COMP1003/1433 Introduction to Data Analytics Midterm Test (Solution)

1. (True/false) Gradient is a vector that points in the direction of the steepest increase of a function.

True. The gradient is a vector that points in the direction of the steepest increase of a function. It is a mathematical concept used in optimization algorithms such as gradient descent, where it is used to find the direction of the largest increase in the cost function. The gradient is also used to find the direction of the steepest decrease, which is the opposite direction of the gradient. (Page 21, Lecture 5).

2. (True/false) Data mining is a subset of data analytics that uses algorithms to extract valuable information from large datasets.

True. Data mining is a subset of data analytics involving various techniques and algorithms to extract useful insights and knowledge from large datasets. These techniques include statistical analysis, machine learning, artificial intelligence, and pattern recognition. Data mining aims to uncover hidden patterns, trends, and relationships in the data that can be used to make better decisions or predictions. (Page 55-57, Lecture 1)

3. (True/false) If two events A and B are independent, then the probability of A given B has occurred is equal to the probability of A.

True. If two events A and B are independent, then the probability of A given B has occurred is equal to the probability of A. This is the definition of independence: the occurrence of one event does not affect the probability of the other. Therefore, the statement is true. (Page 23, Lecture 2)

4. (True/false) A probabilistic language model assigns a probability to a sequence of words in a language.

True. A probabilistic language model assigns a probability to a sequence of words in a language. It is a statistical model that uses probability theory to predict the likelihood of a sequence of words based on the probability distribution of the language. This model is used in many natural language processing tasks, such as speech recognition, machine translation, and text generation. (Page 42, Lecture 2)

5. (True/false) The p-value is the probability that the null hypothesis is true.

False. The p-value is the probability of obtaining a test statistic as extreme as or more extreme than the observed result, assuming that the null hypothesis is true. It is not the probability that the null hypothesis itself is true. (Page 31-32, Lecture 3)

6. (True/false) Descriptive statistics can be generalized beyond the data sample, whereas inferential statistics only describe the data you have.

False. Descriptive statistics only describe the data you have and cannot be generalized beyond that data set. For example, if you calculate the median of a set of values, that median is only representative of that particular set of values.

In contrast, inferential statistics enable you to make inferences about the population beyond the data sample you have. For example, if you perform hypothesis testing on a sample of data, you can use the results to make inferences about the population from which the sample was drawn.

Both descriptive and inferential statistics are important for many Machine Learning techniques as they provide insights into the data and enable modeling and prediction.

(Page 34, Lecture 3)

7. (True/false) The sample variance is a measure of how spread out the data is around the sample mean.

True. The sample variance is a measure of how spread out the data is around the sample mean. It is calculated by taking the sum of the squared differences between each sample value and the sample mean, dividing by the sample size minus 1, and taking the square root of the result. The larger the variance, the more spread out the data is. (Page 11, Lecture 3)

8. (True/false) The cosine similarity of two vectors can be negative.

True. The cosine similarity of two vectors is a measure of the similarity between the vectors, which ranges from -1 to 1. A value of -1 indicates that the two vectors are perfectly dissimilar, 0 indicates that they are orthogonal (i.e., have no similarity), and 1 indicates that they are perfectly similar. A cosine similarity of less than 0 indicates that the angle between the vectors is greater than 90 degrees, which means that they are pointing in opposite directions and are dissimilar. (Page 24, Lecture 4)

9. (True/false) The decision function and the loss function are two distinct components of a machine learning model, with the decision function used to make predictions and the loss function used to measure the accuracy of those predictions.

True. The decision function and the loss function are two distinct components of a machine learning model. The decision function is a mathematical function that maps inputs to outputs, which in the context of supervised learning, is used to make predictions on unseen data. The loss function, on the other hand, is a measure of how well the model is performing in terms of prediction accuracy. It quantifies the difference between the predicted output and the true output for a given input.

The loss function is typically used in the training phase to update the model parameters such that the prediction accuracy improves over time. The decision function, on the other hand, is used in the testing phase to make predictions on unseen data. (Page 5, Lecture 5).

10. (True/false) Logistic regression is a generative classifier.

False. Logistic regression is a discriminative classifier, not a generative classifier.

Generative classifiers learn the joint probability distribution of the input features and the target class, and use this knowledge to make predictions about new data. Examples of generative classifiers include Naive Bayes and Gaussian Mixture Models.

On the other hand, discriminative classifiers learn the decision boundary between different classes directly. Logistic regression is an example of a discriminative classifier. It models the conditional probability of the target class given the input features directly, without modeling the joint probability distribution. (Page 29-34, Lecture 5)

11. (Multi-choice) Which of the following scenarios is an example of the application of conditional probability?
    A) Rolling a dice and getting a number less than 6.
    B) Flipping a coin and getting heads or tails.
    C) Drawing a card from a deck and getting a heart.
    D) Drawing a card from a deck and getting an ace given that it is a face card.

The correct answer is D. This scenario is an example of conditional probability because it involves the probability of one event (getting an ace) given that another event (getting a face card) has already occurred. The other scenarios are examples of simple or unconditional probability because they do not depend on any other event. (Page 14, Lecture 2)

12. (Multi-choice) Which of the following best describes a variable in data analytical processes?
    A) A scientific tool used for data collection.
    B) A unit of measurement for time intervals .
    C) A single quality or quantity of some object or phenomenon
    D) A term used to describe financial transactions.

The answer to the multichoice question is C) A single quality or quantity of some object or phenomenon.

In data analytical processes, a variable is a characteristic or attribute that can take on different values or categories within a dataset. It represents a single quality or quantity of an object or phenomenon that is being studied. Variables can be numerical, such as age or income, or categorical, such as gender or marital status. They are essential in data analysis as they provide a way to organize, summarize, and compare different aspects of the data. Therefore, option C is the correct answer. Options A, B, and D do not accurately describe what a variable is in the context of data analytics. (Page 26, Lecture 1)

Option A describes a scientific tool used for data collection, but it does not accurately describe what a variable is in data analytical processes. Option B describes a unit of measurement for time intervals, but this is not the same as a variable. Option D describes a term used to describe financial transactions, which is not related to what a variable is in data analytical processes. Therefore, by process of elimination, the correct answer is C, which accurately describes a variable as a single quality or quantity of some object or phenomenon in data analytical processes.

13. (Multi-choice) A student is taking a multiple-choice test. Each question on the test has four possible answers, and only one choice is correct. The probability that this student knows the correct answer is 80%. If he does not know the answer, he guesses at random. What is the probability that he answers a question correctly?
    A) 0.2
    B) 0.25

C) 0.8
D) 0.85

The correct answer is D. The probability that the student answers a question correctly depends on whether they know the answer or not. If they know the answer, the probability is 80%. If they do not know the answer, they have a 1/4 chance of guessing correctly. Since the probability that they know the answer is 80%, the overall probability that they answer a question correctly is:

$$0.8 * 1 + 0.2 * 1/4 = 0.8 + 0.05 = 0.85$$

Law of Total Probability (Page 17, Lecture 2)

14. (Multi-choice) Suppose a clinic offers genetic testing for a certain disease, which affects 1 in 1000 people in the general population. The test is 99% accurate for both positive and negative results. If a patient tests positive for the disease, what is the probability that they have the disease?
    A) 99%
    B) 90%
    C) 9.1%
    D) 0.1%

Answer: C) 9.1%

Explanation: Let A be the event that a patient has the disease, and B be the event that a patient tests positive for the disease. Using Bayes' formula, we have:

P(A|B) = P(B|A) x P(A) / P(B)

where P(A) = 1/1000, P(B|A) = 0.99, and P(B|not A) = 0.01 (since the test is 99% accurate for negative results). We can then calculate P(B) using the law of total probability:

P(B) = P(B|A) x P(A) + P(B|not A) x P(not A) = 0.99 x 1/1000 + 0.01 x 999/1000 = 0.01098

Substituting these values into Bayes' formula, we get:

P(A|B) = 0.99 x 1/1000 / 0.01098 = 0.0902 = 9.1% (rounded to one decimal place)

Therefore, the probability that a patient actually has the disease given a positive test result is 9.1%. Answer choice C is correct. (Page 20, Lecture 1)

15. (Multi-choice) Which of the following is the correct indefinite integral of the function $f(x) = 3x^2 + 2x + 1$?
    A) $\int (3x^2 + 2x + 1)\,dx = x^3 + x^2 + x + C$
    B) $\int (3x^2 + 2x + 1)\,dx = x^3 + x^2 + C$
    C) $\int (3x^2 + 2x + 1)\,dx = x^4 + x^2 + x + C$
    D) $\int (3x^2 + 2x + 1)\,dx = 3x^3 + x^2 + x + C$

The correct answer is A) $\int (3x^2 + 2x + 1)dx = x^3 + x^2 + x + C$. The integral of each term of the given function is:

- $\int 3x^2 dx = x^3 + C1$
- $\int 2x\, dx = x^2 + C2$
- $\int 1\, dx = x + C3$

where C1, C2, and C3 are arbitrary constants of integration. Therefore, the indefinite integral of the function $f(x) = 3x^2 + 2x + 1$ is:

$$\int (3x^2 + 2x + 1)dx = \int 3x^2 dx + \int 2x\, dx + \int 1\, dx = x^3 + x^2 + x + C$$

where $C = C1 + C2 + C3$ is an arbitrary constant of integration. So the correct answer is A) $\int (3x^2 + 2x + 1)dx = x^3 + x^2 + x + C$. (Page 42, Lecture 5).

16. (Multi-choice) Suppose $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and $B = \begin{pmatrix} 2 & 1 \\ 4 & 3 \end{pmatrix}$, what is the transpose of $A + B$, i.e., $(A + B)^T$?
    A) $\begin{pmatrix} 3 & 3 \\ 7 & 7 \end{pmatrix}$
    B) $\begin{pmatrix} 10 & 7 \\ 22 & 16 \end{pmatrix}$
    C) $\begin{pmatrix} 10 & 22 \\ 7 & 16 \end{pmatrix}$
    D) $\begin{pmatrix} 3 & 7 \\ 3 & 7 \end{pmatrix}$

The correct answer is D. $A + B = \begin{pmatrix} 1+2 & 2+1 \\ 3+4 & 4+3 \end{pmatrix} = \begin{pmatrix} 3 & 3 \\ 7 & 7 \end{pmatrix}$. So, its transpose is $\begin{pmatrix} 3 & 7 \\ 3 & 7 \end{pmatrix}$. (Page 54-55, Lecture 4).

17. (Multi-choice) Which of the following best defines sampling?
    A) Measuring the height of all adults in a large population.
    B) Measuring the starting salary of all CS students.
    C) Gathering random data from a large population.
    D) Recording the results of a single experiment.

The correct answer is C: Gathering random data from a large population. Sampling involves selecting a smaller subset from a larger population to gather data from. The examples provided in the assumptions, such as measuring the breaking strength of randomly selected bolts, illustrate this concept. Options A and B describe measuring data from an entire population, while option D refers to recording the results of a single experiment, which may or may not involve sampling. (Page 8, Lecture 3)

18. (Multi-choice) Which of the following represents the derivative of the function $f(x) = (3x^2 + 5x)^4$ using the chain rule?
    A) $12x^3(3x^2 + 5x)^3$
    B) $4(3x^2 + 5x)^3\ (6x + 5)$
    C) $12(3x^2 + 5x)^3(6x + 5)$
    D) $24(3x^2 + 5x)^3(6x + 5)$

The correct answer is B. To use the chain rule, we need to identify the outer function and the inner function. In this case, the outer function is the power of 4 and the inner function is the expression $g(x) = (3x\text{^}2 + 5x)$.

Let $g(x) = 3x^2 + 5x$. Then we have $f(x) = g(x)^4$. Using the chain rule, we have:

$$f'(x) = 4g(x)\text{^}3 * g'(x)$$

To find g'(x), we can differentiate the expression $3x^2 + 5x$: $g'(x) = 6x + 5$

Substituting g(x) and g'(x) into the formula for f'(x), we get:

$$f'(x) = 4(3x^2 + 5x)^3 (6x + 5)$$

(Page 17, Lecture 5)

19. (Multi-choice) Suppose u and v are two-dimensional vectors. Which of the following expressions gives the norm (length) of the vector $u + v$?

A) $|u + v|$
B) $|u| + |v|$
C) $|u| - |v|$
D) $|u|^2 + |v|^2$

Answer: A. $|u + v|$

Explanation: To find the magnitude of the vector u + v, we need to take the square root of the sum of the squares of its components. That is, $|u + v| = sqrt((u1 + v1)^2 + (u2 + v2)^2)$

Option A is the only one that represents this calculation. Option B gives the sum of the norms of the individual vectors, which is not the same thing. Option C gives the difference of the norms, which is also not what we're looking for. Option D gives the sum of the squares of the norms, which is related to the dot product of the vectors, but not directly to their sum. (Page 20, Lecture 4)

20. (Multi-choice) Suppose we have the following R code:

```
x <- c(1, 2, 3, 4, 5)
y <- x^2
```

What will be the output of y? Please select the correct answer from the choices below.

A) [1] 1 4 9 16 25
B) [1] 1 2 3 4 5
C) [1] 2 4 6 8 10
D) [1] 0.5 1 1.5 2 2.5

The correct answer to the multichoice question is: A) [1] 1 4 9 16 25

The code creates a numeric vector x with five elements, and then creates another vector y by squaring each element of x using the ^ operator. The resulting values are [1] 1 4 9 16 25. Therefore, option A is the correct answer.

Option B is incorrect because it simply lists the values of the x vector, not the y vector that was created by squaring x.

Option C is incorrect because it lists the doubled values of x, not the squared values.

Option D is incorrect because it lists the half of the squared values of x, which is not what the code does.

(Page 49, Lecture 6)

21. (Multi-answer) Which of the following are data types in R?
    A) Scalars
    B) Vectors (numerical, character, logical)
    C) Matrices
    D) Dataframes

ABCD. R has a wide variety of data types including scalars, vectors (numerical, character, logical), matrices, arrays, dataframes, and lists. (Page 31, Lecture 6)

22. (Multi-answer) Which of the following statements are true regarding K-means clustering? Select all that apply.
    A) K-means is an unsupervised learning algorithm.
    B) The number of clusters in k-means is determined by the value of K.
    C) K-means always converges to the global minimum of the objective function.
    D) The initial placement of centroids in k-means has no impact on the final clustering result.

The correct answer is AB.

A) K-means is an unsupervised learning algorithm. This means that it does not use any labeled data to train the model, but instead finds patterns and groups in the data based on the similarity between data points.

B) The number of clusters in k-means is determined by the value of K. K-means tries to partition the data into K clusters, where K is a user-specified priorly-given parameter.

C) K-means does not always converge to the global minimum. The algorithm can get stuck in local minima, which means that the clustering result may not be optimal.

D) The initial placement of centroids in k-means can have an impact on the final clustering result. The algorithm can converge to different local minima depending on the initial placement of centroids. Therefore, it is common to run K-means multiple times with different initializations to find the best clustering result.

(Page 24-31, Lecture 4)

23. (Multi-answer) What are some properties of the sigmoid function? Select all that apply.
    A) The sigmoid function is defined as $f(x) = 1 / (1 + e^{-x})$.
    B) The sigmoid function can be applied in logistic regression.
    C) The sigmoid function has a range between 0 and 1.
    D) The sigmoid function is not linear.

The correct answer is ABCD.

A) The sigmoid function is defined as f(x) = 1 / (1 + e^(-x)) because this is the mathematical formula for the sigmoid function. It takes any real number as input and outputs a value between 0 and 1.

B) The sigmoid function can be applied in logistic regression because it maps any input to a value between 0 and 1, which can be interpreted as a probability. Logistic regression is a binary classification algorithm that predicts the probability of an instance belonging to a particular class.

C) The sigmoid function has a range between 0 and 1 because of its mathematical formula. As x approaches negative infinity, e^(-x) approaches infinity, making the denominator approach infinity and the value of the sigmoid function approach 0. As x approaches positive infinity, e^(-x) approaches 0, making the denominator approach 1 and the value of the sigmoid function approach 1.

D) The sigmoid function is not linear because it does not satisfy the property of linearity, which requires f(ax) = af(x) for all real numbers a and x. The sigmoid function's output does not scale linearly with its input, and it is not a straight line on a graph.

(Page 32-34, Lecture 5)

24. (Multi-answer) Which of the following statements about the application of data analytics are true? Select all that apply.
    A) Data analytics can be used to identify patterns and trends in large datasets.
    B) Data analytics can be used to optimize business processes and improve efficiency.
    C) Data analytics can be used to make predictions and inform decision-making.
    D) Data analytics can be used in various industries, including healthcare, finance, and marketing.

The correct answer is A, B, C, and D. These statements about applying data analytics are true. Data analytics can help organizations and individuals to analyze large amounts of data, find patterns and trends, optimize processes and efficiency, make predictions, and inform decision-making, and apply to various domains and industries. (Lecture 1)

25. (Multi-answer) Which of the following factors have contributed to the huge success of data science in recent years? Select all that apply.
    A) The availability of the rule-based approaches.
    B) More powerful RAM, CPU, GPU, etc.
    C) Huge volume of data.
    D) Better models to fit the data.

The correct answer is B, C, and D. These factors have contributed to the huge success of data science in recent years. Data science relies on more powerful hardware, such as RAM, CPU, and GPU, to process large amounts of data efficiently. Data science also benefits from the availability of a huge volume of data from various sources, such as social media, sensors, transactions, etc. Data science also uses better models to fit the data, such as machine learning and deep learning algorithms that can learn from data and make predictions. (Page 47, Lecture 1)

26. (Multi-answer) Which of the following statements regarding the law of large numbers are true? Select all that apply.

A) The law of large numbers states that as the sample size increases, the sample mean approaches the population mean.
B) The law of large numbers applies only to discrete random variables.
C) The law of large numbers guarantees that individual outcomes will always be close to the mean.
D) The law of large numbers is a theoretical concept but can benefit practical applications.

The correct answer is AD.

A) is true because the law of large numbers states that as the sample size increases, the sample mean approaches the population mean. This means that if we take many random samples from a population, the average of these samples will be close to the true population mean.

B) is false because the law of large numbers applies to both discrete and continuous random variables.

C) is false because the law of large numbers does not guarantee that individual outcomes will be close to the mean, but rather that the average of a large number of outcomes will be close to the mean. As the sample size increases, the variability of the sample mean decreases, which means that the individual outcomes are more likely to be closer to the mean.

D) is true because the law of large numbers has many practical applications, such as in statistical quality control, finance, and insurance.

(Page 15, Lecture 3)

27. (Multi-answer) Suppose you have a set of data samples with the following values: 4, 5, 6, 7, 8, 9, 10. Which of the following statements is true? Select all that apply.

A) The sample mean is 6.5.
B) The sample range is 6.
C) The sample median is 7.
D) The sample mean equals to the sample median.

The correct answer is BCD.

A) The mean of the data set is calculated by adding all the values and dividing by the total number of values: (4+5+6+7+8+9+10) / 7 = 7. This statement is false.

B) The range is the difference between the largest and smallest values in the data set. In this case, the largest value is 10 and the smallest value is 4, so the range is 10-4 = 6. This statement is true.

C) The median is the middle value of a sorted data set. When the data set has an odd number of values, the median is the middle value. When the data set has an even number of values, the median is the average of the two middle values. In this case, the data set has an odd number of values, so the median is the middle value, which is 7. This statement is true, as the median is actually 7.

D) In this case, the mean is 7 and the median is 7, so the mean equals the median. This statement is true.

(Page 10, Lecture 3)

28. (Multi-answer) Which of the following statements about gradient descent are true? Select all that apply.
    A) Gradient descent is an optimization algorithm used to minimize the loss function in machine learning models.
    B) The learning rate in gradient descent controls the step size taken in each iteration towards the minimum.
    C) Gradient descent always guarantees convergence to the global minimum of the loss function.
    D) The choice of initialization of the parameters in gradient descent can affect the convergence speed and final solution.

The correct answer is ABD.

A) Gradient descent is an optimization algorithm used to minimize the cost function in machine learning models. This statement is true as gradient descent is a widely used optimization algorithm that is used to minimize the cost function in machine learning models. The cost function is typically used to measure how well the model is performing, and minimizing it is necessary to achieve the best possible performance of the model.

B) The learning rate in gradient descent controls the step size taken in each iteration towards the minimum. This statement is true as the learning rate in gradient descent controls the size of the steps taken towards the minimum of the cost function. A higher learning rate will result in larger steps, which may cause the algorithm to overshoot the minimum, while a lower learning rate may cause the algorithm to converge too slowly.

C) Gradient descent always guarantees convergence to the global minimum of the cost function. This statement is false as there may be situations where gradient descent converges to a local minimum instead of the global minimum. This happens when the cost function is non-convex, meaning it has multiple local minima.

D) The choice of initialization of the parameters in gradient descent can affect the convergence speed and final solution. This statement is true as the choice of initialization of the parameters in gradient descent can affect the convergence speed and final solution. A poor initialization can lead to slow

convergence and getting stuck in local minima. Therefore, it is essential to choose the right initialization of the parameters to optimize the performance of the model.

(Page 21, Lecture 5)

29. (Multi-answer) Which of the following statements are true about naming objects in R? Select all that apply.
    A) Object names can contain any symbol, including !, +, - and #
    B) Both dot and underscore can be used in object names.
    C) Object names cannot contain numbers at any position.
    D) R is case sensitive, so objects named temp and temP are considered different.

The correct answer is BD.

Option A is incorrect because object names in R cannot contain "strange" symbols such as !, +, -, or #. In R, object names can only contain letters, numbers, periods, and underscores. Any other symbols are not allowed and would result in an error message. Therefore, option A is not a correct statement about naming objects in R.

Option B is correct because a dot and an underscore are allowed in object names in R.

Option C is incorrect because Object names in R can contain numbers, but they cannot start with a number.

Option D is correct because R is case sensitive, so objects named temp and temP are considered different. R is a case-sensitive language, which means that it distinguishes between uppercase and lowercase letters in object names. Therefore, objects named temp and temP are considered different because they have different capitalization.

(Page 17, Lecture 6)

30. (Multi-answer) Which of the following statements are true about Naive Bayes classifiers? Select all that apply.
    A) They are based on Bayes' Formula, which allows us to invert conditional probabilities.
    B) They are models that assign class labels to problem instances, represented as vectors of feature values.
    C) Naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.
    D) They are very effective when the features strongly correlate with each other.

The correct answer is ABC.

A) Naive Bayes classifiers are based on Bayes' Formula, which allows us to invert conditional probabilities. Bayes' Formula is used to calculate the probability of a hypothesis given some observed evidence. In the case of Naive Bayes classifiers, the hypothesis is the class label, and the evidence is the feature values. The formula is used to calculate the probability of the class label given the observed feature values.

B) Naive Bayes classifiers are models that assign class labels to problem instances, represented as vectors of feature values. The input to a Naive Bayes classifier is a set of features or variables that describe the instance being classified. The classifier then assigns a probability to each possible class label and selects the label with the highest probability as the output.

C) Naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. This is known as the "naive" assumption, and it simplifies the computation of probabilities. Despite this simplifying assumption, Naive Bayes classifiers have been shown to perform well in many real-world applications.

D) The statement "They are very effective when the features strongly correlate with each other" is false. Naive Bayes classifiers assume that features are independent, so they do not work well when the features strongly correlate with each other. In such cases, the classifier may make incorrect predictions due to the violation of the independence assumption.

(Page 34, Lecture 2)