CISC 7700X Final Exam

Pick the best answer that fits the question. Not all of the answers may be correct. If none of the answers fit, write your own answer.

1. A model is:
   (a) A guarantee that predictions will always be correct
   (b) A simplified representation of a real-world process or system
   (c) A database used to store training data
   (d) A programming language designed for machine learning

2. An investment has annual returns of +25%, +25%, and −50% over three consecutive years. Which statement is correct?
   (a) The arithmetic mean return is 0%, and the geometric mean return is 0%
   (b) The arithmetic mean return is negative, and the geometric mean return is 0%
   (c) Both the arithmetic mean return and the geometric mean return are positive
   (d) The arithmetic mean return is 0%, and the geometric mean return is negative

3. What is the primary purpose of bootstrapping in data science and statistics?
   (a) To increase the size of a dataset by collecting additional real-world observations
   (b) To remove outliers before training a machine learning model
   (c) To estimate the variability of a statistic by repeatedly sampling with replacement from the observed data
   (d) To normalize variables so they follow a standard normal distribution

4. What is the main idea behind a permutation test?
   (a) Randomly shuffle the labels or group assignments to determine whether an observed difference could have occurred by chance
   (b) Repeatedly sample observations with replacement to estimate confidence intervals
   (c) Split the dataset into training and testing subsets to evaluate model accuracy
   (d) Rearrange the order of variables to improve model performance

5. Which statement correctly compares standard deviation and interquartile range (IQR)?
   (a) Standard deviation is resistant to outliers, while IQR is highly sensitive to outliers
   (b) Both standard deviation and IQR measure only the center of a distribution
   (c) IQR is generally more robust to outliers than standard deviation
   (d) Standard deviation and IQR will always have the same value for a dataset

6. Suppose X and Y are independent random variables with variances $Var(X) = 4$ and $Var(Y) = 9$. What is $Var(X + Y)$?

   (a) 5

   (b) 13

   (c) 36

   (d) 81

7. You discover a randomly selected widget with serial number 800. Assuming serial numbers start at 1 and are assigned sequentially without gaps, approximately how many total widgets could exist with 50% confidence?

   (a) About 400

   (b) About 800

   (c) About 1600

   (d) About 3200

8. Events A and B are independent, with $P(A) = 0.4$ and $P(B) = 0.5$. What is $P(A \cap B)$?

   (a) 0.1

   (b) 0.2

   (c) 0.4

   (d) 0.9

9. Let A and B be discrete random variables. Which statement correctly describes marginalization?

   (a) $P(A) = P(A \mid B)$

   (b) $P(A) = \sum_b P(A, B = b)$

   (c) $P(A, B) = P(A) + P(B)$

   (d) $P(A) = \frac{P(A,B)}{P(B)}$

10. Which expression correctly represents Bayes' rule?

    (a) $P(A \mid B) = \frac{P(A)P(B)}{P(A \cap B)}$

    (b) $P(A \mid B) = P(A) + P(B)$

    (c) $P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$

    (d) $P(A \mid B) = \frac{P(A \cap B)}{P(A)}$

11. In Bayes' rule, which statement correctly describes the roles of likelihood and prior?

    (a) The likelihood represents the prior belief before seeing data, while the prior represents the probability of the data

    (b) The likelihood and prior are always equal in Bayes' rule

    (c) The prior is computed after observing data, while the likelihood is independent of the data

    (d) The likelihood is the probability of the observed data given a hypothesis, while the prior represents belief about the hypothesis before seeing data

12. Which statement correctly distinguishes likelihood from conditional probability?

    (a) Likelihood treats the observed data as fixed and varies the parameter, while conditional probability treats both events as random variables in a probability statement

    (b) Likelihood is $P(A \mid B)$ and conditional probability is $P(B \mid A)$

    (c) Likelihood and conditional probability are identical concepts with different names

    (d) Conditional probability ignores observed data, while likelihood does not

13. At a car dealership, approximately 1% of visitors end up purchasing a car. Of those who purchase, 70% go directly to the car they ultimately buy. Of those who do not purchase, 90% continue walking around the entire store rather than going directly to a specific car. A customer enters the dealership and goes directly to a specific make/model. What is the probability that they will end up purchasing that car?

(answer)

14. Continuing the previous question, 60% of customers who purchase a car come with family, while customers who do not purchase come with family only 5% of the time. A customer enters the dealership with family. What is the probability that they will purchase a car?

(answer)

15. Continuing the previous question, a customer enters the dealership with family and goes directly to a single make/model. What is the probability that they will purchase a car?

(answer)

16. Continuing the previous question, a customer enters the dealership with family and goes directly to a single make/model. Use Naive bayes to calculate probability that they will purchase a car?

(answer)

17. Which statement best describes the difference between correlation and causation?

    (a) Correlation implies that one variable directly causes changes in another variable

    (b) Causation means two variables are unrelated, while correlation means they are related

    (c) Correlation indicates an association between variables, but does not necessarily imply one causes the other

    (d) Causation can only be measured using correlation coefficients

18. Which approach provides the strongest evidence that one variable causes another?

    (a) Observing a strong correlation between the two variables in a large dataset

    (b) Increasing the sample size of an observational study

    (c) Fitting a more complex predictive model to the data

    (d) Running a randomized controlled experiment where the explanatory variable is randomly assigned

19. You want to determine whether the difference in average outcomes between two groups is statistically significant, but you do not want to assume any specific distribution for the data. Which method is most appropriate?

    (a) A permutation test, where group labels are randomly shuffled to generate a null distribution

    (b) A z-test, assuming the data are normally distributed with known variance

    (c) A linear regression with a single binary predictor, assuming homoscedastic errors

    (d) A t-test, assuming equal variances and normality in both groups

20. Consider a fair coin-flipping game where you start with \$1. Each round, you flip a fair coin: if it lands heads, your wealth increases by 50%; if it lands tails, your wealth decreases by 50%. After 3 rounds, what is the geometric mean of your wealth?

(answer)