**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.** A **target function**, in machine learning, is a method for solving a problem that an AI algorithm parses its training data to find.The target function is essentially the formula that an algorithm feeds data to in order to calculate predictions.

Analyzing the massive amounts of data related to its given problem, an AI derives understanding of previously unspecified rules by detecting consistencies in the data. The observations of inherent rules about how the studied subject operates inform the AI on how to process future data that does not include an output by applying this previously unknown function.

**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.**

**A.** In short, 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.

The three main types of descriptive studies are **Case studies, Naturalistic observation, and Surveys**.

Some examples of descriptive research are: A specialty food group launching a new range of barbecue rubs would like to understand what flavors of rubs are favored by different people.

Case Studies are a type of observational research that involve a thorough descriptive analysis of a single individual, group, or event. There is no single way to conduct a case study so researchers use a range of methods from unstructured interviewing to direct observation.

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

**A.** Logarithmic loss (or log loss) measures the performance of a classification model where the prediction is a probability value between 0 and 1.

Log loss increases as the predicted probability diverge from the actual label. Log loss is a widely used metric for Kaggle competitions. Input on the most important basics for the measurement of the physical parameters: Temperature, flow velocity, humidity, pressure, CO2 and infrared. Tips on correct measurement and for avoiding measurement errors.

**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.**

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

Underfitting is a scenario in data science where a data model is unable to capture the relationship between the input and output variables accurately, generating a high error rate on both the training set and unseen data.

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

Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model.

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

The bias is known as the difference between the prediction of the values by the ML model and the correct value. Being high in biasing gives a large error in training as well as testing data. By high bias, the data predicted is in a straight line format, thus not fitting accurately in the data in the data set.

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

**A.** Building a machine learning model is not enough to get the right predictions, as you have to check the accuracy and need to validate the same to ensure get the precise results. And validating the model will improve the performance of the ML model.

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

**A.**  In case of supervised learning, it is mostly done by measuring the performance metrics such as accuracy, precision, recall, AUC, etc. on the training set and the holdout sets.

Few examples of such measures are:

* Silhouette coefficient.
* Calisnki-Harabasz coefficient.
* Dunn index.
* Xie-Beni score.
* Hartigan index.

**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.**

**A.**  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?**

**A.**  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.**

* **A. 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.

* **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.

* **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.

**10. Make quick notes on:**

**1. The process of holding out**

**2. Cross-validation by tenfold**

**3. Adjusting the parameters**

**A.**

* **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.
* **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.
* **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.

**11. Define the following terms:**

**1. Purity vs. Silhouette width**

**2. Boosting vs. Bagging**

**3. The eager learner vs. the lazy learner**

**A.**

**1. Purity vs. Silhouette width**

**Silhouette** refers to a method of interpretation and validation of consistency within [clusters of data](https://en.wikipedia.org/wiki/Cluster_analysis). The technique provides a succinct graphical representation of how well each object has been classified.

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

The silhouette can be calculated with any [distance](https://en.wikipedia.org/wiki/Distance) metric, such as the [Euclidean distance](https://en.wikipedia.org/wiki/Euclidean_distance) or the [Manhattan distance](https://en.wikipedia.org/wiki/Manhattan_distance).

**Purity** is a simple and transparent evaluation measure that evaluates how well the clustering matches the gold standard classes ( a set of classes in an evaluation benchmark).

To compute *purity* , each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by $N$. Formally:

|  |  |
| --- | --- |
| \begin{displaymath} \mbox{purity}( \Omega,\mathbb{C} ) = \frac{1}{N} \sum_k \max_j \vert\omega_k \cap c_j\vert \end{displaymath} | (182) |

where $\Omega = \{ \omega_1, \omega_2, \ldots, \omega_K \}$is the set of clusters and $\mathbb{C} = \{ c_1,c_2,\ldots,c_J \}$ is the set of classes. We interpret $\omega_k$ as the set of documents in $\omega_k$ and $c_j$ as the set of documents in $c_j$.

**2. Boosting vs. Bagging :**

| **Bagging** | **Boosting** |
| --- | --- |
| The simplest way of combining predictions that  belong to the same type. | A way of combining predictions that  belong to the different types. |
| Aim to decrease variance, not bias. | Aim to decrease bias, not variance. |
| Each model receives equal weight. | Models are weighted according to their performance. |
| Each model is built independently. | New models are influenced  by the performance of previously built models. |
| Different training data subsets are selected using row sampling with replacement and random sampling methods from the entire training dataset. | Every new subset contains the elements that were misclassified by previous models. |
| Bagging tries to solve the over-fitting problem. | Boosting tries to reduce bias. |
| If the classifier is unstable (high variance), then apply bagging. | If the classifier is stable and simple (high bias) the apply boosting. |
| In this base classifiers are trained parallelly. | In this base classifiers are trained sequentially. |
| Example: The Random forest model uses Bagging. | Example: The AdaBoost uses Boosting techniques |

**3. The eager learner vs. the lazy learner :**

|  |  |
| --- | --- |
| **Eager learner** | **lazy learner** |
| Eager learner generalizes during the training phase. Incorporating the information from new data afterward requires retraining. | **Lazy learner adapts to new data quickly** as it’s during the query phase when it determines which objects to use for prediction. So, all that we need to do to use new data is to store them |
| **eager learner has a training phase in which it uses the entire training dataset**to construct a comprehensive model for prediction. | When it comes to training time, lazy learniner requires less time as the algorithms focus on storing data so that they can be easily retrieved later |
| eager learner has a faster query time as completed models are ready to use and usually produce predictions quickly | **Lazy learner is slower at making predictions** since it searches through data to find the query object’s neighborhood. |