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?

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.

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

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

reason for underfitting?

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

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

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

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

success indicators for an unsupervised learning model?

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.

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

categorical predictive modeling?

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.

10. Make quick notes on:

1. The process of holding out

2. Cross-validation by tenfold

3. Adjusting the parameters

11. Define the following terms:

1. Purity vs. Silhouette width

2. Boosting vs. Bagging

3. The eager learner vs. the lazy learner

### **1. Definition of a Target Function and Its Assessment**

**Target Function:**

* **Definition:** A target function in machine learning is the function that maps input features to the output labels or predictions. It represents the true relationship between the input variables and the output variable that the model aims to approximate.
* **Real-life Example:** In a housing price prediction model, the target function might map features like the number of bedrooms, location, and square footage to the house price. The target function describes how these features influence the price.
* **Fitness Assessment:** The fitness of a target function is assessed by evaluating how well the machine learning model approximates this function. This is typically done using metrics like accuracy, precision, recall, F1-score, and mean squared error (MSE) on a test dataset. Cross-validation techniques can also be used to assess the model's performance and generalizability.

### **2. Predictive vs. Descriptive Models**

**Predictive Models:**

* **Definition:** Predictive models are used to forecast future outcomes based on historical data. They make predictions about unknown or future events by learning patterns from past data.
* **How They Work:** Predictive models learn from historical data by identifying relationships between input features and target variables. They use these relationships to make predictions on new, unseen data.
* **Examples:**
  + **Regression Model:** Predicting house prices based on features such as location and size.
  + **Classification Model:** Predicting whether an email is spam or not based on its content.

**Descriptive Models:**

* **Definition:** Descriptive models summarize and describe patterns or relationships in data without making predictions. They help in understanding the underlying structure and characteristics of the data.
* **How They Work:** Descriptive models analyze data to identify patterns, trends, and correlations. They provide insights into the data's distribution and relationships.
* **Examples:**
  + **Clustering:** Grouping customers based on purchasing behavior to identify different market segments.
  + **Association Rules:** Discovering frequent itemsets in transaction data (e.g., market basket analysis).

**Distinction:**

* **Predictive Models:** Focus on forecasting future outcomes based on patterns learned from past data.
* **Descriptive Models:** Focus on summarizing and understanding existing data without making predictions.

### **3. Assessing a Classification Model's Efficiency**

**Method:**

1. **Confusion Matrix:** A table used to describe the performance of a classification model. It shows the counts of true positives, true negatives, false positives, and false negatives.
2. **Measurement Parameters:**
   * **Accuracy:** The proportion of correctly classified instances out of all instances.  
     Accuracy=TP+TNTP+TN+FP+FN\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}Accuracy=TP+TN+FP+FNTP+TN​
   * **Precision:** The proportion of true positive predictions among all positive predictions made by the model.  
     Precision=TPTP+FP\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}Precision=TP+FPTP​
   * **Recall (Sensitivity):** The proportion of true positive predictions among all actual positives.  
     Recall=TPTP+FN\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}Recall=TP+FNTP​
   * **F-Measure:** The harmonic mean of precision and recall.  
     F-Measure=2×Precision×RecallPrecision+RecallF\text{-Measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}F-Measure=2×Precision+RecallPrecision×Recall​
   * **Specificity:** The proportion of true negative predictions among all actual negatives.  
     Specificity=TNTN+FP\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}Specificity=TN+FPTN​

### **4. Underfitting, Overfitting, and Bias-Variance Trade-Off**

**i. Underfitting:**

* **Definition:** Underfitting occurs when a model is too simple to capture the underlying patterns in the data. It results in poor performance on both the training and test datasets.
* **Common Reason:** The most common reason for underfitting is having an overly simplistic model or insufficient features to represent the data adequately.

**ii. Overfitting:**

* **Definition:** Overfitting occurs when a model is too complex and captures noise or random fluctuations in the training data. It performs well on the training data but poorly on unseen data due to poor generalization.
* **When It Happens:** Overfitting typically occurs when the model has too many parameters or when there is insufficient data for training.

**iii. Bias-Variance Trade-Off:**

* **Definition:** The bias-variance trade-off is the balance between two types of errors in a model:
  + **Bias:** Error due to overly simplistic assumptions in the learning algorithm, leading to underfitting.
  + **Variance:** Error due to excessive complexity in the model, leading to overfitting.
* **Trade-Off:** Reducing bias typically increases variance, and reducing variance typically increases bias. The goal is to find a balance that minimizes the total error.

### **5. Boosting the Efficiency of a Learning Model**

**Methods to Boost Efficiency:**

1. **Hyperparameter Tuning:** Adjusting the hyperparameters of the model to improve performance.
2. **Feature Engineering:** Creating and selecting relevant features to enhance model performance.
3. **Ensemble Methods:** Combining predictions from multiple models to improve accuracy (e.g., boosting, bagging).
4. **Regularization:** Applying techniques to prevent overfitting and improve generalization (e.g., L1/L2 regularization).
5. **Cross-Validation:** Using techniques like k-fold cross-validation to assess and refine model performance.

### **6. Rating an Unsupervised Learning Model's Success**

**Success Indicators:**

1. **Cluster Quality Metrics:**
   * **Silhouette Score:** Measures how similar an instance is to its own cluster compared to other clusters.
   * **Davies-Bouldin Index:** Measures the average similarity ratio of each cluster with its most similar cluster.
2. **Intra-cluster vs. Inter-cluster Distances:** Lower intra-cluster distances and higher inter-cluster distances indicate better clustering.
3. **Domain-Specific Metrics:** Depending on the application, such as the usefulness of clusters in customer segmentation.

### **7. Classification vs. Regression Models**

**Classification Model for Numerical Data:**

* **Possibility:** Classification models are generally used for categorical data, but they can handle numerical data by discretizing it into categories (e.g., age ranges).

**Regression Model for Categorical Data:**

* **Possibility:** Regression models are typically used for continuous data, but logistic regression can handle categorical outcomes by modeling the probability of each category.

### **8. Predictive Modeling for Numerical Values**

**Predictive Modeling for Numerical Values:**

* **Definition:** Predictive modeling for numerical values involves predicting a continuous outcome based on input features. This is typically done using regression techniques.

**Distinction from Categorical Predictive Modeling:**

* **Numerical Predictive Modeling:** Involves predicting continuous variables (e.g., house prices, temperature).
* **Categorical Predictive Modeling:** Involves predicting discrete categories (e.g., spam or not spam, disease presence).

### **9. Calculations Based on Classification Results**

**Data:**

* **Accurate Estimates:** 15 cancerous, 75 benign
* **Wrong Predictions:** 3 cancerous, 7 benign

**Metrics:**

* **Error Rate:**Error Rate=False Positives+False NegativesTotal Predictions=3+715+75+3+7=10100=0.10\text{Error Rate} = \frac{\text{False Positives} + \text{False Negatives}}{\text{Total Predictions}} = \frac{3 + 7}{15 + 75 + 3 + 7} = \frac{10}{100} = 0.10Error Rate=Total PredictionsFalse Positives+False Negatives​=15+75+3+73+7​=10010​=0.10
* **Kappa Value:**Observed Agreement (Po)=15+75100=0.90\text{Observed Agreement (Po)} = \frac{15 + 75}{100} = 0.90Observed Agreement (Po)=10015+75​=0.90 Expected Agreement (Pe)=(15+3100)×(15+7100)+(75+7100)×(75+3100)=0.18×0.22+0.82×0.78=0.0396+0.6396=0.6792\text{Expected Agreement (Pe)} = \left(\frac{15 + 3}{100}\right) \times \left(\frac{15 + 7}{100}\right) + \left(\frac{75 + 7}{100}\right) \times \left(\frac{75 + 3}{100}\right) = 0.18 \times 0.22 + 0.82 \times 0.78 = 0.0396 + 0.6396 = 0.6792Expected Agreement (Pe)=(10015+3​)×(10015+7​)+(10075+7​)×(10075+3​)=0.18×0.22+0.82×0.78=0.0396+0.6396=0.6792 Kappa=0.90−0.67921−0.6792=0.22080.3208≈0.688\text{Kappa} = \frac{0.90 - 0.6792}{1 - 0.6792} = \frac{0.2208}{0.3208} \approx 0.688Kappa=1−0.67920.90−0.6792​=0.32080.2208​≈0.688
* **Sensitivity:**Sensitivity=1515+3=1518≈0.833\text{Sensitivity} = \frac{15}{15 + 3} = \frac{15}{18} \approx 0.833Sensitivity=15+315​=1815​≈0.833
* **Precision:**Precision=1515+7=1522≈0.682\text{Precision} = \frac{15}{15 + 7} = \frac{15}{22} \approx 0.682Precision=15+715​=2215​≈0.682
* **F-Measure:**F-Measure=2×Precision×SensitivityPrecision+Sensitivity=2×0.682×0.8330.682+0.833=2×0.5681.515≈0.749F\text{-Measure} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} = 2 \times \frac{0.682 \times 0.833}{0.682 + 0.833} = 2 \times \frac{0.568}{1.515} \approx 0.749F-Measure=2×Precision+SensitivityPrecision×Sensitivity​=2×0.682+0.8330.682×0.833​=2×1.5150.568​≈0.749

### **10. Quick Notes:**

**1. The Process of Holding Out:**

* **Definition:** Splitting the dataset into a training set and a hold-out test set. The model is trained on the training set and evaluated on the hold-out test set to estimate performance on unseen data.

**2. Cross-Validation by Tenfold:**

* **Definition:** A cross-validation technique where the dataset is divided into 10 equal parts (folds). The model is trained on 9 folds and tested on the remaining fold, and this process is repeated 10 times. Results are averaged to estimate model performance.

**3. Adjusting the Parameters:**

* **Definition:** The process of tuning hyperparameters of a model to improve its performance. This can be done through techniques like grid search or random search.

### **11. Definitions:**

**1. Purity vs. Silhouette Width:**

* **Purity:** A measure of the extent to which clusters contain only members of a single class.  
  Purity=1N∑i=1Kmaxj∣Cij∣∣Ci∣\text{Purity} = \frac{1}{N} \sum\_{i=1}^K \text{max}\_{j} \frac{|C\_{ij}|}{|C\_i|}Purity=N1​i=1∑K​maxj​∣Ci​∣∣Cij​∣​
* **Silhouette Width:** A measure of how similar an instance is to its own cluster compared to other clusters. Values range from -1 to 1.

**2. Boosting vs. Bagging:**

* **Boosting:** An ensemble method that combines weak learners sequentially, each one correcting the errors of its predecessor.
* **Bagging:** An ensemble method that combines predictions from multiple models trained on different subsets of the data to reduce variance.

**3. Eager Learner vs. Lazy Learner:**

* **Eager Learner:** Learns a model during training time and makes predictions quickly during testing (e.g., decision trees, neural networks).
* **Lazy Learner:** Delays the learning process until a prediction is needed, often storing training data and using it to make predictions (e.g., k-nearest neighbors).