    A. export
    B. pdf
    C. output
    D. print

    (B)

18. (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=?)


for(i in 0:n){
    print(i)
  }
```

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)

19. (1 correct choice only) We applied Naïve Bayes to classify some tweets into positive, neutral, and negative classes. Then, we can use the ____ graph in ggplot2 R package to visualize the distribution of tweet number over the three classes.

A. Barplot

B. Histogram

C. Scatterplot

D. None of the above

(A) Barplot would be a good alternative because it is the frequency over discrete classes (x-axis indicates discrete categories).

20. (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].

21.(1 or multiple correct choice(s)) Which of the following are NOT allowed to name an object in the R system.

A. #abc
B. .abc
C. 0abc
D. Abc

(AC) P 17, Lecture 6.

22. (1 or multiple correct choice(s)) Naïve Bayes is a(n) ____ classifier.

A. discrete
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) P42, Lecture 2). It is a linear classifier because the model just maximizes the sum of weights (P49, Lecture 2).

23. (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.

24. (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.

25. (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) (P36, Lecture 2). K-means is unsupervised learning, where only the data is given without labels (P39, Lecture 3).

26. (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]

    (ABC) P18, Lecture 6.

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

28. (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.

29. (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:

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

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

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

30. (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 $P(\overline{X} \leq 0.095) = P\left(\frac{\overline{X}-\mu}{\sigma} \leq \frac{0.095-0.1}{\frac{0.1}{\sqrt{400}}}\right) = \phi(-1) = 0.1587 > 0.1 > 0.05$.