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?

The target function is the main objective function in machine learning. It is a mathematical function that shows the relationship between the inputs and outputs. For example, in the task of predicting the price of a house, the model try to find the function which relate variables such as the number of bedrooms, bathrooms, and square footage to the price of home and based on these relation predict the price of a home when these data prepare for it. To find the best function the model tries to minimize difference between predicted value and true value which is calculated by cost and loss functions.

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 predictive models, the target vale of training data exists and the model try to find the pattern and relation between target value and other variables. After training, if we give those variables model try to use same patterns to predict target value such as fraud detection. While in descriptive, the target value is not clear ad model try to group data based on their similarity and difference, this will append for new data too, based on the similarity and difference of new data to each of groups try to categorise new given data such as customer segmentation problem.

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

There are various metrics to assess classification model such as accuracy, precision, recall and F1 score. The most common one is accuracy which is calculated by summation of number of TP and TN divide to total sample of data which show how many cases predicted correctly.

The accuracy is not complete and it will be high even in case of overfitting, specially, when dataset is imbalance, thus other factors such as precision (TP/TP+FP) and recall (TP / TP+FN) and the F1-score which cover all other metrics and it is the best one for showing real performance which calculated through (precision\* recall)/ (precision + recall)

4. explain following items:

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

reason for underfitting?

A machine learning algorithm is underfit when it cannot capture the underlying trend of the data and it only performs well on training data but performs weekly on testing data. in this case accuracy is high on training and low on test data. It usually happens when we have fewer data to build an accurate model and also when we try to build a linear model with fewer non-linear data.

* High bias and low variance
* The size of the training dataset used is not enough.
* The model is too simple.
* Training data is not cleaned and also contains noise in it.

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

The model does not make accurate predictions on testing data, when we train a model with so much data, it starts learning from the noise and inaccurate data entries in our data set then the model does not categorize the data correctly, because of too many details and noise.

* High variance and low bias
* The model is too complex
* The size of the training data

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

Actions to decrease bias which makes a better fit to the training data and it will increase the variance in the model which lead to higher risk of poor predictions, it can happen in reverse when the actions are taken to reduce variance will inherently increase bias.

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

Yes, it is possible by providing more data samples, add context to the data by adding new data or creating new data from previous one, cross validation which help models to train with variety of data, hyperparameter tuning to select best match parameters for model based on the data and also try other algorithm the find best one.

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

In unsupervised learning the performance is evaluated based on some similarity or dissimilarity metrics such as the distance between cluster points. If the clustering algorithm separates dissimilar observations apart and similar observations together, then it has performed well. The methods of evaluating these algorithms are:

- Silhouette Coefficient

- Calisnki-Harabasz coefficient

- Dunn index

- Ball-Hall index

- Hartigan index

Also, precision, recall and other classification metrics can calculate but their formula will have different method for selecting TP, TN, FP, FN

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.

It is possible to do classification with some of regression algorithm if they are able to calculate probability of an output then we can put threshold and change that probability to classes. For example, logistic regression is a kind of regression which can calculate probability. But the thing is the performance may always not be optimum as compared to a classification algorithm.

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

categorical predictive modelling?

The predictive method for numerical values, try to predict cautious values such as price of home but classical methods try to predict distinct values such as fraud detection. In classical predictive, model should predict from specific values while is not same in numerical data, in numerical predictive we model should find the nearest match of previous data but in classification, the model defines some limit and criteria which help to classify data based on them.

9. The following data were collected when using a classification model to predict the malignancy of a group of patients’ tumours:

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.

|  |  |  |
| --- | --- | --- |
|  | 1 | 0 |
| 1 | 15 | 7 |
| 0 | 3 | 75 |

TP = 15, TN = 7, FP = 3, FN = 75

Acc= 15 +75 / 15+7+3+75= 0.9, Expected\_Acc= (82\*78 / 100) + (22 \*18 / 100) /100 = 0.679

Kappa = (Acc – Expected\_Acc) / (1-Eexpected\_Acc) = 0.68

Sensitivity = TP/ TP+FN = 0.166

Precision= TP / TP + FP = 0.83

F1-Score= 2 x [(Precision x Recall) / (Precision + Recall)] = 0.2682

10. Make quick notes on:

1. The process of holding out

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. The hold-out method is used for both model evaluation and model selection.

2. Cross-validation by tenfold

10-fold cross validation mean split data into 10 partition, then train the model 10 times and each time select one of the parts for validation and other parts for training.

3. Adjusting the parameters

The parameters that can get different values and during training we have to test various value for them and select the best value which give highest performance based on current data.

11. Define the following terms:

1. Purity vs. Silhouette width

silhouette width is a measure of how similar a data point is inside the cluster (cohesion) compared to the similarity of datapoints with other clusters (separation). In purity we assume a label for each cluster with the most frequent class label for those datapoint and calculate the number of correctly matched class and divide cluster labels by the number of total data points.

2. Boosting vs. Bagging

In bagging we use different algorithm parallelly to train the data and with the aim of mx voting or averaging calculate the final output of model while in boosting we use a series of a week algorithm and selecting data for training each step by bootstrap technique. in each step we train the wrong prediction data and unused data in previous training step. In this way we can improve performance of week algorithm.

3. The eager learner vs. the lazy learner

The eager learning is a learning technique where the system strives to generalize the training data before receiving queries, whereas lazy learning delays generalization beyond the training data until a query is made to the system