Machine learning 15

1. Recognize the differences between supervised, semi-supervised, and unsupervised learning.

Ans:- Within the field of machine learning, there are three main types of tasks: supervised, semi-supervised, and unsupervised. The main difference between these types is **the level of availability of ground truth data**, which is prior knowledge of what the output of the model should be for a given input

2. Describe in detail any five examples of classification problems.

Ans:- Examples of classification problems include: Given an example, **classify if it is spam or not**. Given a handwritten character, classify it as one of the known characters. Given recent user behavior, classify as churn or not.

3. Describe each phase of the classification process in detail.

Ans:- Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known

4. Go through the SVM model in depth using various scenarios.

Ans:- Deep learning and SVM are different techniques. But thinking SVM as deep learning has misconceptions too. They can not be same but **can be used together**. Deep learning is more powerfull classifier than SVM.

5. What are some of the benefits and drawbacks of SVM?

Ans:- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

6. Go over the kNN model in depth.

Ans:- The value of k in the KNN algorithm is related to the error rate of the model. **A small value of k** could lead to overfitting as well as a big value of k can lead to underfitting. Overfitting imply that the model is well on the training data but has poor performance when new data is coming.

7. Discuss the kNN algorithm's error rate and validation error.

Ans:- By observing validation error rate we can interpret that At K=1, we were over fitting the boundaries. **In Validation graph Error rate initially decreases and reaches a minima.** **After the minima point, it then increase with increasing K**. This value of K where error reaches minima should be used for all predictions.

8. For kNN, talk about how to measure the difference between the test and training results.

Ans:- Specifically, four different distance functions, which are **Euclidean distance, cosine similarity measure, Minkowsky, correlation, and Chi square**, are used in the k-NN classifier respectively

9. Create the kNN algorithm.

Ans:- With the help of KNN algorithms, we can **classify a potential voter into various classes like “Will Vote”, “Will not Vote”, “Will Vote to Party 'Congress', “Will Vote to Party 'BJP'**. Other areas in which KNN algorithm can be used are Speech Recognition, Handwriting Detection, Image Recognition and Video Recognition.

11. Describe the different ways to scan a decision tree.

Ans:- A decision tree model is formed using a hierarchy of branches. Each path from the root node through internal nodes to a leaf node represents a classification decision rule. These decision tree pathways can also be represented as '**if-then' rules**.

12. Describe in depth the decision tree algorithm.

Ans:- The decision tree **splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes**. The ID3 algorithm builds decision trees using a top-down greedy search approach through the space of possible branches with no backtracking.

13. In a decision tree, what is inductive bias? What would you do to stop overfitting?

Ans:- Before learning a model given a data and a learning algorithm, there are a few assumptions a learner makes about the algorithm. These assumptions are called the inductive bias. It is like the property of the algorithm. For eg. in the case of decision trees, **the depth of the tress** is the inductive bias.

14.Explain advantages and disadvantages of using a decision tree?

Ans:- They are very fast and efficient compared to KNN and other classification algorithms. Easy to understand, interpret, visualize. The data type of decision tree can handle any type of data whether it is numerical or categorical, or boolean. Normalization is not required in the Decision Tree.

15. Describe in depth the problems that are suitable for decision tree learning.

Ans:- Decision tree learning is generally best suited to problems with the following characteristics: **Instances are represented by attribute-value pairs**. There is a finite list of attributes (e.g. hair colour) and each instance stores a value for that attribute (e.g. blonde).

16. Describe in depth the random forest model. What distinguishes a random forest?

Ans:- **A decision tree combines some decisions, whereas a random forest combines several decision trees**. Thus, it is a long process, yet slow. Whereas, a decision tree is fast and operates easily on large data sets, especially the linear one. The random forest model needs rigorous training.

17. In a random forest, talk about OOB error and variable value.

Ans:- The out-of-bag (OOB) error is **the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample**. This allows the RandomForestClassifier to be fit and validated whilst being trained [1].