

In Q1 to Q8, only one option is correct, Choose the correct option:

- Answer: A)**

- Answer:Random Forest**

- Answer:B)**

- Answer:C)**

- Answer:A)**

- Answer:C)**

- Answer:C)**

- Answer:C)**

Q9 to Q15 are subjective answer type questions, Answer them briefly.

9. Suppose we have a dataset which have two classes A and B. The percentage of class A is 40% and percentage of class B is 60%. Calculate the Gini index and entropy of the dataset.

Answer: Both gini and entropy are measures of impurity of a node. A node having multiple classes is impure whereas a node having only one class is pure. Entropy in statistics is analogous to entropy in thermodynamics where it signifies disorder. If there are multiple classes in a node, there is disorder in that node.

10. What are the advantages of Random Forests over Decision Tree?

Answer: Random forests consist of multiple single trees each based on a random sample of the training data. They are typically more accurate than single decision trees. **Trees are unpruned.** While a single decision tree like CART is often pruned, a random forest tree is fully grown and unpruned, and so, naturally, the feature space is split into more and smaller regions. **Trees are diverse.** Each random forest tree is learned on a random sample, and at each node, a random set of features are considered for splitting. Both mechanisms create diversity among the trees.

Unpruned and diverse trees lead to a high resolution in the feature space. For continuous features, it means a smoother decision boundary

11. What is the need of scaling all numerical features in a dataset? Name any two techniques used for scaling.

Real world dataset contains features that highly vary in magnitudes, units, and range. Normalisation should be performed when the scale of a feature is irrelevant or misleading and not should Normalise when the scale is meaningful. The algorithms which use Euclidean Distance measure are sensitive to Magnitudes. Here feature scaling helps to weigh all the features equally. Formally, If a feature in the dataset is big in scale compared to others then in algorithms where Euclidean distance is measured this big scaled feature becomes dominating and needs to be normalized.

Featurization and Normalization are 2 scaling techniques

- 12.
-

13. Write down some advantages which scaling provides in optimization using gradient descent algorithm.

Optimization is a big part of machine learning.

Gradient descent is a simple optimization procedure that you can use with many machine learning algorithms.

Batch gradient descent refers to calculating the derivative from all training data before calculating an update.

Stochastic gradient descent refers to calculating the derivative from each training data instance and calculating the update immediately.

14. In case of a highly imbalanced dataset for a classification problem, is accuracy a good metric to measure the performance of the model. If not, why?

classification accuracy is almost universally inappropriate for imbalanced classification. The reason is, a high accuracy (or low error) is achievable by a no skill model that only predicts the majority class.

As such, it is natural to use it on imbalanced classification problems, where the distribution of examples in the training dataset across the classes is not equal. When the class distribution is slightly skewed, accuracy can still be a useful metric. When the skew in the class distributions are severe, accuracy can become an unreliable measure of model performance.

15. What is “f-score” metric? Write its mathematical formula.

The F-score, also called the F1-score, is a measure of a model’s accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into ‘positive’ or ‘negative’.

The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model’s precision and recall.

The F-score is commonly used for evaluating information retrieval systems such as search engines, and also for many kinds of machine learning models, in particular in natural language processing.

16. What is the difference between `fit()`, `transform()` and `fit_transform()`?

"fit" computes the mean and std to be used for later scaling. (just a computation), nothing is given to you.

"transform" uses a previously computed mean and std to autoscale the data (subtract mean from all values and then divide it by std). "fit_transform" does both at the same time.
