UM-DAE-Centre for Excellence in Basic Sciences
End - Semester Examination: Semester X
Machine Learning & Artificial Intelligence (C1004)

Name & Roll No: **May 6, 2024** **Time: 2 hrs**

1. You have two machine learning algorithms. Algorithm A is a decision tree; Algorithm B is a k-nearest neighbors classifier. Which of the following statements is likely to be true about their inductive biases?

(a) Algorithm A has a stronger bias towards linear patterns than Algorithm B.
(b) Algorithm B has a stronger bias towards local patterns than Algorithm A.
(c) Algorithm A and Algorithm B have the same type of inductive bias.
(d) Inductive bias does not play a significant role with either algorithm.

2. You're developing a model to predict housing prices in various cities. You notice your model consistently underestimates prices in certain neighborhoods. To address this, which of the following actions is MOST likely to improve performance?

(a) Increase the model's overall complexity.
(b) Apply stronger regularization techniques.
(c) Review the data collection process and the assumptions in your model.
(d) Train the model for a longer period of time.

3. When selecting a machine learning algorithm, the bias-variance trade-off suggests that:

(a) Algorithms with strong inductive bias are always more accurate.
(b) Algorithms with weak inductive bias will never overfit.
(c) Models with a strong bias may suffer from underfitting.
(d) Models with a weaker bias generally memorize the training data.

4. What does the rank of a matrix tell you about its properties?

(a) The number of linearly independent columns or rows
(b) The number of zero entries in the matrix
(c) The sum of the diagonal elements
(d) The maximum determinant of any submatrix

5. In gradient descent, if the learning rate is too large, what might happen?

(a) The model converges to the optimal solution quickly.
(b) The model might overshoot the minimum and diverge.
(c) The model will get stuck in a local minimum.
(d) Training time will increase significantly.

6. A dataset has an imbalanced class distribution: 95% of instances are 'Positive' and 5% are 'Negative'. Which situation is likely if you train a standard classifier on this data?

(a) The model will have high recall for the 'Negative' class.
(b) The model might predict all instances as 'Positive'.
(c) The model will learn the underlying patterns very effectively.
(d) Accuracy will be a very unreliable metric.

7. What is the key assumption of linear regression?

(a) Variables have a linear relationship
(b) Errors are normally distributed
(c) Features are independent of each other
(d) All of the above

8. A machine learning problem has a non-convex cost function. What does this potentially imply?

(a) Gradient descent will always find the global optimum.
(b) The optimization problem might have multiple local optima.
(c) There is a unique optimal solution.
(d) The cost function is not differentiable.

9. You flip a coin twice. What is the probability of getting at least one heads?

(a) 1/2
(b) 3/4
(c) 1/4
(d) 1

10. Which of the following best characterizes an ill-posed problem in machine learning?

(a) A problem where the optimal solution is computationally difficult to find.
(b) A problem lacking a unique, stable solution.
(c) A problem where the available data is insufficient in quantity or quality.

(d) A problem specifically related to supervised learning tasks.

11. Which of the following describes a continuous random variable?

(a) The number of students in a class
(b) The outcome of a coin toss
(c) The temperature outside
(d) The rating of a movie (1-5 stars)

12. In a Naive Bayes classifier, features are often assumed to follow what type of distribution?

(a) Gaussian
(b) Poisson
(c) Uniform
(d) Exponential

13. You're building a spam filter. P(spam) = 0.2, P(word "free" | spam) = 0.6, P(word "free" | not spam) = 0.05. A new email contains the word "free". What is the probability it's spam? (approximately)

(a) 0.75
(b) 0.6
(c) 0.2
(d) 0.05

14. MLE is a common method for parameter estimation in machine learning models. What is the underlying principle of MLE?

(a) Finding parameters through a random search process
(b) Finding parameters that minimize the distance between predictions and actual values
(c) Finding parameters that give equal probability to all possible outcomes
(d) Finding parameters that maximize the likelihood of the observed data

15. In classification problems, why is it important to consider the probability of an instance belonging to a particular class rather than just the raw predicted value?

(a) Probabilities allow for better calibration and decision-making
(b) Probabilities are necessary for calculating the gradient during training
(c) Probabilities ensure the model is not overfitting
(d) Probabilities improve the computational efficiency of the model

16. What is the expected value (E[X]) of a discrete random variable X?

(a) The most likely outcome of X
(b) The weighted average of possible outcomes of X
(c) The spread or dispersion of X
(d) None of the above

17. How does the concept of entropy find usage in machine learning?

(a) Lower entropy indicates a more overfit model.
(b) Higher entropy indicates greater uncertainty in a probability distribution.
(c) Entropy is used to calculate the gradient for backpropagation
(d) Entropy is only relevant for continuous random variables.

18. You have a mix of structured data (customer demographics, transaction history) and unstructured data (customer support tickets as text). How might you derive features that combine information from both sources for a churn prediction (e.g. customer leaving a service) model?

(a) Apply topic modeling to support tickets, then join topics with structured data
(b) Use one-hot encoding across all features, structured and unstructured
(c) Focus exclusively on the structured data, as unstructured is too noisy
(d) Calculate TF-IDF for the text data and average it with numerical features

19. In natural language processing, how might a probability distribution over words be used to help a language model generate the next word in a sequence?

(a) To prevent the model from repeating the same words too often
(b) To ensure that all words in the vocabulary have an equal chance of being generated
(c) To select the word with the highest probability
(d) To calculate the gradient for backpropagation during training

20. What is the purpose of the activation function in a neural network?

(a) To introduce non-linearity
(b) To control the learning rate
(c) To initialize the weights
(d) To regularize the network

21. In Bayesian inference, what is the term for the distribution used to represent uncertainty about a parameter before observing any data?

(a) Posterior distribution
(b) Likelihood function

(c) Prior distribution
(d) Marginal distribution

22. Given two random variables X and Y, if Cov(X,Y)=0, what can be inferred?
(a) X and Y are independent
(b) X and Y are dependent
(c) X and Y are uncorrelated
(d) X and Y have zero variance

23. What is the formula for the mutual information between two discrete random variables X and Y?
(a) $I(X;Y)=H(X)-H(Y|X)$
(b) $I(X;Y)=H(X)+H(Y)-H(X,Y)$
(c) $I(X;Y)=\sum p(x,y)\log(p(x,y)/p(x)p(y))$
(d) $I(X;Y)=\int p(x,y)\log(p(x,y)/p(x)p(y))\,dx\,dy$

24. What is the name of the measure of the degree of asymmetry of a probability distribution?
(a) Variance
(b) Skewness
(c) Kurtosis
(d) Entropy

25. If P(A)=0.4 and P(B)=0.6, what is the maximum possible value for P(A∩B)?
(a) 0.2
(b) 0.4
(c) 0.6
(d) 0.24

26. You have a mix of numerical, categorical, and text features. Which feature selection method would NOT be directly suitable?
(a) Correlation-based methods
(b) Mutual information
(c) L2 regularization (Ridge Regression)
(d) Select K Best with chi-squared scoring

27. In a group of 50 students, 30 study Mathematics and 20 study Physics. If 15 students study both Mathematics and Physics, what is the probability that a randomly chosen student studies at least one of the subjects?
(a) 4/5
(b) 3/5
(c) 7/10
(d) 2/5

28. Which of the following factors can CONTRIBUTE to a strong inductive bias in a machine learning model?
(a) A large and diverse training dataset
(b) A complex model architecture with many parameters
(c) Prior knowledge about the relationships between features
(d) Using a learning rate that is too high

29. Consider a scenario where you have infinite data and infinite computational resources. How would the concept of inductive bias be relevant in this situation?
(a) Inductive bias becomes irrelevant because any pattern can be learned.
(b) Inductive bias becomes even more critical for guiding efficient learning.
(c) Models built with a strong inductive bias will perform poorly.
(d) Only models built with weak inductive bias will perform well.

30. Why is it important to consider covariance/correlation when performing feature selection for machine learning?
(a) Highly correlated features can provide redundant information.
(b) Features with low covariance are unimportant.
(c) Correlation directly indicates the predictive power of a feature.
(d) Covariance and correlation only matter for unsupervised learning.

31. Which technique addresses the issue of different scales or ranges among features?
(a) One-Hot Encoding
(b) Normalization/Standardization
(c) Imputation
(d) Discretization

32. Why is feature selection important?
(a) Always required for machine learning models
(b) Essential for handling missing data
(c) Reduces overfitting, can make models faster
(d) Always required for deep learning models

33. What challenge does imbalanced data present for classification problems?
(a) Increases the risk of the model overfitting
(b) Always leads to poor performance, regardless of the algorithm
(c) Makes accurate visualization impossible
(d) Makes it difficult to compute the loss function

34. How would you handle text data to make it suitable for machine learning algorithms?

(a) Directly feed sentences into a neural network
(b) Remove punctuation and stop words, use techniques like TF-IDF
(c) Convert text to ordinal data based on alphabetical order
(d) Use feature selection techniques designed for numerical data

35. What might contribute to bias in a machine learning model?
(a) Bias in the training data itself
(b) Using a model that is too simple for the problem
(c) Having too much data
(d) Focusing on accuracy as the primary metric

36. You convert the categorical feature "City" with values ["New York", "London", "Paris"] into numerical features. Which encoding technique is likely the best choice?
(a) Label Encoding
(b) One-Hot Encoding
(c) Binary Encoding
(d) Target Encoding

37. Image reconstruction from blurry or noisy images is a classic example of an ill-posed problem because:
(a) There might be multiple valid images consistent with the input data.
(b) Image processing is only possible with deep learning.
(c) The underlying true image is impossible to recover perfectly.
(d) These problems require large amounts of computational resources

38. What feature scaling is and why it might be used.
(a) Features to have similar ranges, important for some algorithms
(b) Creating more features by combining existing ones
(c) Encoding categorical features into a numerical format
(d) Removing features that are not statistically significant

39. You're working with text data and create a "bag-of-words" representation. What does this typically involve?
(a) Converting text into fixed-length numerical vectors
(b) Replacing all words with their synonyms
(c) Counting how often each word appears in a document
(d) Assigning a unique numerical ID to each word

40. Consider the problem of overfitting in machine learning. This problem arises primarily due to limitations in:
(a) The amount of available training data
(b) The reliance on inductive reasoning
(c) The use of deductive reasoning for feature selection
(d) The lack of domain knowledge in the model

41. Which type of feature interaction might a decision tree model naturally capture that a linear model would not?
(a) A feature having a different effect depending on the value of another feature
(b) Features having a perfectly linear relationship
(c) Features that are statistically uncorrelated
(d) Features with a very high number of unique values

42. You're working with text data containing many misspellings, abbreviations, and informal language. Which of the following is NOT a common technique to address these issues:
(a) Spell checking and correction
(b) Stemming or lemmatization
(c) Replacing all text with its numerical sentiment score
(d) Normalizing text to lowercase

43. You want to create new features based on the combination of two existing numerical features, 'X' and 'Y'. Which of the following might be suitable?
(a) X + Y, X - Y, X * Y
(b) Encoding X and Y using Label Encoding
(c) Calculating the correlation between X and Y
(d) Applying PCA to X and Y

44. Your model is trained on historical product reviews but will be used to classify reviews for new products with potentially different vocabulary. What could help mitigate this difference?
(a) Using exclusively character-level n-grams
(b) Training a word embedding model on a large, diverse text corpus
(c) Decreasing the model's learning rate
(d) Removing all stop words from the data

45. You're building a classification model in a setting where false negatives (failing to identify a true positive instance) have a much higher cost than false positives. Which strategy might you consider?

(a) Increasing the dataset size, regardless of class balance

(b) Adjusting the decision threshold of the model

(c) Using a model specifically designed for high-recall scenarios

(d) Applying under-sampling to the majority class

46. You calculate the pairwise correlations between all numerical features in your dataset. Which of the following scenarios could potentially lead to misleading interpretations?

(a) The presence of categorical features

(b) A small number of outliers in the data

(c) Features with non-linear relationships

(d) Features with a high degree of linear correlation

47. Your machine learning model is deployed in a real-world application where the data distribution gradually changes over time. Which of the following is NOT an appropriate strategy to maintain model performance?

(a) Retraining the model periodically on fresh data

(b) Using online learning algorithms for updates

(c) Monitoring model performance for signs of drift

(d) Relying on the model's initial validation score

48. You're collecting data for a supervised learning task, and the labelling process is prone to occasional errors. Which strategy could potentially help identify and mitigate the impact of mislabelled examples?

(a) Decreasing the regularization strength of your model

(b) Using robust loss functions less sensitive to outliers

(c) Focusing exclusively on precision as an evaluation metric

(d) Removing all instances where the feature values seem unusual

49. Which statement is TRUE about cross-entropy as a loss function?

(a) It penalizes incorrect predictions more heavily than correct ones.

(b) It is only used for regression problems.

(c) It is always minimized when a model outputs the true probability distribution

(d) It directly measures the accuracy of a model.

50. You have a six-sided die, but two of the sides are labelled "3." What is the entropy of rolling this die?

(a) Less than the entropy of a fair die

(b) Equal to the entropy of a fair die

(c) Greater than the entropy of a fair die

(d) It depends on the exact numbers on the other sides

51. A dataset contains both numerical and categorical features for a classification problem. Which feature selection method is more suitable:

(a) Correlation analysis

(b) Principal Component Analysis

(c) Mutual Information

(d) Chi-Square test

52. How can information theory concepts be used for feature selection?

(a) Mutual information can capture non-linear relationships between features and the target variable

(b) Correlation measures are computationally expensive

(c) Mutual information select features which are non-redundant

(d) Removing features that cause overfitting

53. Define the KL-divergence (Kullback-Leibler divergence) between two probability distributions P and Q.

(a) A measure of the similarity between P and Q

(b) A measure of the dissimilarity between P and Q

(c) The average difference between the probabilities in P and Q

(d) The amount of information lost when using Q to approximate P

54. Describe the concept of a gradient in the context of optimization.

(a) The direction of steepest increase of the loss function.

(b) The rate of change of the loss function with respect to the parameters.

(c) The point where the loss function reaches its minimum value.

(d) A measure of how well the model fits the data.

55. Which of the following best describes the general goal of optimization in the context of supervised machine learning?

(a) Find the model parameters that best fit the training data according to a chosen objective.

(b) To ensure that the model should not make mistakes on unseen data.
(c) Create the computationally efficient model possible.
(d) Ensure that the model follows a set of pre-defined rules.

56. Consider a random variable X with four possible outcomes having probabilities: 1/2, 1/4, 1/8, 1/8. What is the entropy of X (in bits)?
(a) 1 bit
(b) 1.5 bits
(c) 1.75 bits
(d) 2 bits

57. The mutual information between two random variables X and Y, I(X;Y), is equal to zero. This implies that:
(a) H(X) = H(Y)
(b) X and Y are perfectly correlated
(c) X and Y are independent
(d) One of the variables has zero entropy

58. Why do distance measures become less meaningful in high-dimensional spaces?
(a) Distance measures are only relevant for spatial data.
(b) Distance calculations become computationally intractable.
(c) High-dimensional data always lies on a curved manifold.
(d) Distances between all points tend to become similar.

59. Which of the following is a common technique to combat the curse of dimensionality?
(a) Increasing the size of the training dataset
(b) Dimensionality reduction
(c) Using more complex models
(d) Collecting data with correlated features

60. What is the purpose of MAP estimation in machine learning?
(a) To randomly sample parameters from a distribution
(b) To find parameters that minimize the distance between predictions and true values
(c) To find parameters that maximize the posterior distribution
(d) To normalize the predicted outputs of a model

61. You have a "date" feature in the format YYYY-MM-DD. Which feature engineering technique could derive useful information?
(a) Extract day of the week, month, or seasonal information
(b) Convert the date into an ordinal number
(c) Apply one-hot encoding
(d) Discretize the date into bins

62. When creating new features by combining existing ones, it's important to consider:
(a) Introducing redundancy into the dataset
(b) The potential for interactions between features
(c) Making the features normally distributed
(d) Removing all outliers before creating new features

63. Which technique is particularly helpful when dealing with outliers?
(a) Mean
(b) Median
(c) Standard Deviation
(d) Maximum

64. When might you choose feature selection over dimensionality reduction?
(a) When model interpretability is important
(b) When you want to visualize the dataset in 2D
(c) When dealing with extremely high-dimensional data
(d) When the original features are very noisy

65. Why is it important to evaluate the impact of feature engineering on model performance?
(a) Feature engineering might introduce unintended biases
(b) To decrease model complexity
(c) To catch errors in the implementation of the techniques
(d) Decreasing the training time of the model

66. You have highly skewed numerical features. Which transformation might be appropriate?
(a) Log transformation
(b) Polynomial transformation
(c) One-hot encoding
(d) Min-Max Scaling

67. How might you address the issue of concept drift in feature engineering and selection?
(a) Using static feature selection methods
(b) Retraining models frequently and monitoring feature distributions
(c) Focusing on creating a large number of features
(d) Applying advanced dimensionality reduction techniques

68. Your model is overfitting, and you have a large number of features. You try L1 regularization (LASSO) for feature selection but find it too drastic (many features zeroed out). What might you consider next?

(a) Decrease the regularization strength of LASSO
(b) Switch to L2 regularization (Ridge)
(c) Use a dimensionality reduction technique like PCA
(d) Remove features with low variance

69. You're developing image recognition software for a self-driving car. Standard image features aren't performing well. What advanced feature engineering might significantly improve your model?

(a) Increase image resolution before extracting features
(b) Use pre-trained convolutional neural networks for feature extraction
(c) Convert images to grayscale to reduce dimensionality
(d) Apply a combination of edge detection and color histograms

70. You have highly imbalanced data for a fraud detection problem. Which feature selection strategy needs careful consideration in this scenario?

(a) Using accuracy as the primary evaluation metric for feature selection
(b) Correlation-based feature selection
(c) Selecting features based on domain expertise
(d) Using embedded feature selection methods

71. A machine learning model trained on weather data is biased towards predicting sunny days more often, even in regions with a higher chance of rain. This situation could be caused by:

(a) Imbalanced data
(b) An inductive bias favoring simpler models
(c) Not enough training data for rainy days
(d) All of the above

72. When evaluating the results of feature engineering on model performance, why is using only a single train/test split problematic?

(a) You might overestimate the true benefits due to dataset specifics
(b) Feature engineering always makes it necessary to use larger datasets
(c) Feature engineering invalidates the use of standard train/test splits
(d) Models will become too sensitive to small changes in the features

73. You're working with highly noisy, non-stationary time-series data. Which feature engineering approach might be helpful before applying a standard forecasting model?

(a) Extract simple moving averages
(b) Discretize the data into bins
(c) Wavelet transforms to decompose into different frequency bands
(d) Directly apply PCA for dimensionality reduction

74. When analysing social network data to predict the spread of information, which feature engineering technique might create informative features from the network structure?

(a) Graph embedding techniques
(b) Creating lagged features based on timestamps
(c) One-hot encoding of user attributes
(d) Applying TF-IDF to the text of posts

75. You suspect interactions between features are crucial for prediction, but there are too many features to explore all combinations manually. Which strategy could help?

(a) Use a decision tree-based model to uncover interactions
(b) Focus on features with the highest correlation to the target variable
(c) Remove features that exhibit non-linearity
(d) Apply L2 regularization (Ridge regression) for feature selection

76. Consider a decision tree with a maximum depth of 3. This depth limitation is an example of:

(a) Overfitting
(b) Underfitting
(c) Inductive bias
(d) Regularization

77. How can dimensionality reduction techniques help visualize high-dimensional data?

(a) Project the data onto a 2D or 3D space
(b) Directly reveal the features most important for classification
(c) Remove noisy samples from the data that distort visualizations
(d) Ensure that visualizations are aesthetically pleasing

78. In the context of using mutual information for feature selection, focusing on features with high scores. What potential limitation should you keep in mind?

(a) Mutual information cannot handle categorical features.
(b) Mutual information ignores feature interactions.
(c) Mutual information is computationally very expensive.
(d) Mutual information only works for regression problems.

79. You're engineering features for a loan approval model. You discover that using a feature derived from the applicant's address leads to a slight performance increase but could perpetuate geographic biases. What's the most responsible action?

(a) Use the feature, but downsample the majority group to address imbalance
(b) Ignore the performance increase and exclude the feature
(c) Add additional features that correlate with address, hoping to dilute the bias
(d) Use the feature and carefully monitor performance for different groups

80. You're working with sensitive medical data for disease diagnosis. Due to the sensitive nature, your dataset is quite small. Which feature engineering approach might be particularly risky in this scenario?

(a) Creating hand-crafted features based on domain knowledge
(b) Using data augmentation to generate synthetic samples
(c) Creating complex feature interactions
(d) Removing features that are missing in a significant number of cases

81. A machine learning model exhibits high error on both the training set and unseen data. This is most likely an indication of:

(a) Underfitting
(b) Overfitting
(c) Low variance
(d) Insufficient training data

82. You observe that increasing model complexity initially improves performance, but then leads to a decline in performance on unseen data. This is a classic sign of:

(a) Insufficient features in the dataset
(b) The need for feature selection
(c) Overfitting
(d) High bias in the model

83. You're working with a very noisy dataset. Which of the following is likely to be an effective strategy for improving model performance?

(a) Focus on a highly complex, flexible model
(b) Choose a model with higher bias (e.g., simpler model)
(c) Include as many features as possible
(d) Train the model on a very small subset of the data

84. The "No Free Lunch" theorem in machine learning suggests that:

(a) There's no single algorithm that performs best universally across all problems
(b) Overfitting is impossible to avoid completely
(c) Complex models will always outperform simpler models
(d) Feature engineering has great impact on model performance

85. You decompose the error of a machine learning model into bias, variance, and Bayes error. Which component of the error cannot be reduced by the model itself?

(a) Bias
(b) Variance
(c) Bayes error
(d) Total error

86. Describe a scenario where choosing a model with slightly higher bias might be a good trade-off.

(a) You have a very large and clean dataset
(b) You need the highest possible accuracy on unseen data.
(c) You have few features and suspect non-linear relationships
(d) Your primary concern is computational efficiency.

87. You're trying to improve a model that suffers from overfitting. Which of these is NOT likely to be helpful?

(a) Collecting more data
(b) Introducing regularization
(c) Increasing the number of features
(d) Switching to a simpler model class

88. When using bagging (e.g., Random Forests) to reduce overfitting, the primary mechanism at work is:
(a) Reducing the complexity of individual models
(b) Decreasing the bias of the ensemble
(c) Decreasing the variance of the ensemble
(d) Creating more balanced datasets

89. Which of the following situations is likely to result in an ill-posed machine learning problem?
(a) Having more features than data samples.
(b) Using a decision tree with a limited maximum depth.
(c) Applying L2 regularization (Ridge regression).
(d) Having a large and perfectly clean dataset.

90. Feature engineering can involve using domain knowledge to create new features likely to be informative. This process reflects elements of:
(a) Purely inductive reasoning
(b) Purely deductive reasoning
(c) A combination of inductive and deductive reasoning
(d) Neither inductive nor deductive reasoning

91. A trained decision tree model classifies new data points by following a series of decision rules derived from the training data. This is an example of:
(a) Inductive reasoning
(b) Deductive reasoning
(c) Analogy-based reasoning
(d) Reinforcement learning

92. Which of the following techniques directly aims to prevent overfitting and improve generalization?
(a) Data augmentation
(b) Feature scaling
(c) Hyperparameter optimization
(d) Regularization

93. Debates about the interpretability and explainability of machine learning models touch upon philosophical questions related to:
(a) The role of aesthetics in model design
(b) The existence of objective truth
(c) Causation and understanding
(d) The limits of human intelligence

94. Which of the following is a core concept shared by statistics and machine learning?
(a) Hypothesis testing
(b) Deductive reasoning
(c) Metaphysics
(d) Determinism

95. In which scenario might a stochastic optimization method like stochastic gradient descent be preferred over full-batch gradient descent?
(a) You have a small dataset and want to avoid overfitting.
(b) You need the highest possible accuracy, regardless of computational cost.
(c) Your dataset is very large and redundant.
(d) Your optimization problem is guaranteed to be convex.

96. Optimization in machine learning often aims to find a balance between:
(a) Underfitting and overfitting
(b) Computational speed and memory usage
(c) Interpretability and fairness
(d) Using supervised and unsupervised learning

97. Which of the following phenomena is a core challenge associated with high-dimensional spaces in machine learning?
(a) Visualization of data becomes impossible.
(b) All data points tend to lie very close to the center of the space.
(c) Data points become sparse, impacting the effectiveness of distance-based algorithms.
(d) Computational complexity is guaranteed to decrease with more dimensions.

98. How does feature selection differ from dimensionality reduction in addressing high-dimensional data?
(a) Feature selection selects a subset of existing features; dimensionality reduction creates new features.
(b) Feature selection is only for supervised learning; dimensionality reduction is for unsupervised.
(c) Dimensionality reduction is computationally more expensive than feature selection.
(d) Feature selection introduces redundancy; dimensionality reduction removes it.

99. Which matrix decomposition is used in Principal Component Analysis (PCA) for dimensionality reduction?
(a) LU Decomposition
(b) QR Decomposition
(c) Singular Value Decomposition (SVD)
(d) Eigen decomposition

100. Consider cross-entropy loss, commonly used in classification. Which statement best connects it to information theory concepts?

(a) It's guaranteed to be minimized when the model perfectly predicts the true probability of each class.

(b) It's directly proportional to the variance of the model's predictions.

(c) It measures the average number of bits needed to encode data coming from the true distribution using a model's predicted distribution.

(d) It's only useful for binary classification problems.