Question 1

Difference between supervised and unsupervised learning?

|  |  |
| --- | --- |
| SUPERVISED | UNSUPERVISED |
| In this type of learning, the target variable is clearly defined. If not defined, we are expected to create one. We build an algorithm to learn the mapping function from input to output i.e y = f(x) | In this type of learning, the target variable is not required. |
| It is used to map inputs to outputs. The outputs can be continuous or discrete. The outputs are the predictions. | Here inputs are not mapped to outputs. Rather, here patterns are discovered. Here we try to understand distribution of data. |
| The main types of supervised learnings are Classification, Regression. | The main types of unsupervised learnings are Clustering, Dimension Reduction and Association Rules. |

Question 2

What is the difference between pruned decision tree and unpruned tree?

Answer: The main difference is that in a pruned tree, from the original tree, the leaf nodes are removed, and some branches are converted to leaf nodes. The branches that are removed are done so because they do not contribute that well to the result. This essentially reduces the size of the tree. Trees that are very deep, tend to be overfitting i.e do well on training data but perform badly on testing data. In decision trees, this problem is avoided by pruning.

Question 3

What is overfitting? How to avoid that?

The algorithm used to learn the mapping function from the inputs to the outputs may perform better on training data i.e the predictions made with training data are almost close to the target variable but perform very badly on the testing set i.e low accuracy. This is low bias, high variance. Any new point that comes, the model will not be able to get a proper prediction.

Regularization is one of the techniques to avoid overfitting. It reduces the coefficient estimates. Also applying a k-fold cross validation technique also allows the model to be trained much better.

Question 4

What is the output of the Logistic regression (Binary)?

Answer: For a Binary Logistic model, there are two outputs/ target classes available. Ex 0 and 1. In order to determine a given data point has either 0 or 1, a probability for the point is generated. We fix a threshold value by observing the ROC curve where we see high True Positive Rate and low/minimum False Positive Rate for the given data set. The threshold value determines which target class the data point falls under. This is explained with an example below.

Example: Classifying whether customer will default on loan.

A screenshot of a map

Description automatically generated

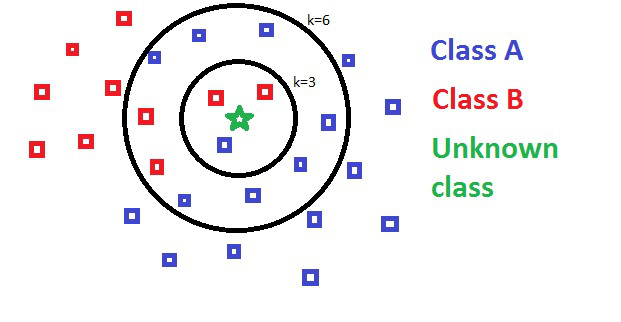
The above is a plot between the True Positive Rate and False Positive Rate. This is generated for all the training data provided in this data set. Here we can observe that, 0.2 is a good threshold value which satisfies the condition of high True Positive Rate and low False Positive Rate. This threshold value is used to predict the output a given data point as either belonging to class A or class B. In this example, any probability less than 0.2, means the customer is unlikely to default. Any value greater than 0.2 and less than 1 means, customer likely to default.

Question 5

What is k-NN, k-means, k-fold cross validation?

Answers: k-NN (k nearest neighbors) is a supervised learning approach. Here, the target variable/ output of a given data point depends on its K nearest neighbors. The way it does this is by measuring the similarity between the given point and its neighbors. The more similar it’s with respect to its neighbors, more likely it will share the same value as output. K-NN can be used in case of regression as well as classification.

Example: Determining if a given point belongs to ‘Red’ color or ‘Blue’ color



In the above diagram, the aim is to determine the class of a given point. Here if we keep the value of k = 3, which means we look at three nearest neighbors. We have 2 points that belong to class Red and 1 point which belongs to class Blue. Therefore, the point is classified as ‘Red’.

k-means: This is a unsupervised learning technique with the objective of classifying the given data points into k clusters. Here, depending on the value of k, centroids are randomly chosen. If there are 3 clusters, then 3 centroid points are randomly chosen. Next, points fall into a cluster depending on the distance between the centroid and the point. The clusters are formed in such a way that there is clear distinction i.e no overlapping clusters. The initial clustering will have some data points classified incorrectly. Keep repeating this process, until the all the points are correctly grouped i.e no changes to the centroids. For a given cluster, we calculate the centroid value by taking the mean of the data points in that cluster.

k-fold cross validation: It’s a resampling procedure in which the given data points are split into multiple mini train-test splits. This is done so that the model learns more from the data. It’s a good approach to reduce overfitting.

Example : For a given data set, if we apply k-fold cross validation and choose the value of k = 5 i.e 5 training-test samples. So at any given point, 4 groups will be chosen for training and the remaining will be used as a testing set. This is done till we have iterated over all the groups.

A picture containing different

Description automatically generated