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, in machine learning, is a method for solving a problem that an AI algorithm parses its training data to find. Once an algorithm finds its target function, that function can be used to predict results (predictive analysis).

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 modeling is a commonly used statistical technique to predict future behavior. Predictive modeling solutions are a form of data-mining technology that works by analyzing historical and current data and generating a model to help predict future outcomes. In predictive modeling, data is collected, a statistical model is formulated, predictions are made, and the model is validated (or revised) as additional data becomes available.

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

How to Best Evaluate a Classification Model

* Classification accuracy.
* Confusion matrix.
* Precision and recall.
* F1 score.
* Sensitivity and specificity.
* ROC curve and AUC.

4.

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

Underfitting occurs when our machine learning model is not able to capture the underlying trend of the data. To avoid the overfitting in the model, the fed of training data can be stopped at an early stage, due to which the model may not learn enough from the training data. As a result, it may fail to find the best fit of the dominant trend in the data. In the case of underfitting, the model is not able to learn enough from the training data, and hence it reduces the accuracy and produces unreliable predictions.

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

Overfitting is a concept in data science, which occurs when a statistical model fits exactly against its training data. When this happens, the algorithm unfortunately cannot perform accurately against unseen data, defeating its purpose.

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

There is a tradeoff between a model's ability to minimize bias and variance. Gaining a proper understanding of these errors would help us not only to build accurate models but also to avoid the mistake of overfitting and underfitting.

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

Boosting is an ensemble learning method that combines a set of weak learners into a strong learner to minimize training errors. In boosting, a random sample of data is selected, fitted with a model and then trained sequentially—that is, each model tries to compensate for the weaknesses of its predecessor.

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

Twin sample validation can be used to validate results of unsupervised learning. It should be used in combination with internal validation. It can prove to be highly useful in case of time-series data where we want to ensure that our results remain same across time.

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 predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

The output variables are often called labels or categories. The mapping function predicts the class or category for a given observation.

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

Predictive modeling is a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data. It works by analyzing current and historical data and projecting what it learns on a model generated to forecast likely 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.

10. Make quick notes on:

1. The process of holding out

The hold-out method for training the machine learning models is a technique that involves splitting the data into different sets: one set for training, and other sets for validation and testing. The hold-out method is used to check how well a machine learning model will perform on the new data.

2. Cross-validation by tenfold

10-fold cross validation would perform the fitting procedure a total of ten times, with each fit being performed on a training set consisting of 90% of the total training set selected at random, with the remaining 10% used as a hold out set for validation.

3. Adjusting the parameters

11. Define the following terms:

1. Purity vs. Silhouette width

The main difference between the cluster purity and silhouette width is that the former ignores the intra-cluster variance. This provides a simpler interpretation of cluster separation; a low silhouette width may still occur in well-separated clusters if the internal heterogeneity is high, while no such complication exists for the cluster purity.

2. Boosting vs. Bagging

Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data. Boosting is an iterative technique which adjusts the weight of an observation based on the last classification.

3. The eager learner vs. the lazy learner

A lazy learner delays abstracting from the data until it is asked to make a prediction while an eager learner abstracts away from the data during training and uses this abstraction to make predictions rather than directly compare queries with instances in the dataset.