Top of Form

Bottom of Form

Top of Form

Bottom of Form

|  |  |
| --- | --- |
|  | 1. The value of correlation coefficient will always be:  **C) between -1 and 1** |
|  |  |

2. Which of the following cannot be used for dimensionality reduction? :

**D) Ridge Regularization**

3. Which of the following is not a kernel in Support Vector Machines? :

**C) Hyper plane**

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries? :

**D) Support Vector Classifier**

5. In a Linear Regression problem, ‘X’ is independent variable and ‘Y’ is dependent variable, where ‘X’ represents weight in pounds. If you convert the unit of ‘X’ to kilograms, then new coefficient of ‘X’ will be :

|  |
| --- |
|  |
| **A) 2.205 × old coefficient of ‘X’** |
|  |
|  |

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?

**B) increases**

7. Which of the following is not an advantage of using random forest instead of decision trees?

**C) Random Forests are easy to interpret**

8. Which of the following are correct about Principal Components?

**B) Principal Components are calculated using unsupervised learning techniques**

**C) Principal Components are linear combinations of Linear Variables**

9. Which of the following are applications of clustering?

**A,B,C,D**

10.Which of the following is(are) hyper parameters of a decision tree?

**A,B,D**

**11.What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.**

**Outliers:**

An outlier is a data point in a data set that is distant from all other observations. A data point that lies outside the overall distribution of the data set.

**Eg:** 20,34,12,50,800

"800" is an outlier because it lies outside the overall distribution. Also, these outliers may impact the data set and their model accuracy, in linear regression if there are any outliers then our best fit line get diverted and due to which there is problem of high bias and high variance. Thereis multiple model which get impacted by the outliers like:

Linear Regression

Logistic Regression

K-Means

**Detect Outliers:**

We can see the outliers present in the data set by using scatter plot and box plot. Also, by checking the range or difference between the 75% percentile and max value.

**Techniques to remove outliers:**

\* calculating z score

\* IQR (inter Quantile Range)

Box plot use the IQR method to display data and outliers (shape of the data) but in order to be get a list of identified outliers, we will need to use the mathematical formula and retrieve the outlier data.

The IQR, fall under spread, which is a statically dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q3 – Q1.

|  |  |
| --- | --- |
|  |  |
|  |  |

**12. What is the primary difference between bagging and boosting algorithms?**

Both bagging and boosting are fall under ensemble techniques, which are used to reduce overfitting and improve the score of the model. There is various algorithm fall under this technique like: Random Forest, XG Boost, AdaBoost, Gradient Boosting, All These techniques used Decision tree internally.

There are 2 types of ensemble techniques which are: Bagging and boosting.

**Bagging :** Main aim of Bagging is to decrease variance not bias

In this technique the dataset is divided into multiple models and each model provide or calculate the output and at last all the model output is aggregate and generate the final output

The techniques by which the dataset is divided into various models are called row sampling + Feature Selection , in which the number of records or number of rows are divided by sampling and provided to different models and then each model generate the output and at last we calculate the final output.

Bagging (Bootstrap Aggregation) is used when our goal is to reduce the variance of a decision tree. Here idea is to create several subsets of data from training sample chosen randomly with replacement. Now, each collection of subset data is used to train their decision trees. As a result, we end up with an ensemble of different models. Average of all the predictions from different trees are used which is more robust than a single decision tree.

Some of Advantage’s of Random Forest:

* Handles High Dimension data easily.
* Handles missing values and maintain accuracy of the model

**Boosting:** Main aim of Boosting is to decrease bias not variance

Here are some steps to follow :

*1.*  The base learner takes all the distributions and assign equal weight or attention to each observation.

*2.* If there is any prediction error caused by first base learning algorithm, then we pay higher attention to observations having prediction error. Then, we apply the next base learning algorithm.

*3.* Iterate Step 2 till the limit of base learning algorithm is reached or higher accuracy is achieved.

Finally, it combines the outputs from weak learner and creates a strong learner which eventually improves the prediction power of the model. Boosting pays higher focus on examples which are mis-classiﬁed or have higher errors by preceding weak rules.

**13.What is adjusted R2 in logistic regression. How is it calculated?**

**Adjusted R Square:**

Before Adjusted R Square, lets understand R square, this is used for regression type of problem to calculate the accuracy of model.

Formula:

More near value of R square value to 1 means the line is best.

So, this feature also tries to reduce the overfitting problem and can also determines to reduce the curse of dimension

**14.What is the difference between standardization and normalization ?**

Both techniques are fall under feature scaling.Having features on a similar scale can help the gradient descent converge more quickly towards the minima or the minimum global loss.

**Normalization:**

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here’s the formula for normalization:

X’=(X-Xmin)/(Xmax-Xmin)

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

* When the value of X is the minimum value in the column, the numerator will be 0, and hence X’ is 0
* On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X’ is 1
* If the value of X is between the minimum and the maximum value, then the value of X’ is between 0 and 1

## What is Standardization?

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here’s the formula for standardization:

X’=(X-Mean)/Feature scaling: Sigma

Feature scaling: Sigma  is the standard deviation of the feature values

**15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.**

**Cross – Validation:**

It is primarily used in Machine Learning to estimate the skill of a ML model on unseen data, it used a limited sample in order to estimate that how the model is expected to perform in general when used to make prediction

|  |
| --- |
| While doing train – test split, we use random state, so at every random state the score values is different, so the accuracy values is not consistent or generalized to make it generalized for every dataset or upcoming values we use cross validation.  Advantage : It prevents our model from over-fitting the training data set  Disadvantage: Cross validation increases the training time |

|  |
| --- |
|  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |