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).

he target variable of a dataset is the feature of a dataset about which you want to gain a deeper understanding. A supervised machine learning algorithm uses historical data to learn patterns and uncover relationships between other features of your dataset and the target.

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.

In short, predictive modelling 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 analysing current and historical data and projecting what it learns on a model generated to forecast likely outcomes

Descriptive, or qualitative, methods include the case study, naturalistic observation, surveys, archival research, longitudinal research, and cross-sectional research. Experiments are conducted in order to determine cause-and-effect relationships.

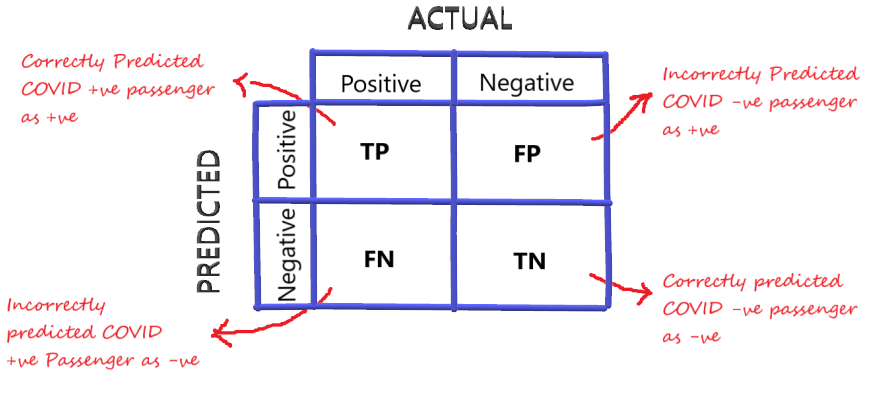
A descriptive model will exploit the past data that are stored in databases and provide you with the accurate report. In a Predictive model, it identifies patterns found in past and transactional data to find risks and future outcomes.

Descriptive Model: Sales report, revenue of a company, performance analysis, etc.

Predictive Model: Sentimental analysis, credit score analysis, forecast reports for a company, etc.

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

Efficiency measure of classification model using Confusion Metrix.



Accuracy = Number of correct predictions / Total number of predictions.

Recall (Sensitivity or True positive rate) = Recall gives the fraction you correctly identified as positive out of all positives.

Precision = Precision gives the fraction of correctly identified as positive out of all predicted as positives.

F1 Score = It is defined as the harmonic mean of the model’s precision and recall.

ROC/AUC Curve =

The receiver operator characteristic is another common tool used for evaluation. It plots out the sensitivity and specificity for every possible decision rule cutoff between 0 and 1 for a model. For classification problems with probability outputs, a threshold can convert probability outputs to classifications. we get the ability to control the confusion matrix a little bit. So by changing the threshold, some of the numbers can be changed in the confusion matrix. But the most important question here is, how to find the right threshold? Of course, we don’t want to look at the confusion matrix every time the threshold is changed, therefore here comes the use of the ROC curve.

For each possible threshold, the ROC curve plots the False positive rate versus the true positive rate.

False Positive Rate: Fraction of negative instances that are incorrectly classified as positive.

True Positive Rate: Fraction of positive instances that are correctly predicted as positive.

4.

i. 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.

ii. 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 mode

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.

Add more data, Treat missing and Outlier values, Feature Engineering, Feature Selection, Multiple algorithms, Algorithm Tuning, Ensemble methods, Cross Validation

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.

NO.

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

Predictive models use known results to develop (or train) a model that can be used to predict values for different or new data. The modeling results in predictions that represent a probability of the target variable (for example, revenue) based on estimated significance from a set of input variables.

There are many different types of predictive modeling techniques including ANOVA, linear regression (ordinary least squares), logistic regression, ridge regression, time series, decision trees, neural networks, and many more.

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= 11%, Kappa value, sensitivity= 83%, precision = 68%, and F-measure=0.75

10. Make quick notes on:

1. The process of holding out

Hold-out is when you split up your dataset into a 'train' and 'test' set. The training set is what the model is trained on, and the test set is used to see how well that model performs on unseen 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

The data analysis computer program must change the parameter values to achieve a minimum value for the weighted sum of the squared residuals (WSS). This can be illustrated by changing the slope and intercept for the equation for a straight line.

11. Define the following terms:

1. Purity vs. Silhouette width

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.