**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 is the function for solving a problem that an AI algorithm is trained to find. Once target function is made, it is used to predict results i.e. output data is calculated based on its input features. Consider we want to establish a relationship between experience and salary. A linear regression model can be made based on this:

Salary = Slope(Experience) + constant

The target’s function fitness is accessed by calculating the loss 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.**

Predictive models are those which uses the past data or evaluating what might can happen in future. Steps that must be followed

1) Define business objective.

2) Explore & profile your data

3) Gather, Clean and integrate your data

4) Built the predictive model

5) Incorporate analytics into business process

6) Monitor and measure results.

Descriptive type of models are those models which uses the stored data to know the past analytics that might helps with the future decisions of the company. Steps followed by this type of model are

1) Define objective

2) Collect and explore data

3) Clean the data

4) Data analytics

5) Interpret results

6) Communicate results

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

The various methods adopted in order to access the efficiency of a classification model are

**Confusion Matrix** – Confusion Matrix is a kind of 2 x 2 matrix in case of binary classification.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual |  |
|  |  | 1 | 0 |
| Pred | 1 | TP | FP |
|  | 0 | FN | TN |

The focus is always on reducing Type 1 error & Type 2 error. Since, it is a balanced dataset, we use accuracy as metrics.

**Accuracy = (TP+TN) / (TP+FP+FN+TP)**

If the dataset is unbalanced, we go with

**Recall or sensitivity** – Out of the total actual positive values, how many we have predicted positively. This is mainly used in healthcare sector eg. Cancer Detection.

**Recall = TP / (TP + FN)**

**Precision –** Out of the total actual predicted values how many was actually positive. This application is found in spam detection

**Precision = TP / (TP + FP)**

**F Beta Score -** When FP and FN both are important. It is given by

**((1+beta²) (Precision \* Recall))/( Precision + Recall)**

* If FP & FN both are important, beta = 1
* If FP is more important than FN, beta = 0.5 , then F0.5 score
* If FN is more important than FP, beta = 2 , then F2 score

**ROC & AUC curve** – It establishes the relationship between TPR and FPR at various threshold levels. By this method they try to visualise the accuracy of classification model.

4.

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

Underfitting is the situation in which our model does not get trained well in training dataset. Also it does not perform well in our test dataset. This basically leads to a model with high bias and high variance. The most common reasons for underfitting are:

1) The model being too simple.

2) Having data with feature.

3) The duration for which data is trained being too small.

4) Having noise in the data.

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

Overfitting is the situation in which our model get trained very well in training dataset. But it performs poorly in our test dataset. This basically leads to a model with low bias and high variance. The most common reasons for overfitting are:

1) The model being too complex.

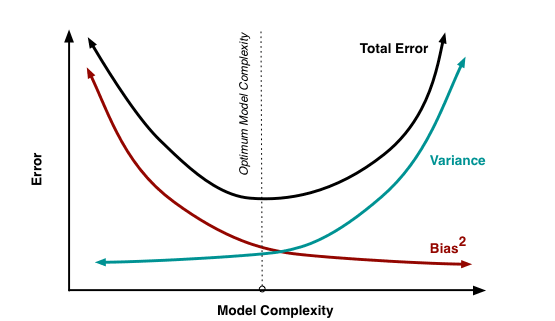
2) Small training data and trying to learn from it.

3) Data with noise i.e. like those with outliers & errors in the data.

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

Bias is basically the error that we get as expected model prediction differ from the actual one. Model with high bias pays very little attention to the model and oversimplifies it.

Variance is the variability in the predicted model that we get when we introduce the model with different set of data. Model with high variance pays very huge attention to training dataset but does not generalise it for dataset not seen before.

****

Thus it is a problem to have a model with high bias and high variance. As we increase bias the variance will reduce and vice versa. Hence it is important to maintain an optimum value for both which is known as “Bias Variance Tradeoff”.

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

Yes, it is possible to boost the efficiency of a learning model. Boosting algorithm makes this possible. In these kind of models several weak learners/base estimators are combined together to form a strong models.

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

Clustering is the most common form of unsupervised model. Our goal in this is to create clusters that have similar observations which can be further clubbed together and dissimilar observations kept as far as possible. Clusters are evaluated based on some similarity and dissimilarity measures suchj as distance between clusters points.

Sillhoute coefficients : s=(b-a)/max(a,b)

a = mean distance between a sample and all other points in the same clusters.

b = mean distance between a sample and all other points in the neighbouring clusters.

-1 for incorrect clustering, 0 for overlapping clusters and +1 for highly dense clusters.

**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.**

Yes, it is possible to use a classification model for numerical data or a regression model for categorical data with a classification model. In case, For numerical data if we can define the various categories for the output data for example for binary classification, data above a certain level is defined a value and rest another. In such a case classification model can be used.

Similarly, for categorical data if we are able to define a continuous value as an output, then in such a case regression model can be used.

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

Predictive modelling method for numerical data involves forecasting the output with a given set of inputs. Example include various regression models.

Categorical modelling involves predicting the predefined class/categories to which our output belongs to. Example for these include various classification methods.

**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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Actual** |  |  |
|  |  | **Yes** | **No** |  |
|  | **Yes** | 15 | 3 | 18 |
| **Pred** | **No** | 7 | 75 | 82 |
|  |  | 22 | 78 | 100 |
|  |  |  |  |  |

**Error Rate** = Classified Incorrectly / Total = (10/100)\*100 = 10%

**Kappa Value** = P(o) = Observed Agreement = 90/100 = 0.9

P(e) = Agreement By Chance = (22/100) \* (18/100)+ (78/100) \* (82/100) = 6792/10000 =0.68

Kappa Value = (0.9-0.68)/(1-0.68) = 0.22/0.32 = 0.69

Thus there is substancial agreement with the results.

**Precision** = TP/(TP+FP) = 15/18 = 0.833

**Recall (sensitivity)** = TP/(TP+FN) = 15/(15+7) = 15/22 = 0.681

**F1 Score** = (2 \* 0.5672)/1.514 = 0.749

**10. Make quick notes on:**

**1. The process of holding out**

When you split your dataset into train and test dataset, the process is known as “Process of Holding Out”. The training set is what the model is trained on and test set is used to see how well the model has performed.

**2. Cross-validation by tenfold**

In this method, we have one dataset which we divide randomly into 10 parts. We use 9 of those for training and reserve one tenth for testing. We repeat this procedure 10 times so that different dataset is selected for testing each time.

**3. Adjusting the parameters**

Hyperparameter tuning is the process adopted for adjusting the parameters. In this, accuracy is checked by manipulating the parameters values. A higher accuracy model is selected which can perform well with our test data.

**11. Define the following terms:**

**1. Purity vs. Silhouette width**

**2. Boosting vs. Bagging**

Boosting is the process in which several weak learners/base estimators are combined in sequential mannar to form a strong model. Only erroneous records are passed by updating their weights.

Bagging is the process in which on dataset is passed through many models connected in parallel. This process is also known as “Bootstrap Aggregation”.

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

Eager Learner – When these kinds of algorithms are provided with training data, they constructs a classification model before receiving new data to classify.

Lazy Learner – These algorithm simply stores the training data and waits until provided with the test data.