Artificial Intelligence and Machine Learning: Key Concepts and Explanations

1. Introduction to AI, ML, DL, and DS

1. **Define Artificial Intelligence (AI).**
   AI refers to the simulation of human intelligence in machines that can perform tasks such as learning, reasoning, and problem-solving.

2. **Explain the differences between AI, ML, DL, and DS.**

   - **AI:** Broad field enabling machines to perform intelligent tasks.
   - **ML:** Subset of AI that uses algorithms to learn from data.
   - **DL:** A further subset of ML using deep neural networks.
   - **DS:** Extracting insights from data using AI/ML methods.
3. **How does AI differ from traditional software development?**
   AI learns from data and adapts, while traditional software follows explicitly programmed rules.

4. **Provide examples of AI, ML, DL, and DS applications.**

   - **AI:** Chatbots, self-driving cars.
   - **ML:** Fraud detection, recommendation systems.
   - **DL:** Image classification, voice recognition.
   - **DS:** Data visualization, predictive analytics.
5. **Discuss the importance of AI, ML, DL, and DS in today's world.**
   These technologies improve automation, decision-making, and efficiency in healthcare, finance, and other industries.

2. Machine Learning Paradigms

6. **What is Supervised Learning?**
   A learning method using labeled data for training.

7. **Provide examples of Supervised Learning algorithms.**

    - Linear Regression
    - Decision Trees
    - Support Vector Machines
8. **Explain the process of Supervised Learning.**

    - Collect labeled data
    - Train the model
    - Evaluate performance
    - Make predictions
9. **What are the characteristics of Unsupervised Learning?**

    - Works with unlabeled data
    - Finds patterns and structures
    - Used for clustering and association tasks
10. **Give examples of Unsupervised Learning algorithms.**

- K-Means Clustering
- DBSCAN
- Principal Component Analysis (PCA)

11. **Describe Semi-Supervised Learning and its significance.**
    Uses a mix of labeled and unlabeled data, beneficial when labeled data is scarce.

12. **Explain Reinforcement Learning and its applications.**
    An agent learns by interacting with an environment and receiving rewards (e.g., robotics, game AI).

13. **How does Reinforcement Learning differ from Supervised and Unsupervised Learning?**
    It learns through trial and error with feedback rather than labeled or unlabeled data.

3. Model Training and Validation

14. **What is the purpose of the Train-Test-Validation split in machine learning?**
   Ensures models generalize well by evaluating performance on separate datasets.

15. **Explain the significance of the training set.**
   Used to optimize model parameters.

16. **How do you determine the size of the training, testing, and validation sets?**
   Common splits: 70-20-10 or 80-10-10, based on dataset size and complexity.

17. **What are the consequences of improper Train-Test-Validation splits?**
   Overfitting or underfitting leading to poor performance.

18. **Discuss the trade-offs in selecting appropriate split ratios.**
   Larger training sets improve learning, but sufficient validation is needed for tuning.

4. Model Performance and Generalization

19. **Define model performance in machine learning.**
   It refers to how well a model predicts outcomes on new data.

20. **How do you measure the performance of a machine learning model?**

- Accuracy
- Precision, Recall, F1-score
- ROC-AUC

21. **What is overfitting and why is it problematic?**
A model memorizes training data but fails on new data.

22. **Provide techniques to address overfitting.**

- Regularization
- Dropout (for neural networks)
- Cross-validation

23. **Explain underfitting and its implications.**
A model is too simple to learn patterns, leading to poor accuracy.

24. **How can you prevent underfitting in machine learning models?**

- Use a more complex model
- Train for more epochs
- Feature engineering

25. **Discuss the balance between bias and variance in model performance.**

- **High bias:** Underfitting, too simplistic.
- **High variance:** Overfitting, too sensitive to data.
- Goal: Find a trade-off.

5. Data Handling and Preprocessing

26. **What are the common techniques to handle missing data?**

- Deletion
- Mean/Median/Mode imputation
- KNN imputation

27. **Explain the implications of ignoring missing data.**
It can introduce bias and reduce model accuracy.

28. **Discuss the pros and cons of imputation methods.**

- **Mean:** Simple but loses variance.

- **KNN:** Retains patterns but is computationally expensive.

29. **How does missing data affect model performance?**
    It leads to biased results and poor generalization.

30. **Define imbalanced data in the context of machine learning.**
    A dataset where some classes have significantly more samples than others.

31. **Discuss the challenges posed by imbalanced data.**

- Bias towards majority class
- Poor recall for minority class

32. **What techniques can be used to address imbalanced data?**

- Up-sampling
- Down-sampling
- SMOTE

33. **Explain the process of up-sampling and down-sampling.**

- **Up-sampling:** Duplicating minority class examples.
- **Down-sampling:** Reducing majority class examples.

34. **When would you use up-sampling versus down-sampling?**

- **Up-sampling:** When data is limited.
- **Down-sampling:** When majority class dominates too much.

35. **What is SMOTE and how does it work?**
    Synthetic Minority Over-sampling Technique generates synthetic samples.

36. **Explain the role of SMOTE in handling imbalanced data.**
    It balances the dataset by creating synthetic examples.

37. **Discuss the advantages and limitations of SMOTE.**

- **Advantages:** Balances classes, improves recall.
- **Limitations:** May introduce noise.

38. **Provide examples of scenarios where SMOTE is beneficial.**
    Fraud detection, rare disease prediction.

Define data interpolation and its purpose.

Data interpolation estimates unknown values within the range of known data points, used to fill in missing data or smooth data.

40. **What are the common methods of data interpolation?**

Common methods include linear interpolation, polynomial interpolation, and spline interpolation.

41. **Discuss the implications of using data interpolation in machine learning.**

Interpolation can improve data quality but may introduce bias if the assumptions are incorrect.

42. **What are outliers in a dataset?**

Outliers are data points that significantly differ from other observations, potentially due to variability or errors.

43. **Explain the impact of outliers on machine learning models.**

Outliers can skew model training, leading to poor performance and inaccurate predictions.

44. **Discuss techniques for identifying outliers.**

Techniques include Z-score, IQR (Interquartile Range), and visualization methods like box plots.

45. **How can outliers be handled in a dataset?**

Outliers can be handled by removing them, transforming the data, or using robust algorithms that are less sensitive to outliers.

46. **Compare and contrast Filter, Wrapper, and Embedded methods for feature selection.**

- **Filter**: Selects features based on statistical measures (e.g., correlation).
- **Wrapper**: Uses a model to evaluate feature subsets.
- **Embedded**: Feature selection is part of the model training process.

47. Provide examples of algorithms associated with each method.

- **Filter**: Chi-square test, Pearson correlation.
- **Wrapper**: Recursive Feature Elimination (RFE).
- **Embedded**: LASSO, Ridge Regression.

48. Discuss the advantages and disadvantages of each feature selection method.

- **Filter**: Fast but may ignore feature interactions.
- **Wrapper**: Considers feature interactions but computationally expensive.
- **Embedded**: Efficient but specific to the model used.

49. Explain the concept of feature scaling.

Feature scaling standardizes the range of independent variables, ensuring that no feature dominates due to its scale.

50. Describe the process of standardization.

Standardization transforms data to have a mean of 0 and a standard deviation of 1.

51. How does mean normalization differ from standardization?

Mean normalization scales data to a range of [-1, 1], while standardization scales data to have a mean of 0 and a standard deviation of 1.

52. Discuss the advantages and disadvantages of Min-Max scaling.

- **Advantages**: Preserves the shape of the original distribution.

- **Disadvantages**: Sensitive to outliers.

53. What is the purpose of unit vector scaling?

Unit vector scaling normalizes data to have a length of 1, useful for algorithms sensitive to the magnitude of data.

54. Define Principle Component Analysis (PCA).

PCA is a dimensionality reduction technique that transforms data into a set of orthogonal components, reducing the number of features while retaining most of the variance.

55. Explain the steps involved in PCA.

1. Standardize the data.
2. Compute the covariance matrix.
3. Calculate eigenvalues and eigenvectors.
4. Select principal components.
5. Transform the data into the new subspace.

56. Discuss the significance of eigenvalues and eigenvectors in PCA.

Eigenvalues indicate the amount of variance captured by each principal component, while eigenvectors define the direction of the new feature space.

57. How does PCA help in dimensionality reduction?

PCA reduces the number of features by projecting data onto a lower-dimensional space, retaining the most important information.

58. Define data encoding and its importance in machine learning.

Data encoding converts categorical data into numerical format, making it suitable for machine learning algorithms.

59. Explain Nominal Encoding and provide an example.

Nominal encoding assigns a unique number to each category without implying any order. Example: Encoding colors as Red=1, Blue=2, Green=3.

60. Discuss the process of One Hot Encoding.

One Hot Encoding creates binary columns for each category, where only one column is "hot" (1) for each sample.

61. How do you handle multiple categories in One Hot Encoding?

Each category is represented by a separate binary column, and the presence of a category is indicated by a 1 in the corresponding column.

62. Explain Mean Encoding and its advantages.

Mean Encoding replaces categories with the mean of the target variable for that category. Advantages include capturing target information and reducing dimensionality.

63. Provide examples of Ordinal Encoding and Label Encoding.

- **Ordinal Encoding**: Encoding education levels as High School=1, Bachelor=2, Master=3.
- **Label Encoding**: Encoding categories as unique integers, e.g., Dog=1, Cat=2.

64. What is Target Guided Ordinal Encoding and how is it used?

Target Guided Ordinal Encoding orders categories based on the mean of the target variable, useful for ordinal data with a relationship to the target.

65. Define covariance and its significance in statistics.

Covariance measures the relationship between two variables, indicating how they change together. It is significant for understanding variable dependencies.

66. Explain the process of correlation check.

Correlation check involves calculating the correlation coefficient to measure the strength and direction of the relationship between two variables.

67. What is the Pearson Correlation Coefficient?

The Pearson Correlation Coefficient measures the linear relationship between two variables, ranging from -1 (perfect negative) to +1 (perfect positive).

68. How does Spearman's Rank Correlation differ from Pearson's Correlation?

Spearman's Rank Correlation measures the monotonic relationship (not necessarily linear) between two variables, while Pearson's measures linear relationships.

69. Discuss the importance of Variance Inflation Factor (VIF) in feature selection.

VIF measures multicollinearity among features, helping to identify and remove redundant features that can degrade model performance.

70. Define feature selection and its purpose.

Feature selection is the process of selecting the most relevant features for model training, improving performance and reducing overfitting.

71. Explain the process of Recursive Feature Elimination.

Recursive Feature Elimination (RFE) recursively removes the least important features and builds the model until the desired number of features is reached.

72. How does Backward Elimination work?

Backward Elimination starts with all features and iteratively removes the least significant features based on p-values until only significant features remain.

73. Discuss the advantages and limitations of Forward Elimination.

- **Advantages**: Simple, computationally efficient.
- **Limitations**: May not find the optimal feature subset, sensitive to initial feature set.

74. What is feature engineering and why is it important?

Feature engineering involves creating new features or transforming existing ones to improve model performance. It is important because better features lead to better models.

75. Discuss the steps involved in feature engineering.

Steps include data cleaning, feature creation, transformation, scaling, and selection.

76. Provide examples of feature engineering techniques.

Examples include creating interaction terms, polynomial features, and log transformations.

77. How does feature selection differ from feature engineering?

Feature selection involves choosing the best subset of features, while feature engineering involves creating or transforming features.

78. Explain the importance of feature selection in machine learning pipelines.

Feature selection improves model performance, reduces overfitting, and speeds up training by focusing on the most relevant features.

79. Discuss the impact of feature selection on model performance.

Proper feature selection can improve accuracy, reduce training time, and make models more interpretable.

80. How do you determine which features to include in a machine-learning model?

Features can be selected based on domain knowledge, statistical tests, feature importance scores, or automated feature selection techniques.