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&#39;s fitness assessed?

A:

Machine learning algorithms are described as learning a target function (f) that best maps input variables (X) to an output variable (Y). Y = f(X) This is a general learning task where we would like to make predictions in the future (Y) given new examples of input variables (X).

Fitness Function (also known as the Evaluation Function) evaluates how close a given solution is to the optimum solution of the desired problem. It determines how fit a solution is.

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:

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.

Image result for What are descriptive types, and how do you
use them?

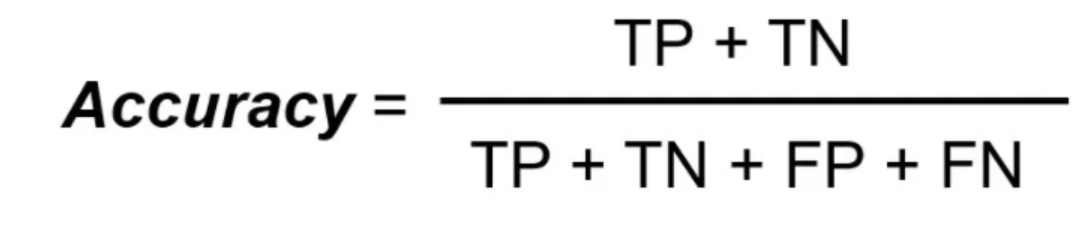
There are 3 main types of descriptive statistics: The distribution concerns the frequency of each value. The central tendency concerns the averages of the values. The variability or dispersion concerns how spread out the values are.

A descriptive model will exploit the past data that are stored in databases and provide you with the accurate report. In a Predictive model, it identifies patterns found in past and transactional data to find risks and future outcomes.

3. Describe the method of assessing a classification model&#39;s efficiency in detail. Describe the various

measurement parameters.

A:



A confusion matrix is defined as thetable that is often used to describe the performance of aclassification model on a set of the test data for which the true values are known.

A picture containing letter

Description automatically generated

Text

Description automatically generated

A picture containing text

Description automatically generated

4.

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

reason for underfitting?

A:

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. Like overfitting, when a model is underfitted, it cannot establish the dominant trend within the data, resulting in training errors and poor performance of the model.

High bias and low variance are good indicators of underfitting.

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

A:

Overfitting occurs when the model cannot generalize and fits too closely to the training dataset instead. 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.

A:

By high bias, the data predicted is in a straight line format, thus not fitting accurately in the data in the data set. Such fitting is known as **Underfitting of Data**.

When a model is high on variance, it is then said to as **Overfitting of Data**.

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

A:

* Reframe the problem. Sometimes, improving a model may have nothing to do with the data or techniques used to train the model
* Provide more data samples
* Add context to the data
* Use meaningful data and features
* Cross-validation
* Hyperparameter tuning
* Choose a different algorithm

6. How would you rate an unsupervised learning model&#39;s success? What are the most common

success indicators for an unsupervised learning model?

A:

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

External validation

* **TP**: Number of pairs of records which are in the same cluster, for both S and P
* **FP**: Number of pairs of records which are in the same cluster in S but not in P
* **FN**: Number of pairs of records which are in the same cluster in P but not in S
* **TN**: Number of pairs of records which are not in the same cluster S as well as P

Diagram

Description automatically generated with medium confidence

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:

* For discrete variable classification model can be used
* . Logistic regression can be used for classification model

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

categorical predictive modelling?

A:

Predictive modeling is a statistical technique in which an organization references known results and historical data to develop predictions for future events. Predictive models analyze patterns and observe trends within specific conditions to determine the most likely outcome. For example, financial organizations may use predictive modeling to gather a customer's credit history and other customer data. They might then use this information to determine an individual's credit score and the likelihood of them submitting credit payments on time. Another name for this technique is predictive analytics, which also includes machine learning and data mining to gather and assess data.

**Some of the more common predictive algorithms are:**

* **Random Forest**: This algorithm is derived from a combination of decision trees, none of which are related, and can use both classification and regression to classify vast amounts of data.
* **Generalized Linear Model (GLM)** for Two Values: This algorithm narrows down the list of variables to find “best fit.” It can work out [tipping points](https://www.merriam-webster.com/dictionary/tipping%20point) and [change data capture](https://www.hvr-software.com/blog/change-data-capture/) and other influences, such as [categorical predictors](https://online.stat.psu.edu/stat462/node/86/), to determine the “best fit” outcome, thereby overcoming drawbacks in other models, such as a regular linear regression.
* **Gradient Boosted Model**: This algorithm also uses several combined decision trees, but unlike Random Forest, the trees are related. It builds out one tree at a time, thus enabling the next tree to correct flaws in the previous tree. It’s often used in rankings, such as on search engine outputs.
* K**-Means:** A popular and fast algorithm, K-Means groups data points by similarities and so is often used for the clustering model. It can quickly render things like personalized retail offers to individuals within a huge group, such as a million or more customers with a similar liking of lined

9. The following data were collected when using a classification model to predict the malignancy of a

group of patients&#39; tumors:

i. Accurate estimates – 15 cancerous, 75 benign

ii. Wrong predictions – 3 cancerous, 7 benign

Determine the model&#39;s error rate, Kappa value, sensitivity, precision, and F-measure.

10. Make quick notes on:

1. The process of holding out

Holdout Method is the simplest sort of method to evaluate a classifier. In this method, the data set (a collection of data items or examples) is separated into two sets, called the Training set and Test set.

A classifier performs function of assigning data items in a given collection to a target category or class.

Example –  
E-mails in our inbox being classified into spam and non-spam.

Classifier should be evaluated to find out, it’s accuracy, error rate, and error estimates. It can be done using various methods. One of most primitive methods in evaluation of classifier is ‘Holdout Method’.

2. Cross-validation by tenfold

With this method we have one data set which we divide randomly into 10 parts. We use 9 of those parts for training and reserve one tenth for testing. We repeat this procedure 10 times each time reserving a different tenth for testing.

3. Adjusting the parameters

Example:

**Hyperparameters**

These are adjustable parameters that must be tuned in order to obtain a model with optimal performance.

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.

2. Boosting vs. Bagging

Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions. Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

3. The eager learner vs. the lazy learner

