1. What is the definition of a target function? In the sense of a real-life example, express the target function. How is a target function's fitness assessed?

A target function maps inputs to outputs in a predictive model. For example, predicting house prices based on features like size and location. Fitness is assessed by performance metrics like accuracy, RMSE, or R-squared.

2. What are predictive models, and how do they work? What are descriptive types, and how do you use them? Examples of both types of models should be provided. Distinguish between these two forms of models.

Predictive models forecast outcomes using historical data (e.g., linear regression). Descriptive models summarize data patterns (e.g., clustering). Predictive models focus on future predictions, while descriptive models provide insights into existing data.

3. Describe the method of assessing a classification model's efficiency in detail. Describe the various measurement parameters.

Efficiency is assessed using metrics like accuracy, precision, recall, F1 score, and AUC-ROC. These parameters evaluate different aspects like true positives, false positives, and overall model performance.

4. i. In the sense of machine learning models, what is underfitting? What is the most common reason for underfitting?

Underfitting occurs when a model is too simple to capture data patterns, often due to insufficient complexity or training data.

ii. What does it mean to overfit? When is it going to happen?

Overfitting happens when a model learns noise and details in the training data, leading to poor generalization to new data. It occurs with overly complex models or insufficient data.

iii. In the sense of model fitting, explain the bias-variance trade-off.

The bias-variance trade-off balances model simplicity (bias) and complexity (variance). High bias leads to underfitting, while high variance leads to overfitting. Optimal models minimize both.

5. Is it possible to boost the efficiency of a learning model? If so, please clarify how.

Yes, efficiency can be boosted by techniques like feature selection, hyperparameter tuning, regularization, and using ensemble methods like boosting and bagging.

6. How would you rate an unsupervised learning model's success? What are the most common success indicators for an unsupervised learning model?

Success is rated by metrics like silhouette score, Davies-Bouldin index, and purity. These indicators evaluate clustering quality and separation.

7. Is it possible to use a classification model for numerical data or a regression model for categorical data with a classification model? Explain your answer.

Classification models aren't suitable for numerical predictions; regression models aren't for categorical data. They require converting data types, but results may be suboptimal.

8. Describe the predictive modeling method for numerical values. What distinguishes it from categorical predictive modeling?

Predictive modeling for numerical values uses regression techniques, focusing on continuous outcomes. Categorical predictive modeling uses classification techniques for discrete outcomes.

9. The following data were collected when using a classification model to predict the malignancy of a group of patients' tumors: i. Accurate estimates – 15 cancerous, 75 benign ii. Wrong predictions – 3 cancerous, 7 benign Determine the model's error rate, Kappa value, sensitivity, precision, and F-measure.

Error rate = (3+7)/100 = 10%

Kappa = (Observed - Expected accuracy)/(1 - Expected accuracy)

Sensitivity = 15/(15+3) = 83.33%

Precision = 15/(15+7) = 68.18%

F-measure = 2(PrecisionSensitivity)/(Precision+Sensitivity) ≈ 75%

10. Make quick notes on: 1. The process of holding out 2. Cross-validation by tenfold 3. Adjusting the parameters

1. Holding out: Splitting data into training and test sets to evaluate model performance.

2. Tenfold Cross-validation: Dividing data into 10 parts, training on 9, testing on 1, and repeating.

3. Adjusting parameters: Fine-tuning model hyperparameters to optimize performance.

11. Define the following terms: 1. Purity vs. Silhouette width 2. Boosting vs. Bagging 3. The eager learner vs. the lazy learner

1. Purity vs. Silhouette width: Purity measures clustering homogeneity; silhouette width assesses clustering quality by comparing intra-cluster and inter-cluster distances.

2. Boosting vs. Bagging: Boosting sequentially improves weak models; bagging combines multiple models in parallel to reduce variance.

3. Eager learner vs. Lazy learner: Eager learners build a model before prediction (e.g., decision trees); lazy learners delay processing until a query is made (e.g., k-NN).