## Date: February 22, 2024

## SMART INTERNZ - APSCHE
## AI / ML Training

## Assessment 2

**1. In logistic regression, what is the logistic function (sigmoid function) and how is it used to compute probabilities?**

The logistic function, also known as the sigmoid function, is a mathematical function represented as ( σ(z) =1/(1 + e^(-z)), where z  is the linear combination of input features and their corresponding weights. In logistic regression, this function is used to map the output of a linear equation to a probability score between 0 and 1. It transforms the output of the linear equation into a probability of belonging to a certain class.

**2. When constructing a decision tree, what criterion is commonly used to split nodes, and how is it calculated?**

The criterion commonly used to split nodes in constructing a decision tree is typically the information gain or Gini impurity. Information gain measures the reduction in entropy or disorder in the data after a particular split, while Gini impurity measures the probability of incorrectly classifying a randomly chosen element if it were randomly labelled according to the distribution of labels in the node.

**3. Explain the concept of entropy and information gain in the context of decision tree construction.**

Entropy is a measure of impurity or disorder in a dataset. In decision tree construction, entropy is used to quantify the uncertainty in the data at a given node. Information gain is the measure of the effectiveness of a particular attribute in classifying the data points. It is calculated as the difference between the entropy of the parent node and the weighted average of the entropies of the child nodes after a split.

**4. How does the random forest algorithm utilize bagging and feature randomization to improve classification accuracy?**

The random forest algorithm utilizes bagging (bootstrap aggregation) to improve classification accuracy. It constructs multiple decision trees on different subsets of the training data and aggregates their predictions. Feature randomization is also employed, where only a random subset of features is considered for splitting at each node of the decision tree. This randomness helps to reduce overfitting and increases the robustness of the model.

**5. What distance metric is typically used in k-nearest neighbours (KNN) classification, and how does it impact the algorithm's performance?**

The typical distance metric used in k-nearest neighbours (KNN) classification is Euclidean distance. However, other distance metrics such as Manhattan distance or Minkowski distance can also be used. The choice of distance metric impacts the algorithm's performance by determining how "near" or "far" points are in the feature space.

**6. Describe the Naïve-Bayes assumption of feature independence and its implications for classification.**

The Naïve-Bayes assumption of feature independence assumes that the features used for classification are conditionally independent given the class label. This assumption simplifies the calculation of probabilities in the model. Despite its simplicity, Naïve-Bayes classifiers often perform well in practice, especially when the assumption approximately holds true.

**7. In SVMs, what is the role of the kernel function, and what are some commonly used kernel functions?**

In Support Vector Machines (SVMs), the kernel function is used to map the input data into a higher-dimensional space where it becomes easier to find a hyperplane that separates the classes. Commonly used kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid kernels. These kernels allow SVMs to handle non-linear decision boundaries effectively.

**8. Discuss the bias-variance tradeoff in the context of model complexity and overfitting.**

The bias-variance tradeoff refers to the balance between bias (error due to overly simplistic assumptions) and variance (error due to sensitivity to fluctuations in the training data) in model performance. Increasing model complexity typically reduces bias but increases variance, leading to overfitting. Conversely, reducing model complexity increases bias but decreases variance, leading to underfitting. Finding the right balance is crucial for avoiding overfitting and achieving optimal model performance.

**9. How does TensorFlow facilitate the creation and training of neural networks?**

TensorFlow facilitates the creation and training of neural networks by providing a comprehensive framework for building and optimizing computational graphs. It offers a wide range of tools and functionalities for defining neural network architectures, handling data, implementing optimization algorithms, and deploying models on various platforms.

**10. Explain the concept of cross-validation and its importance in evaluating model performance.**

Cross-validation is a technique used to evaluate model performance by splitting the dataset into multiple subsets (folds), training the model on some folds, and evaluating it on the remaining fold(s). This process is repeated multiple times, rotating which fold is used for evaluation. Cross-validation helps to estimate how well a model will generalize to new, unseen data and provides a more robust evaluation of its performance compared to a single train-test split.

**11. What techniques can be employed to handle overfitting in machine learning models?**

Techniques for handling overfitting in machine learning models include:

- Regularization: Introducing a penalty term to the loss function to discourage complex models.
- Cross-validation: Using techniques like k-fold cross-validation to assess model performance on multiple subsets of the data.
- Feature selection: Selecting only the most relevant features to reduce model complexity.
- Early stopping: Monitoring the model's performance on a validation set and stopping training when performance starts to degrade.
- Ensemble methods: Combining multiple models to reduce individual model variance.

**12. What is the purpose of regularization in machine learning, and how does it work?**

Regularization in machine learning is a technique used to prevent overfitting by adding a penalty term to the loss function that penalises large coefficients or complex model structures. Regularization encourages simpler models that generalize better to new data. Common regularization techniques include L1 regularization (Lasso), L2 regularization (Ridge), and Elastic Net regularization.

**13. Describe the role of hyper-parameters in machine learning models and how they are tuned for optimal performance.**

Hyperparameters in machine learning models are parameters that are not directly learned from the data but are set prior to model training. These parameters control aspects of the learning process such as model complexity, regularization strength, learning rate, etc. Hyperparameters need to be tuned to optimize model performance, typically through techniques like grid search, random search, or Bayesian optimization.

**14. What are precision and recall, and how do they differ from accuracy in classification evaluation?**

Precision and recall are evaluation metrics used in classification tasks, especially in situations with imbalanced class distributions. Precision measures the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positive predictions among all actual positive instances. Accuracy, on the other hand, measures the overall correctness of the predictions regardless of class imbalance.

**15. Explain the ROC curve and how it is used to visualize the performance of binary classifiers.**

The ROC (Receiver Operating Characteristic) curve is a graphical representation of the performance of a binary classifier at various classification thresholds. It plots the true positive rate (TPR or recall) against the false positive rate (FPR) at different threshold values. The area under the ROC curve (AUC) is a commonly used metric to quantify the overall performance of the classifier, with a higher AUC indicating better discrimination between the classes.