**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?**

In machine learning, a target function is a mathematical function that maps input data to an output value. It's also known as an "objective function" or "cost function"

The target function is a method for solving a problem that an AI algorithm uses to find. Once the algorithm finds the target function, it can be used to predict results

The target function is a mathematical representation of the desired outcome or the performance metric used to measure the algorithm's performance. The specific form of the target function depends on the type of machine learning task being performed.

In the sense of a real-life example, the target function can be expressed as the formula that an algorithm feeds data to in order to calculate predictions. For example, if you are trying to train a machine learning algorithm to predict the price of a house, the target function would be the actual price of the house.

The fitness of a target function is assessed by how well it predicts the output values for the given input data. The goal is to find the target function that minimizes the error between the predicted and actual values.

**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 mathematical process used to predict future events or outcomes by analyzing patterns in a given set of input data.

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.

Descriptive Analytics tells you what happened in the past. Diagnostic Analytics helps you understand why something happened in the past. Predictive Analytics predicts what is most likely to happen in the future.

Descriptive data mining is used to summarize and describe the data, while predictive data mining is used to make predictions about future events. Both techniques have their own advantages and applications, and the choice of technique depends on the specific problem and the nature of the data

Descriptive Analysis: Before using a machine learning algorithm, it is very important to acquire abstract knowledge of the problem. The goal of descriptive analysis is to find an accurate understanding of the problem by asking questions from historical data. Let’s understand the descriptive analysis process using an example. Suppose your task is to optimize the supply chain of a department store, for this task we have [purchase and sales data](https://www.kaggle.com/aungpyaeap/supermarket-sales). After analyzing the data, we make assumptions that sales increase during the day just before the weekend. This means that our machine learning model is based on periodicity. So, descriptive analysis helps us understand the deep patterns from the data to uncover all those special features that were overlooked at the initial stage.

In short, the purpose of descriptive analysis is to enable us to understand whether the machine learning model will perform poorly or whether it is the best model in a particular problem.

Predictive Analysis: Predictive analytics is an important concept in machine learning. What happens is that once we have formed a machine learning model based on descriptive analysis, the next goal is to infer its future steps by giving some initial conditions. Predictive analytics is used to discover and define certain rules that underlie a process for pushing a particular condition on time. For example, the object detector of a [self-driven car](https://thecleverprogrammer.com/2020/10/25/the-machine-learning-behind-self-driving-cars/) can be extremely precise at detecting an obstacle in time, but another model must take action that minimizes the risk of damage and maximizes the likelihood of safe movement.  
Predictive analytics, therefore, means observing a problem in time and taking the most appropriate action as a prescription to avoid any type of risk.

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

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

if the model is performing poorly over the test and the train set, then we call that an underfitting model. An example of this situation would be building a linear regression model over non-linear data

It occurs when a model is too simple, which can be a result of a model needing more training time, more input features, or less regularization.

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

Overfitting is a modeling error in statistics that occurs when a function is too closely aligned to a limited set of data points. As a result, the model is useful in reference only to its initial data set, and not to any other data sets.

overfitting occurs when your model is too complex for your data. Overfitting happens due to several reasons, such as: The training data size is too small and does not contain enough data samples to accurately represent all possible input data values.

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

Bias is the simplifying assumptions made by the model to make the target function easier to approximate. Variance is the amount that the estimate of the target function will change given different training data. Trade-off is tension between the error introduced by the bias and the variance.

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

There are several ways to increase the accuracy of a regression model, such as collecting more data, relevant feature selection, feature scaling, regularization, cross-validation, hyperparameter tuning, adjusting the learning rate, and ensemble methods like bagging, boosting, and stacking.

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

Evaluating the performance of unsupervised learning algorithms can be challenging since there is no clear objective function to optimize. However, there are several methods for evaluating the performance of unsupervised learning algorithms, such as:

1. Visualization: One way to evaluate the performance of an unsupervised learning algorithm is to visualize the results. For example, if the algorithm is clustering data points, you can plot the clusters and see if they make sense.
2. Silhouette score: The silhouette score is a metric that measures how similar an object is to its own cluster compared to other clusters. It ranges from -1 to 1, where 1 indicates that the object is well-matched to its own cluster and poorly-matched to neighboring clusters.
3. Elbow method: The elbow method is used to determine the optimal number of clusters in a dataset. It involves plotting the explained variation as a function of the number of clusters and selecting the number of clusters where the change in explained variation begins to level off.
4. Reconstruction error: If the unsupervised learning algorithm involves dimensionality reduction or feature extraction, the reconstruction error can be used to evaluate its performance. The reconstruction error measures how well the algorithm can reconstruct the original data from the reduced feature space.
5. Domain-specific evaluation metrics: Depending on the application, domain-specific evaluation metrics may be used to evaluate the performance of unsupervised learning algorithms. For example, in anomaly detection, the false positive rate and false negative rate may be used to evaluate the performance of the algorithm.

**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.**

Categorical Data is the data that generally takes a limited number of possible values. Also, the data in the category need not be numerical, it can be textual in nature. All machine learning models are some kind of mathematical model that need numbers to work with. This is one of the primary reasons we need to pre-process the categorical data before we can feed it to machine learning models.

If a categorical target variable needs to be encoded for a classification predictive modeling problem, then the LabelEncoder class can be used.

**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.

Classification is the process of identifying the category or class label of the new observation to which it belongs.Predication is the process of identifying the missing or unavailable numerical data for a new observation. That is the key difference between classification and prediction.

**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 machine learning model is the process of splitting the data in different splits and using one split for training the model and other splits for validating and testing the models. The hold-out method is used for both model evaluation and model selection.

**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:** A fancy name for training: the selection of parameter values, which are optimal in some desired sense (eg. minimize an objective function you choose over a dataset you choose). The parameters are the weights and biases of the network.

**11. Define the following terms:**

1. **Purity vs. Silhouette width**: Purity is a measure of the extent to which clusters contain a single class. Its calculation can be thought of as follows: For each cluster, count the number of data points from the most common class in said cluster.

The silhouette width is also an estimate of the average distance between clusters. Its value is comprised between 1 and -1 with a value of 1 indicating a very good cluster.

**2. Boosting vs. Bagging:**

* **Data partition |**whole data vs. bias  
  While bagging uses random bags out of the training data for all models independently, boosting puts higher importance on misclassified data of the upcoming models. Therefore, the data partition is different here.
* **Models | i**ndependent vs. sequences  
  Bagging creates independent models that are aggregated together. However, boosting updates the existing model with the new ones in a sequence. Therefore, the models are affected by previous builds.
* **Goal |**variance vs. bias  
  Another difference is the fact that bagging aims to reduce the variance, but boosting tries to reduce the bias. Therefore, bagging can help to decrease overfitting, and boosting can reduce underfitting.
* **Function |**weighted vs. non-weighted  
  The final function to predict the outcome uses equally weighted average or equally weighted voting aggregations within the bagging technique. Boosting uses weighted majority vote or weighted average functions with more weight to those with better performance on training data.
  1. **The eager learner vs. the lazy learner:** Eager learning is a type of machine learning where the system constructs a generalized model during the training phase, before any queries are made. This approach is in contrast to lazy learning, where the model is not built until a prediction is required.
* Eager methods, like decision trees, require upfront model building, while lazy methods, like k-NN, defer it.