1. (True or false) Multimedia processing falls into the area of big data.

True. Multimedia processing is related to the variety of data in big data. One of the defining characteristics of big data is its variety, which refers to the wide range of data types and formats that are generated from various sources. Multimedia data, including images, videos, audio, and other media types, are prime examples of such diverse data. These types of data require specialized processing techniques and tools that can handle the complexity and variety of the data. Therefore, multimedia processing is an essential part of big data analytics, as it helps organizations to extract insights from the large and diverse datasets that they collect. (Page 50, Lecture 1)

2. (True or false) Data mining only involves finding patterns in data that are already known and easily interpretable by humans.

False. Data mining is a process that involves discovering patterns, relationships, and insights in large datasets using various analytical techniques and methods. One of the key advantages of data mining is its ability to identify patterns and relationships that may not be easily recognizable or interpretable by humans. These patterns may be hidden in large, complex datasets, and data mining algorithms can be used to uncover them. Therefore, the statement "Data mining only involves finding patterns in data that are already known and easily interpretable by humans" is false. (P56, Lecture 1)

3. (True or false) The sample space of an experiment represents all possible outcomes that can occur.

True. The sample space of an experiment is the set of all possible outcomes that can occur during the course of the experiment. It includes every possible result that can be obtained, even if some of those outcomes may be unlikely or impossible in practice. The sample space is a fundamental concept in probability theory and is used to calculate the probabilities of various events and outcomes. (Page 4, Lecture 2)

4. (True or false) K-means clustering is a supervised machine learning algorithm that assigns each observation to the cluster with the nearest mean.

False. k-means clustering is an unsupervised machine learning algorithm, not a supervised one. This means that it does not use labelled data to assign observations to clusters, but rather finds clusters based on their similarity. (Page 27-28, Lecture 4).

5. (True or false) The parameter size will remain unchanged given a larger N in N-gram models because of the Markov assumption.

False. The Markov assumption in N-gram models states that the probability of a word only depends on the preceding n-1 words, where n is the order of the model. This assumption reduces the number of parameters required to estimate the probability distribution of the next word, as only a subset of N-grams needs to be considered. However, the parameter size in N-gram models will generally increase as the training corpus size (N) increases, because larger training corpora will have more unique N-grams. Therefore, the Markov assumption does not imply that the parameter size will remain unchanged given a larger N in N-gram models. (Page 46-49, Lecture 2)

6. (True or false) If a random variable X follows standard normal, its probability density function (PDF) is $\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$. So we can infer the probability of observing $X = 0$ is $\frac{1}{\sqrt{2\pi}}$.

False. The probability of observing a specific value in a continuous distribution, such as the standard normal distribution, is zero. Therefore, the probability of observing $X = 0$ in a standard normal distribution is zero.

At any specific value of x, including x=0, the probability density function (PDF) value represents the "density" of the distribution at that point, but not the probability of observing that value. The probability of observing a range of values, such as between -1 and 1, can be calculated by integrating the PDF over that range. (Page 19-24, Lecture 3).

7. (True or false) The central limit theorem applies to any samples identically distributed, independent, and large enough, regardless of their population distribution.

True. The central limit theorem (CLT) applies to any samples identically distributed, independent, and large enough, regardless of their population distribution. The CLT states that as the sample size increases, the sampling distribution of the mean of the sample approaches a normal distribution, regardless of the distribution of the population from which the sample is drawn. This means that even if the population distribution is not normal, as long as the sample size is large enough and the samples are independently and identically distributed, the distribution of sample means will be approximately normal. (Page 18, Lecture 3)

8. (True or false) Sigmoid functions are commonly used for binary classification.

True. Sigmoid functions are commonly used in binary classification problems, where the goal is to predict a binary output (i.e., 0 or 1) based on input features. The sigmoid function maps any input value to a value between 0 and 1, which can be interpreted as the probability of the output being 1 given the input. Therefore, sigmoid functions are often used as activation functions in the output layer of neural networks for binary classification tasks.

(Page 33, Lecture 5)

9. (True or false) In logistic regression, the parameters of the models can be seen as the weights over features.

True. In logistic regression, the model parameters are often called coefficients or weights. Each feature in the model is assigned a weight that reflects its contribution to the predicted probability of the outcome variable. These weights can be positive or negative, indicating the direction and strength of the relationship between each feature and the outcome.

Therefore, the model’s parameters can be seen as the weights over features. (Page 34, Lecture 5)

10. (True or false) R allows the user to give an object a name that already exists, and R will not warn you when you use an existing name.

True. R allows the user to give an object a name that already exists, and it will not warn you when you use an existing name. This behaviour can potentially lead to unintentional overwriting of existing objects, so it is important to be careful when assigning names to objects in R. (Page 22, Lecture 6)

11. (Multi-choice) Which of the following is true about a random sample from a population?
    A) The sample consists of dependent random variables.
    B) The sample consists of identically distributed random variables.
    C) The outcomes of the experiment are fixed values.
    D) A statistic is a fixed value, not a random variable.

Answer: B) The sample consists of identically distributed random variables.

Option B is correct because the given statement clearly states that the random sample from a population consists of independent, identically distributed random variables, which means that each variable in the sample is independent of the other and has the same probability distribution. Therefore, option B is the correct answer.

A) The sample consists of dependent random variables: This statement is not true because each observation is selected independently and randomly from the population in a random sample. Therefore, the sample's random variables are independent, not dependent.

C) The experiment’s outcomes are fixed values: This statement is not true because the outcomes of a random sample are random variables because they depend on chance. The values are not fixed; they vary from one sample to another.

D) A statistic is a fixed value, not a random variable: This statement is not true because a statistic is a function of the sample data and is a random variable. A statistic's value depends on the sample taken, and therefore it varies from sample to sample.

(Page 9, Lecture 3)

12. (Multi-choice) Suppose a bag contains 5 red balls and 7 blue balls. You randomly choose a ball from the bag, and without replacing it, you choose a second ball. What is the probability that the second ball is red, given that the first ball was blue?

    A) 5/12
    B) 1/3
    C) 5/11
    D) 2/3

The correct answer is C).

To solve this problem using conditional probability, you need to use the formula:

P(A|B) = P(AB) / P(B),

where A is the event that the second ball is red, and B is the event that the first ball is blue.

The probability of selecting a blue ball on the first draw is P(B)=7/12. After this, 5 red balls and 6 blue balls remain in the bag. Therefore, the probability of selecting a red ball in the second draw and a blue ball in the first draw is P(AB)=7/12*5/11.

So, P(A|B) = 5/11.

(Page 14, Lecture 2).

13. (Multi-choice) Which of the following is a characteristic of data analytics?

A. It relies solely on computer science to extract insights from data.
B. It is not concerned with the amount of data used in the analysis.
C. It involves the discovery of knowledge and information from data.
D. It is an isolated field with no connection to other disciplines.

The correct answer is C.

Data analytics examines large and varied data sets to identify patterns, draw conclusions, and make decisions based on the data. Data analytics aims to discover insights and knowledge from the data, which can then be used to inform business decisions, improve processes, or identify new opportunities.

Therefore, option C is the correct answer as it highlights that discovering knowledge and information from data is a key characteristic of data analytics. Option A is incorrect because data analytics involves not just computer science, but also statistical analysis, mathematics, and domain expertise. Option B is incorrect because the amount of data used in the analysis is often a crucial factor in determining the accuracy and reliability of the results. Option D is incorrect because data analytics is a multidisciplinary field that draws on knowledge and methods from various disciplines, including computer science, statistics, mathematics, and domain-specific fields such as business or healthcare.

(Page 29-30, Lecture 1)

14. (Multi-choice) Which of the following BEST describes probabilistic language models?

A) They are used to generate human-like responses in chatbots.
B) They are based on a statistical analysis of large amounts of text data.
C) They can only be trained on a small amount of data.
D) They do not take into account the context of a sentence.

The correct answer is B.

Probabilistic language models are a natural language processing (NLP) technique that uses statistical methods to analyse and generate language. These models are based on analysing large amounts of text data, and they use probability distributions to determine the likelihood of different word sequences or phrases. They are used in various applications, including speech recognition, machine translation, and chatbots, to generate more human-like responses. Therefore, option A is partially correct but not the best answer. Options C and D are incorrect.

Option A is partially correct because while probabilistic language models can generate human-like responses in chatbots, this is not their only use case. Probabilistic language models are used in various natural language processing tasks, including speech recognition, machine translation, and sentiment analysis. Therefore, while generating human-like responses in chatbots is one application of probabilistic language models, it is not the only application. These models can be used to analyse and generate language in many other ways.

Option B is the best answer because probabilistic language models are based on a statistical analysis of large amounts of text data. By analysing a large corpus of text, these models can estimate the probability of different words and word sequences, which allows them to generate more accurate and natural language responses.

Option C is incorrect, as probabilistic language models can be trained on large amounts of data. The more data these models are trained on, the better their performance tends to be.

Option D is also incorrect, as probabilistic language models do take into account the context of a sentence. By analysing the surrounding words and word sequences, these models can estimate the probability of different words in a given context, allowing them to generate more accurate and natural language responses.

(Page 42, Lecture 2)

15. (Multi-choice) Suppose that a company produces two types of products, A and B. The probability of producing a defective product for type A is 0.1, and for type B is 0.15. The proportion of type A products produced is 0.6, and the proportion of type B products produced is 0.4. Given that a randomly selected product is defective, what is the conditional probability that the product is of type A?

    A) 0.32
    B) 0.40
    C) 0.50
    D) 0.60

The correct answer is C).

We can use Bayes' formula for conditional probability to solve the problem.

Let A be the event that the selected product is of type A, and let D be the event that the selected product is defective. We want to find P(A|D), the conditional probability that the selected product is of type A, given that it is defective.

By Bayes' formula, we have:

P(A|D) = P(D|A) * P(A) / P(D)

where P(D|A) is the probability of selecting a defective product given that it is of type A, P(A) is the proportion of type A products produced, and P(D) is the overall probability of selecting a defective product. We can compute these probabilities as follows:

P(D|A) = 0.1 (given)
P(D|B) = 0.15 (given)
P(A) = 0.6 (given)
P(B) = 0.4 (given)
P(D) = P(D|A) * P(A) + P(D|B) * P(B) = 0.1 * 0.6 + 0.15 * 0.4 = 0.12

Now we can substitute these values into the formula:

P(A|D) = P(D|A) * P(A) / P(D)
= 0.1 * 0.6 / 0.12
= 0.5

Therefore, the conditional probability that the selected product is of type A, given that it is defective, is 0.5, which indicates that C is the correct answer.

(Page 20, Lecture 2)

16. (Multi-choice) Which of the following is the correct definition of the derivative of a function?
   A) The slope of the tangent line to the function at a specific point
   B) The area under the curve of the function between two points
   C) The average rate of change of the function over a specific interval
   D) The maximum value of the function over a specific interval

The correct answer is A) The slope of the tangent line to the function at a specific point. The derivative of a function is the rate at which the function changes at a specific point. It is the slope of the tangent line to the function at that point. Option B defines the integral of a function, not the derivative. Option C is related to the average rate of change concept, but not the precise definition of the derivative. Option D is incorrect because the derivative is not the maximum value of a function over a specific interval. (Page 10, Lecture 5).

17. (Multi-choice) What is the value of the integral $\int (x^2 + 2x - 3)\, dx$ from $x = -1$ to $x = 2$?

A) -3
B) 0
C) 3
D) 6

The correct answer is A.

We can find the antiderivative of the integrand $(x^2 + 2x - 3)$ by applying the power rule:

$$\int (x^2 + 2x - 3)\,dx = (\frac{x^3}{3} + x^2 - 3x) + C$$

where C is an arbitrary constant.

Then, we can evaluate the definite integral from $x = -1$ to $x = 2$ by plugging in the upper and lower limits:

$$\int_{-1}^{2} (x^2 + 2x - 3)\,dx$$
$$= [\left(\frac{2^3}{3} + 2^2 - 3 * 2\right) - (-\frac{1^3}{3} + (-1)\hat{}2 - 3 * (-1))]$$
$$= [8/3 + 4 - 6 - (-1/3 + 1 + 3)]$$
$$= [8/3 - 6 + 1/3]$$
$$= -3$$

Therefore, the value of the integral $\int (x^2 + 2x - 3)\,dx$ from $x = -1$ to $x = 2$ is -3. (Page 42-45, Lecture 5)

18. (Multi-choice) Which symbol is used in R to represent missing values, and which symbol is used to represent impossible values?

A) NaN represents missing values, and NA represents impossible values.
B) NA represents missing values, and NaN represents impossible values.
C) NA represents both missing and impossible values.
D) NaN represents both missing and impossible values.

The correct answer is B. Missing values are represented by NA and impossible values are represented by NaN. (Page 45, Lecture 6)

19. (Multi-choice) Which of the following statements regarding decision and loss functions in machine learning training is true?
A) Decision functions make predictions, while loss functions measure the error between the predicted output and the true output.
B) Decision functions measure the error between the predicted and true output, while loss functions make predictions.

C) Decision and loss functions are the same and are used interchangeably in machine learning training.
D) Decision and loss functions are unimportant in machine learning training.

The correct answer is A.

Decision functions are used to predict the output given some input, while loss functions measure the error or the difference between the predicted output and the true output. During training, the goal is to find the decision function that minimizes the loss function, i.e., the function that makes the most accurate predictions. Therefore, both decision functions and loss functions are important in machine learning training.

20. (Multi-choice) Which of the following properties of vectors is NOT true?

A) Vector addition is commutative: $\mathrm{u} + \mathrm{v} = \mathrm{v} + \mathrm{u}$, where $\mathrm{u}$ and $\mathrm{v}$ are both vectors.
B) Vector multiplication by a scalar is distributive: $\mathrm{a}(\mathrm{u} + \mathrm{v}) = \mathrm{au} + \mathrm{av}$, where $\mathrm{u}$ and $\mathrm{v}$ are both vectors and a is a scalar.
C) Vector multiplication by a scalar is associative: $\mathrm{a}(\mathrm{bu}) = (\mathrm{ab})\mathrm{u}$, where $\mathrm{u}$ is a vector and a and b are both scalars.
D) The dot product of two vectors is always in the range of -1 and 1: $\mathrm{u}^{\mathrm{T}}v \in [-1,1]$, where $\mathrm{u}$ and $\mathrm{v}$ are both vectors.

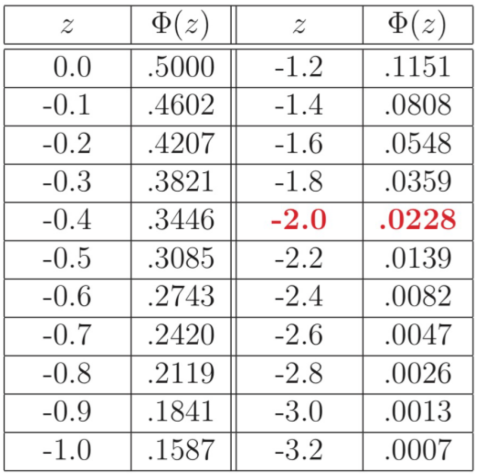
The correct answer is D.

A) Vector addition is commutative: $\mathrm{u} + \mathrm{v} = \mathrm{v} + \mathrm{u}$, where u and v are both vectors. This means that the order in which we add vectors does not affect the result.
B) Vector multiplication by a scalar is distributive: $\mathrm{a}(\mathrm{u} + \mathrm{v}) = \mathrm{au} + \mathrm{av}$, where u and v are both vectors and a is a scalar. This means that we can distribute a scalar factor over a sum of vectors.
C) Vector multiplication by a scalar is associative: $\mathrm{a}(\mathrm{bu}) = (\mathrm{ab})\mathrm{u}$, where u is a vector and a and b are both scalars. This means that the order in which we multiply vectors by scalars does not affect the result.
D) The dot product of two vectors is always in the range of -1 and 1 is NOT true. The dot product of two vectors is defined as the product of their norms and the cosine of the angle between them. The cosine of an angle can range from -1 to 1, but the dot product itself can take on any real value. Therefore, statement D is not true.

(Page 9-24, Lecture 4)

21. (Multi-answer) A researcher wants to test if the mean height of a population is 170 cm (null hypothesis $H_0$). The standard deviation of the population height is known to be 10 cm. He takes a random sample of 100 people and measures their heights. He finds that the sample mean is 172 cm. The standard normal distribution table is shown in the

following, where $\Phi(z)$ is the cumulative distribution function of the standard normal.



| $z$ | $\Phi(z)$ | $z$ | $\Phi(z)$ |
|---|---|---|---|
| 0.0 | .5000 | -1.2 | .1151 |
| -0.1 | .4602 | -1.4 | .0808 |
| -0.2 | .4207 | -1.6 | .0548 |
| -0.3 | .3821 | -1.8 | .0359 |
| -0.4 | .3446 | **-2.0** | **.0228** |
| -0.5 | .3085 | -2.2 | .0139 |
| -0.6 | .2743 | -2.4 | .0082 |
| -0.7 | .2420 | -2.6 | .0047 |
| -0.8 | .2119 | -2.8 | .0026 |
| -0.9 | .1841 | -3.0 | .0013 |
| -1.0 | .1587 | -3.2 | .0007 |

Which of the following statements are correct? Select all that apply.

A) The researcher should reject $H_0$ at the level of significance 0.01.
B) The researcher should reject $H_0$ at the level of significance 0.03.
C) The researcher should reject $H_0$ at the level of significance 0.06.
D) The researcher should reject $H_0$ at the level of significance 0.09.

The correct answer is CD). To conduct the hypothesis test, the researcher can calculate the test statistic, which is the z-score given by:

$$Z = (\bar{X} - \mu) / (\sigma / \sqrt{N})$$

where $\bar{X}$ is the sample mean, $\mu$ is the population mean (under the null hypothesis, $\mu = 170$), $\sigma$ is the population standard deviation (which is unknown and replaced by the sample standard deviation), and n is the sample size.

Plugging in the values, we get:

$$Z = (172 - 170) / (10 / \sqrt{100}) = 2$$

Using the standard normal distribution table, we can find the probability of obtaining a z-score of 2 or higher:

$$P(Z \geq 2) = \Phi(-2) = 0.0228$$

Since this is a two-tailed test, the probability of obtaining a z-score of 2 or higher in either direction is:

$$P(Z \geq 2 \; or \; Z \leq -2) = 2\,\Phi(-2) = 0.0456$$

This is the p-value of the test. It represents the probability of obtaining a sample mean as extreme as 172 cm or more extreme, assuming the null hypothesis is true. We will reject the null hypothesis if the p-value (0.0456) is smaller than the given significant level. So CD is the correct answer. (Page 31, Lecture 3)

22. (Multi-answer) Which of the following statements are true regarding the properties of expected values and variance of discrete random variables? Select all that apply.

A) The expected value of a constant is equal to the constant itself.
B) The expected value of a sum of random variables is equal to the sum of their expected values.
C) The variance of a constant is equal to zero.
D) The variance of a sum of random variables is equal to the sum of their variances.

The correct answer is ABC.

A) True. The expected value of a constant is equal to the constant itself, since the constant is not a random variable and has no variability.
B) True. The expected value of a sum of random variables is equal to the sum of their expected values, regardless of whether they are independent or dependent.
C) True. The variance of a constant is equal to zero, since a constant has no variability.
D) False. The variance of a sum of random variables may not equal to the sum of their variances. For example, let X be the number of heads obtained when tossing a fair coin twice, and let Y be the number of tails obtained in the same tosses. The variance of X is 1/2, the variance of Y is 1/2, and the variance of X + Y is 0, because X + Y is always 2.

(Page 5-6, Lecture 3)

23. (Multi-answer) Which of the following statements about gradients are true? Select all that apply.

A) Gradients are a vector quantity.
B) Gradients 0 indicate that the corresponding point is a global optimal solution.
C) Gradients can be used to optimize loss functions in machine learning.
D) Gradients are closely related to derivatives.

The correct answers are ACD.

Gradients are vectors that represent the direction and length (norm) of the steepest ascent of a function. In machine learning, gradients are used to optimize loss functions by iteratively adjusting model parameters in the direction of decreasing loss. Gradients are closely related to derivatives, which are used to compute the rate of change of a function at a point. The gradient is a generalization of the derivative to functions of multiple variables. However, statement B is not always true. A gradient of 0 at a point only indicates a local optimal solution, not necessarily a global one. (Page 20-25, Lecture 5).

24. (Multi-answer) Which of the following statements are true regarding vectors? Select all that apply.

A) Vectors have length but no direction.
B) Vectors can be added and subtracted using the head-to-tail method.

C) Vectors can be seen as the columns or rows of a matrix.
D) The dot product of two vectors of the same dimension always yields a scalar.

The correct answer is BCD.

A) This statement is false. Vectors have both norm (length) and direction.
B) This statement is true. Vectors can be added and subtracted by placing the tail of one vector at the head of another vector, and the result is a new vector that goes from the tail of the first vector to the head of the second vector.
C) This statement is also true. Vectors can be seen as the columns or rows of a matrix. A matrix is a rectangular array of numbers that can be used to represent a list of vectors. A vector can be written as a matrix with one column (a column vector) or one row (a row vector). The column vectors of a matrix form the column space of the matrix, which is the set of all linear combinations of the column vectors. The row vectors of a matrix form the row space of the matrix, which is the set of all linear combinations of the row vectors.
D) This statement is also True. The dot product of two vectors always results in a scalar. The dot product is defined as the product of the magnitudes of the two vectors and the cosine of the angle between them. It is also equal to the sum of the product of the corresponding components of the vectors. The dot product is a measure of how closely two vectors align, in terms of the directions they point.

(Page 6, Lecture 4)

25. (Multi-answer) Which of the following statements about matrix multiplication are true? Select all that apply.

A) Matrix multiplication is not commutative in general, but it becomes commutative when one of the factors is an identity matrix, such that AB = BA.
B) The product of two matrices with dimensions m x n and p x q is a matrix with dimensions (m+n) x (p+q).
C) The product of two matrices is only defined if the number of columns in the first matrix is equal to the number of rows in the second matrix.

The correct answer is AC.

A) Matrix multiplication is not commutative in general, which means that AB is not necessarily equal to BA. However, if one of the matrices is an identity matrix, then the product is commutative. This is because an identity matrix multiplied by any other matrix returns that matrix unchanged, so if A is an identity matrix, then AB = BA=B.

C) In order to multiply two matrices together, the number of columns in the first matrix must be equal to the number of rows in the second matrix. This is because matrix multiplication is essentially a way of multiplying corresponding row and column entries and summing the results. If the two matrices have different numbers of rows and columns, then there won't be a one-to-one correspondence between entries and the multiplication won't be defined.

B) The dimensions of the product of two matrices with dimensions m x n and p x q are m x q, given that n=p. The resulting matrix has the same number of rows as the first matrix and the same number of columns as the second matrix. For example, if A is a 2x3 matrix and B is a 3x4 matrix, then AB is a 2x4 matrix. The number of rows in the product is the same as the number of rows in the first matrix (A), and the number of columns in the product is the same as the number of columns in the second matrix (B).

(Page 60-62, Lecture 4)

26. (Multi-answer) Which of the following are true regarding the data analysis process? Select all that apply.
   A) The goal of data analysis is to discover useful information, inform conclusions, and support decision-making.
   B) Data analysis involves only inspecting and cleaning data.
   C) The modelling stage of data analysis is not important in discovering useful information.
   D) The data analysis process involves transforming data to make it more useful.

The correct answer is AD.

Option A is correct because the main goal of data analysis is to extract useful information from data, and to use this information to inform conclusions and support decision-making.

Option D is also correct because data transformation is an essential part of the data analysis process. Data may need to be transformed in order to correct errors, remove outliers, standardize values, or prepare it for modelling. Data transformation can help to make the data more useful and relevant for analysis.

Option B is incorrect because data analysis involves more than just inspecting and cleaning data. While inspecting and cleaning data is an important first step, data analysis also involves transforming and modelling the data to extract useful information.

Option C is incorrect because modelling is an important part of the data analysis process. Modelling allows us to create statistical or mathematical models of the data that can be used to predict future outcomes, understand relationships between variables, and test hypotheses. Modelling is a key step in the process of discovering useful information and supporting decision-making.

(Page 29, Lecture 1)

27. (Multi-answer) Which of the following statements are true about cosine similarity of two data samples in vector representation? Select all that apply.

   A) Cosine similarity is a measure of the angle between two vectors.
   B) Cosine similarity is always between 0 and 1.
   C) Cosine similarity is highly sensitive to norm of the two vectors.
   D) Cosine similarity can reflect the data similarity.

The correct answer is AD.

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space. It is defined to equal the cosine of the angle between them, which is also the same as the inner product of the same vectors normalized to both have length. Therefore, A is true.

Cosine similarity can reflect the data similarity because it measures how aligned the two vectors are, regardless of their magnitude. The closer the cosine similarity is to 1, the more similar the two vectors are. The closer the cosine similarity is to 0, the more orthogonal the two vectors are. Therefore, D is true.

Cosine similarity is not always between 0 and 1. It can be negative if the angle between the two vectors is greater than 90 degrees. This means that the two vectors are pointing in opposite directions. Therefore, B is false.

Cosine similarity is not highly sensitive to the norm of the two vectors, because it is normalized by dividing the dot product of the vectors by the product of their lengths. This means that the cosine similarity only depends on the direction of the vectors, not their norm (length). Therefore, C is false.

28. (Multi-answer) Which of the following statements are true about using R? Select all that apply.

    A) You can only enter commands one at a time at the command prompt (>)
    B) You can run a set of commands from a source file
    C) R only supports numerical data types such as vectors
    D) R can support a wide variety of data types such as matrices, dataframes, and lists.

The correct answer is BD. This is because R allows users to both enter commands one at a time at the command prompt (>) and run a set of commands from a source file. Additionally, R supports a wide variety of data types, such as vectors (numerical, character, logical), matrices, dataframes, and lists. (Page 7, Lecture 6)

29. (Multi-answer) Which of the following statements are true about decision function, loss function, machine learning goal, and gradient descent? Select all that apply.
    A) The decision function can be used to map data samples to the classification labels.
    B) The loss function measures the accuracy of the model on the test data.
    C) The machine learning goal is to minimize the loss function on the training data.
    D) Gradient descent is an optimization algorithm used to find the optimal parameters of the model.

The correct answer is ACD.

A) True. The decision function is a mapping from input features to output labels, so it can be used to map data samples to classification labels.

B) False. The loss function measures the error of the model on the training data, not the test data.
C) True. The goal of machine learning is to minimize the loss function on the training data, so that the model can generalize well to unseen test data.
D) True. Gradient descent is an optimization algorithm that can be used to find the optimal parameters of model by iteratively updating the parameters in the direction of the negative gradient of the loss function.

(Page 35, Lecture 5)

30. (Multi-answer) Which of the following statements are true regarding Naive Bayes classifier? Select all that apply.

A) Naive Bayes assumes that the features are conditionally independent given the class.
B) Naive Bayes can be considered as a linear classifier.
C) Naive Bayes is commonly used for unsupervised learning tasks.
D) Naive Bayes can possibly handle missing features (those present in the test set whereas absent in the training set).

The correct answer is ABD.

A) Naive Bayes assumes that the features are conditionally independent given the class. This statement is true. Naive Bayes assumes that each feature is independent of every other feature, given the class variable. This means that the presence or absence of a particular feature does not affect the probability of any other feature occurring.

B) Naive Bayes can be considered as a linear classifier. This statement is true. Naive Bayes can be considered as a linear classifier because it uses a linear decision boundary to separate the classes based on the posterior probabilities of the classes given the input features. The log-likelihoods and log-priors are added together to form the linear function that defines the decision boundary.

C) Naive Bayes is commonly used for unsupervised learning tasks. This statement is false. Naive Bayes is a supervised learning algorithm that requires labelled training data to estimate the parameters of the model. It cannot be used for unsupervised learning tasks, which do not involve labelled data.

D) Naive Bayes can possibly handle missing features (those present in the test set whereas absent in the training set). This statement is true. Naive Bayes can handle missing features by ignoring them during training and classification. The algorithm assumes that the missing values are missing at random and does not attempt to impute them. Therefore, Naive Bayes can still make predictions based on the available features.

(Page 27-39, Lecture 2; Page 43, Lecture 3)