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 functions 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). The target function, also known as the objective function or cost function, is a mathematical function that describes the relationship between the input data and the desired output. The goal of machine learning is to find the values of the model's parameters that minimize the error or difference between the predicted output and the actual output. The target function is used to measure this error or difference.

The target function is specific to the type of machine learning problem being solved. For example, in supervised learning problems, the target function is typically chosen to be the mean squared error (MSE) or the cross-entropy loss. In unsupervised learning problems, the target function is typically chosen to be the sum of squares or the K-L divergence.

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. It is a crucial component of predictive analytics, a type of data analytics which uses current and historical data to forecast activity, behavior and trends.

A descriptive model describes a system or other entity and its relationship to its environment. It is generally used to help specify and/or understand what the system is, what it does, and how it does it. A geometric model or spatial model is a descriptive model that represents geometric and/or spatial relationships.

|  |  |
| --- | --- |
| **Descriptive data mining** | **Predictive data mining** |
| Descriptive mining is usually used to provide correlation, cross-tabulation, frequency, etc. | The term 'Predictive' means to predict something, so predictive data mining is the analysis done to predict the future event or other data or trends. |
| It is based on the reactive approach. | It is based on the proactive approach. |
| It specifies the characteristics of the data in a target data set. | It executes the induction over the current and past data so that prediction can happen. |
| It needs data aggregation and data mining. | It needs statistics and data forecasting procedures. |
| It provides precise data. | It produces outcomes without ensuring accuracy. |

3. Describe the method of assessing a classification model’s efficiency in detail. Describe the various

measurement parameters.

Accuracy

Accuracy is the number of correct predictions made by the model by the total number of records. The best accuracy is 100% indicating that all the predictions are correct.

For an imbalanced dataset, accuracy is not a valid measure of model performance. For a dataset where the default rate is 5%, even if all the records are predicted as 0, the model will still have an accuracy of 95%. But this model will ignore all the defaults and can be very detrimental to the business. So accuracy is not a right measure for model performance in this scenario.

Sensitivity or recall

Sensitivity (Recall or True positive rate) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR).

Specificity

Specificity (true negative rate) is calculated as the number of correct negative predictions divided by the total number of negatives.

Specificity

Precision

Precision (Positive predictive value) is calculated as the number of correct positive predictions divided by the total number of positive predictions.

Precision

KS statistic

KS statistic is a measure of degree of separation between the positive and negative distributions. KS value of 100 indicates that the scores partition the records exactly such that one group contains all positives and the other contains all negatives. In practical situations, a KS value higher than 50% is desirable.

ROC chart & Area under the curve (AUC)

ROC chart is a plot of 1-specificity in the X axis and sensitivity in the Y axis. Area under the ROC curve is a measure of model performance. The AUC of a random classifier is 50% and that of a perfect classifier is 100%. For practical situations, an AUC of over 70% is desirable.

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. 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.

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. Generalization of a model to new data is ultimately what allows us to use machine learning algorithms every day to make predictions and classify data. 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.

If the algorithm is too simple (hypothesis with linear eq.) then it may be on high bias and low variance condition and thus is error-prone. If algorithms fit too complex ( hypothesis with high degree eq.) then it may be on high variance and low bias. In the latter condition, the new entries will not perform well. Well, there is something between both of these conditions, known as Trade-off or Bias Variance Trade-off.

This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can’t be more complex and less complex at the same time.

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

1. Add More Data

Having more data is always a good idea. It allows the “data to tell for itself” instead of relying on assumptions and weak correlations. Presence of more data results in better and more accurate machine-learning models.

I understand we don’t get an option to add more data. For example, we do not get a choice to increase the size of training data in data science competitions. But while working on a real-world company project, I suggest you ask for more data, if possible. This will reduce the pain of working on limited data sets.

2. Treat Missing and Outlier Values

The unwanted presence of missing and outlier values in the training data often reduces the accuracy of a trained model or leads to a biased model. It leads to inaccurate predictions. This is because we don’t analyze the behavior and relationship with other variables correctly. So, it is important to treat missing and outlier values well.

3. Feature Engineering

This step helps to extract more information from existing data. New information is extracted in terms of new features. These features may have a higher ability to explain the variance in the training data. Thus, giving improved model accuracy.

4. Multiple Algorithms

There are many different algorithms in machine learning, but hitting the right machine learning algorithm is the ideal approach to achieve higher accuracy.

5. Algorithm Tuning

We know that machine learning algorithms are driven by hyperparameters. These hyperparameters majorly influence the outcome of the learning process.

The objective of hyperparameter tuning is to find the optimum value for each hyperparameter to improve the accuracy of the model. To tune these hyperparameters, you must have a good understanding of these meanings and their individual impact on the model. You can repeat this process with a number of well-performing models.

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

success indicators for an unsupervised learning model?

Adjusted Rand Index

This is one of the variations of the classic Rand Index that tries to get the proportion of the cluster assignments which are “accurate.” It does so by computing the similarity measure between two different clusterings, taking into consideration all pairs of inputs and numbering those pairs that are assigned in the same or a different cluster, then comparing the same with the random probability of assignment of these clusters. The measuring metric is available in sklearn and can be directly used to create a quantifiable measure for clustering algorithms through the following clustering metrics:

1.) Fowlkes-Mallows Score

It is a bit similar in its results to the ARI as it also attempts to look at the cluster assignments that are accurate. This clustering metric computes the GM (Geometric Mean) between the precision and recall metric, and like most of the supervised learning measuring metrics, it is capped between the values of 0 – 1, with a larger value indicating that the algorithm did a good job at assigning individual samples to respective clusters. This is also available as a metric in sklearn and is therefore easy to implement.

2.) Silhouette Score

This algorithmtries to explain the extent to which two data points are similar to each other, given that they are assigned to the same cluster. The score is computed based on the aggregation of all the data points so we can have an average of the performance of the implementation in the entirety of the dataset. This metric measure lies between negative 1 and positive 1. Positive values are desired in this mechanism as a negative value indicates that the data points in the clusters are not very similar to each other.

Calinski-Harabasz Index

This implementation looks at the ratio of the variance of individual data points compared to the points in the clusters other than the one in which the current instance is attributed and the data points within the current cluster. For this metric, higher values are preferred.

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.

Anything is possible, if the data has categorical values, they can be encoded to convert it to numeric and train regression or classification models.

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

categorical predictive modeling?

Predictive modeling is a mathematical process used to predict future events or outcomes by analyzing patterns in a given set of input data. It is a crucial component of predictive analytics, a type of data analytics which uses current and historical data to forecast activity, behavior and trends.

Convert to Number

Convert to number: As discussed above, some ML libraries do not take categorical variables as input. Thus, we convert them into numerical variables. Below are the methods to convert a categorical (string) input to numerical nature:

Combine Levels

Combine levels: To avoid redundant levels in a categorical variable and to deal with rare levels, we can simply combine the different levels. There are various methods of combining levels. Here are commonly used ones:

Dummy Coding

Dummy Coding: Dummy coding is a commonly used method for converting a categorical input variable into continuous variable. ‘Dummy’, as the name suggests is a duplicate variable which represents one level of a categorical variable. Presence of a level is represent by 1 and absence is represented by 0. For every level present, one dummy variable will be created

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 models error rate, Kappa value, sensitivity, precision, and F-measure.

Precision = TP/TP+FP = 15/15+3 =0.833

Sensitivity/recall = TP/ TP+FN = 15/15+7=0.68

F- measure = 2\*Precision\*Recall/ Precision+Recall =1.12/1.51 =0.80

Error rate= total incorrect predictions/ total number of observations = 10/100 =0.1

10. Make quick notes on:

1. The process of holding out

What is the Hold-out method for training ML models? The hold-out method for training a machine learning model is the process of splitting the data into different splits and using one split for training the model and other splits for validating and testing the models.

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

A model parameter is a configuration variable that is internal to the model and whose value can be estimated from the given data. They are required by the model when making predictions. Their values define the skill of the model on your problem.

11. Define the following terms:

1. Purity vs. Silhouette width

The silhouette width is an established metric for evaluating cluster separation. For each cell, we compute the average distance to all cells in the same cluster. We also compute the average distance to all cells in another cluster, taking the minimum of the averages across all other clusters. The silhouette width for each cell is defined as the difference between these two values divided by their maximum. Cells with large positive silhouette widths are closer to other cells in the same cluster than to cells in different clusters. Thus, clusters with large positive silhouette widths are well-separated from other clusters.

The “clustering purity” is defined for each cell as the proportion of neighboring cells that are assigned to the same cluster, after some weighting to adjust for differences in the number of cells between clusteres. Well-separated clusters should exhibit little intermingling and thus high purity values for all member cells.

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

Eager learning methods construct general, explicit description of the target function based on the provided training examples. Lazy learning methods simply store the data and generalizing beyond these data is postponed until an explicit request is made