Machine Learning Assignment two

Regression & Logistic Regression

1. **What is regression analysis?**
   Regression analysis is a statistical method used to model the relationship between a dependent variable and one or more independent variables.

2. **Difference between linear and nonlinear regression:**

   - *Linear regression*: Assumes a straight-line relationship.

   - *Nonlinear regression*: Models complex, curved relationships between variables.

3. **Simple vs. Multiple Linear Regression:**

   - *Simple Linear Regression*: One independent variable.

   - *Multiple Linear Regression*: Two or more independent variables.

4. **How is regression model performance evaluated?**
   Common metrics:

   - R^2 (coefficient of determination)

   - Mean Absolute Error (MAE)

   - Mean Squared Error (MSE)

   - Root Mean Squared Error (RMSE)

5. **What is overfitting in regression models?**
   Overfitting occurs when a model learns noise from the training data, leading to poor performance on unseen data.

6. **What is logistic regression used for?**
   Logistic regression is used for binary or multiclass classification problems (e.g., spam vs. not spam).

7. **Difference between logistic and linear regression:**
    - *Linear*: Predicts continuous values.
    - *Logistic*: Predicts probabilities (between 0 and 1) for classification.
8. **What is the odds ratio in logistic regression?**
   It quantifies how the odds of an event occurring change with a one-unit change in a predictor.
9. **What is the sigmoid function?**
   A mathematical function that maps any real number into a value between 0 and 1: σ(x)=11+e−x\sigma(x) = \frac{1}{1 + e^{-x}}σ(x)=1+e−x1
10. **Performance evaluation of logistic regression:**
    - Accuracy
    - Precision
    - Recall
    - F1-Score
    - ROC-AUC

Decision Trees

11. **What is a decision tree?**
    A flowchart-like structure used for both classification and regression. It splits data based on feature values.

12. **How does a decision tree make predictions?**
    It follows decision nodes based on feature values until it reaches a leaf node (final prediction).

13. **What is entropy in decision trees?**
    A measure of impurity or randomness. Lower entropy means purer data.

14. **What is pruning in decision trees?**
    The process of cutting back branches to reduce overfitting and improve generalization.

15. **How do decision trees handle missing values?**

    - Assign most frequent value
    - Use surrogate splits
    - Use algorithms like CART which can handle missing data

Support Vector Machine (SVM)

16. **What is a Support Vector Machine (SVM)?**
    A supervised ML algorithm used for classification and regression. It finds the optimal hyperplane to separate data.

17. **What is margin in SVM?**
    The distance between the hyperplane and the nearest support vectors. SVM maximizes this margin.

18. **What are support vectors?**
    Data points closest to the separating hyperplane. They influence the decision boundary.

19. **How does SVM handle non-linearly separable data?**
    By using **kernel functions** (e.g., RBF, polynomial) to transform data into higher dimensions where it becomes linearly separable.

20. **Advantages of SVM over other classifiers:**
    - Effective in high-dimensional spaces
    - Works well for both linear and nonlinear problems
    - Robust to overfitting (especially with regularization)

Naïve Bayes

21. **What is the Naïve Bayes algorithm?**
    A probabilistic classifier based on Bayes' Theorem with the "naïve" assumption that features are independent.

22. **Why is it called "Naïve"?**
    Because it assumes all features are independent of each other — an unrealistic but often effective assumption.

23. **Handling continuous and categorical features:**
    - *Categorical*: Uses frequency-based probabilities.
    - *Continuous*: Assumes normal distribution and uses Gaussian probability.

24. **Prior and Posterior in Naïve Bayes:**
    - *Prior*: Probability of the class before seeing the data.
    - *Posterior*: Updated probability after considering the evidence.

25. **What is Laplace smoothing?**
    A technique to handle zero probabilities by adding 1 to all frequency counts.

26. **Can Naïve Bayes be used for regression?**
    It's rarely used for regression. Naïve Bayes is mainly for classification tasks.

27. **Handling missing values in Naïve Bayes:**
    - Impute missing values
    - Ignore missing features during probability calculation

28. **Common applications of Naïve Bayes:**
    - Spam detection
    - Sentiment analysis
    - Text classification
    - Medical diagnosis

29. **Feature independence assumption in Naïve Bayes:**
    Assumes that all features are conditionally independent given the target class, simplifying computation but not always true in real data.

Naïve Bayes & General Machine Learning Concepts

1. **How does Naïve Bayes handle categorical features with a large number of categories?**
   For categorical features with many categories, Naïve Bayes calculates the probability of each category. With large numbers of categories, this can lead to very sparse data. One way to handle this is through **Laplace smoothing** to ensure that probabilities for rare categories don't become zero.

2. **What is the curse of dimensionality, and how does it affect machine learning algorithms?**
   The curse of dimensionality refers to the challenges that arise when dealing with high-dimensional data. As the number of features increases, the volume of the feature space grows exponentially, leading to sparse data and making it difficult for machine learning algorithms to find patterns.

3. **Explain the bias-variance tradeoff and its implications for machine learning models.**

    - *Bias*: Error due to overly simplistic models that fail to capture the complexity of the data.

    - *Variance*: Error due to overly complex models that capture noise and fluctuations in the data.

    - The tradeoff involves balancing between bias (underfitting) and variance (overfitting). A good model has low bias and low variance.

4. **What is cross-validation, and why is it used?**
   Cross-validation is a technique for evaluating the performance of a model by splitting the data into multiple subsets (folds). The model is trained on some folds and tested on others, providing a more reliable estimate of its performance and reducing overfitting.

5. **Explain the difference between parametric and non-parametric machine learning algorithms.**

    - *Parametric*: Assumes a specific form for the model (e.g., linear regression), with a fixed number of parameters.

    - *Non-parametric*: Makes fewer assumptions and allows the model to learn more flexible relationships (e.g., decision trees, KNN).

6. **What is feature scaling, and why is it important in machine learning?**
   Feature scaling is the process of standardizing or normalizing features so that they have

similar ranges or distributions. It's important because many algorithms (like KNN or gradient descent) perform better when features are on similar scales.

7. **What is regularization, and why is it used in machine learning?**
   Regularization is a technique to prevent overfitting by adding a penalty term to the model's loss function, discouraging the model from fitting the noise in the data. Examples include L1 (Lasso) and L2 (Ridge) regularization.

Ensemble Learning & Gradient Descent

8. **Explain the concept of ensemble learning and give an example.**
   Ensemble learning combines multiple models to improve performance. An example is **Random Forest**, which combines multiple decision trees to reduce overfitting and improve generalization.

9. **What is the difference between bagging and boosting?**

    - *Bagging* (Bootstrap Aggregating) involves training multiple models independently and combining their predictions (e.g., Random Forest).

    - *Boosting* sequentially trains models, with each model focusing on correcting the mistakes of the previous one (e.g., AdaBoost, Gradient Boosting).

10. **What is the difference between a generative model and a discriminative model?**

    - *Generative models* model the joint probability P(x,y)P(x, y)P(x,y), learning how the data is generated (e.g., Naïve Bayes).

    - *Discriminative models* model the conditional probability P(y∣x)P(y|x)P(y∣x), directly learning the decision boundary (e.g., Logistic Regression, SVM).

11. **Explain the concept of batch gradient descent and stochastic gradient descent.**

    - *Batch Gradient Descent*: Computes the gradient over the entire dataset, then updates the model parameters.

    - *Stochastic Gradient Descent (SGD)*: Updates the model parameters for each training example, making it faster but noisier.

KNN, Hyperparameter Tuning, and Model Evaluation

12. **What is the K-nearest neighbors (KNN) algorithm, and how does it work?**
    KNN is a simple classification algorithm where a data point is assigned to the class most common among its nearest neighbors in the feature space.

13. **What are the disadvantages of the K-nearest neighbors algorithm?**

    - Computationally expensive, especially with large datasets
    - Sensitive to irrelevant features and the choice of distance metric
    - Struggles with high-dimensional data due to the curse of dimensionality

14. **Explain the concept of one-hot encoding and its use in machine learning.**
    One-hot encoding is a method for converting categorical variables into binary vectors where each category is represented by a 1 in one position and 0s in others. It's used to make categorical data compatible with machine learning algorithms.

15. **What is feature selection, and why is it important in machine learning?**
    Feature selection involves selecting the most relevant features for model training. It's important because it reduces the complexity of the model, prevents overfitting, and improves computational efficiency.

16. **Explain the concept of cross-entropy loss and its use in classification tasks.**
    Cross-entropy loss is a measure of the difference between the true distribution and the predicted distribution. It's commonly used for classification tasks, particularly in logistic regression and neural networks.

17. **What is the difference between batch learning and online learning?**

    - *Batch Learning*: The model is trained on the entire dataset at once.
    - *Online Learning*: The model is trained incrementally, one data point or a small batch at a time.

Hyperparameter Tuning & Regularization

18. **Explain the concept of grid search and its use in hyperparameter tuning.**
    Grid search is a method to exhaustively search through a manually specified subset of hyperparameters to find the best-performing combination.

19. **What are the advantages and disadvantages of decision trees?**

    - *Advantages*: Easy to understand, interpret, and visualize. Can handle both categorical and numerical data.

    - *Disadvantages*: Prone to overfitting, especially with deep trees. Sensitive to small changes in the data.

20. **What is the difference between L1 and L2 regularization?**

    - *L1 Regularization (Lasso)*: Adds the absolute value of coefficients to the loss function, encouraging sparsity (many coefficients become zero).

    - *L2 Regularization (Ridge)*: Adds the squared value of coefficients, encouraging smaller coefficients but not necessarily zero.

21. **What are some common preprocessing techniques used in machine learning?**
    Common preprocessing techniques include:

    - Normalization and standardization

    - Handling missing data (imputation or removal)

    - One-hot encoding for categorical variables

    - Feature scaling

    - Data augmentation

22. **What is the difference between a parametric and non-parametric algorithm? Give examples of each.**

    - *Parametric*: Assumes a specific form for the model, such as linear regression.

    - *Non-parametric*: Does not assume a specific form, such as decision trees or KNN.

23. **Explain the bias-variance tradeoff and how it relates to model complexity.**
    More complex models tend to have low bias but high variance (overfitting), while simpler models have high bias and low variance (underfitting). Finding the right balance improves generalization.

24. **What are the advantages and disadvantages of using ensemble methods like random forests?**

    - *Advantages*: Better generalization, reduced overfitting, handles high-dimensional data well.

    - *Disadvantages*: Computationally expensive, harder to interpret.

25. **What is the purpose of hyperparameter tuning in machine learning?**
    Hyperparameter tuning optimizes the model's performance by selecting the best combination of hyperparameters, such as learning rate, number of layers, or tree depth.

26. **What is the difference between regularization and feature selection?**

    - *Regularization*: Adds a penalty to the loss function to avoid overfitting.

    - *Feature Selection*: Involves choosing a subset of features that are most relevant to the model.

27. **How does the Lasso (L1) regularization differ from Ridge (L2) regularization?**

    - *Lasso (L1)*: Can set some coefficients to zero, leading to sparse models.

    - *Ridge (L2)*: Shrinks coefficients but doesn't set them to zero, leading to less sparse models.

Cross-Validation & Evaluation Metrics

1. **Explain the concept of cross-validation and why it is used.**
   Cross-validation is a technique used to evaluate the performance of a machine learning model by dividing the dataset into multiple subsets or folds. The model is trained on some folds and tested on the remaining fold. This process is repeated for each fold, and the results are averaged to give a more reliable estimate of model performance. It helps reduce the risk of overfitting and ensures that the model generalizes well on unseen data.

2. **What are some common evaluation metrics used for regression tasks?**
   Common metrics for regression tasks include:

   - **Mean Absolute Error (MAE):** The average of absolute differences between predicted and actual values.
   - **Mean Squared Error (MSE):** The average of squared differences between predicted and actual values.
   - **Root Mean Squared Error (RMSE):** The square root of MSE.
   - **R-squared (R^2):** The proportion of variance in the dependent variable explained by the model.

Naïve Bayes Algorithm

3. **How does the Naïve Bayes algorithm handle categorical features?**
   Naïve Bayes handles categorical features by calculating the probability of each category within a feature for each class. The algorithm assumes that the features are conditionally independent, which simplifies the computation of the joint probability of the data.

4. **Explain the concept of prior and posterior probabilities in Naïve Bayes.**

   - **Prior Probability:** The probability of a class before seeing any data. It's calculated by dividing the number of instances of a class by the total number of instances.

   - **Posterior Probability:** The probability of a class given the observed data, calculated using Bayes' theorem.

5. **What is Laplace smoothing, and why is it used in Naïve Bayes?**
   Laplace smoothing is used to handle zero probabilities in the Naïve Bayes algorithm. When a feature category does not appear in the training data for a particular class, it would normally lead to a zero probability, making the entire probability of that class zero. Laplace smoothing adds a small constant (usually 1) to every feature count to prevent zero probabilities.

6. **Can Naïve Bayes handle continuous features?**
   Yes, Naïve Bayes can handle continuous features by assuming a probability distribution (often Gaussian) for the continuous data. The probability of a continuous feature value is calculated based on the distribution's parameters (mean and variance).

7. **What are the assumptions of the Naïve Bayes algorithm?**

   - **Conditional independence:** Each feature is assumed to be independent of all other features, given the class label.

   - **The feature distributions:** The algorithm assumes that features are distributed according to a known probability distribution (e.g., Gaussian for continuous features, multinomial for discrete features).

8. **How does Naïve Bayes handle missing values?**
   Naïve Bayes can handle missing values by ignoring instances where features are missing or imputing missing values using the mean (for continuous data) or mode (for categorical data) of that feature.

9. **What are some common applications of Naïve Bayes?**
   Naïve Bayes is commonly used in:

   - Text classification (e.g., spam email detection).
   - Sentiment analysis.
   - Document categorization.
   - Medical diagnosis.

Naïve Bayes Classifiers and Decision Boundaries

10. **Explain the difference between generative and discriminative models.**

    - **Generative Models:** These models try to model the distribution of individual classes and generate new data samples (e.g., Naïve Bayes).

    - **Discriminative Models:** These models focus on the boundary between classes and try to directly model the conditional probability of the class given the features (e.g., Logistic Regression, SVM).

11. **How does the decision boundary of a Naïve Bayes classifier look like for binary classification tasks?**
    In binary classification, the decision boundary of Naïve Bayes is typically a linear boundary, as it relies on comparing the posterior probabilities of each class. The boundary is determined by the likelihood ratios of the classes, and it is affected by the distributions of the features.

12. **What is the difference between multinomial Naïve Bayes and Gaussian Naïve Bayes?**

    - **Multinomial Naïve Bayes** is used when features are discrete (e.g., text data with word counts) and models the count distribution of features.

    - **Gaussian Naïve Bayes** is used when features are continuous and assumes that the features follow a Gaussian (normal) distribution.

13. **How does Naïve Bayes handle numerical instability issues?**
    Numerical instability in Naïve Bayes can occur due to very small probability values multiplying together, leading to underflow. This can be handled by using **logarithms** of probabilities, as they turn multiplications into additions, making the computation more stable.

14. **What is the Laplacian correction, and when is it used in Naïve Bayes?**
    The Laplacian correction (also known as Laplace smoothing) is used to prevent the occurrence of zero probabilities for categorical features that do not appear in the training data. It is typically applied when calculating the probabilities for categorical features.

15. **Can Naïve Bayes be used for regression tasks?**
    While Naïve Bayes is primarily used for classification, it can be adapted for regression by using Gaussian Naïve Bayes, where the output is treated as a continuous variable and a normal distribution is assumed.

16. **Explain the concept of conditional independence assumption in Naïve Bayes.**
    The **conditional independence assumption** is the key assumption in Naïve Bayes that says features are independent of each other, given the class label. This assumption simplifies the calculation of the joint probability distribution of the features.

Challenges and Limitations of Naïve Bayes

17. **How does Naïve Bayes handle categorical features with a large number of categories?**
    Naïve Bayes can handle categorical features with a large number of categories, but this can lead to sparse data. The model can still compute probabilities for each category but may struggle with performance if the number of categories is extremely large. Techniques like **feature hashing** or **binning** can help address this issue.

18. **What are some drawbacks of the Naïve Bayes algorithm?**

    - Assumes conditional independence of features, which may not always hold true in practice.
    - Struggles with highly correlated features.
    - Performance may suffer when categorical data has a large number of levels or features have complex dependencies.

19. **Explain the concept of smoothing in Naïve Bayes.**
    Smoothing in Naïve Bayes refers to techniques (like Laplace smoothing) used to handle zero probabilities by adding a small constant to all feature counts, ensuring that no probability is exactly zero.

20. **How does Naïve Bayes handle imbalanced datasets?**
    Naïve Bayes may perform poorly on imbalanced datasets, as it could be biased towards the majority class. Techniques like **class weighting** or **resampling** (under-sampling or over-sampling) can help mitigate this issue.