1. What is the concept of supervised learning? What is the significance of the name?

Supervised learning, also known as supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately.

2. In the hospital sector, offer an example of supervised learning.

Cancer Detection,

3. Give three supervised learning examples.

House price prediction, cancer detection, Mushroom classifier

4. In supervised learning, what are classification and regression?

Classification tries to find the decision boundary, which divides the dataset into different classes. Regression algorithms solve regression problems such as house price predictions and weather predictions

5. Give some popular classification algorithms as examples.

Logistic Regression, Decision tree, SVM, KNN, XGBOOST, Random forest

6. Briefly describe the SVM model.

support vector machine (SVM) is a type of deep learning algorithm that performs supervised learning for classification or regression of data groups. In AI and machine learning, supervised learning systems provide both input and desired output data, which are labeled for classification

7. In SVM, what is the cost of misclassification?

Misclassification costs allow you to specify the relative importance of different kinds of prediction errors. Misclassification costs are basically weights applied to specific outcomes. These weights are factored into the model and may actually change the prediction (as a way of protecting against costly mistakes)

8. In the SVM model, define Support Vectors.

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM

9. In the SVM model, define the kernel.

Kernels are used in Support Vector Machines (SVMs) to solve regression and classification problems. Support Vector Machines use the Kernel Trick to transform linearly inseparable data into linearly separable data, thus finding an optimal boundary for possible outputs

10. What are the factors that influence SVM&#39;s effectiveness?

The effectiveness of SVM depends on the selection of kernel, kernel's parameters and soft margin parameter C. . Each pair of parameters is checked using cross validation, and the parameters with best cross validation accuracy are picked.

11. What are the benefits of using the SVM model?

The advantages of SVM and support vector regression include that they can be used to avoid the difficulties of using linear functions in the high-dimensional feature space, and the optimization problem is transformed into dual convex quadratic programs.

12. What are the drawbacks of using the SVM model?

SVM algorithm is not suitable for large data sets. SVM does not perform very well when the data set has more noise i.e. target classes are overlapping. In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform

13. Notes should be written on

1. The kNN algorithm has a validation flaw.

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

2. In the kNN algorithm, the k value is chosen.

The k value in the k-NN algorithm defines how many neighbors will be checked to determine the classification of a specific query point. For example, if k=1, the instance will be assigned to the same class as its single nearest neighbor.

3. A decision tree with inductive bias

Shorter trees are preferred over longer trees. Trees that place high information gain attributes close to the root are preferred over those that do not.

14. What are some of the benefits of the kNN algorithm?

* Quick calculation time.
* Simple algorithm – to interpret.
* Versatile – useful for regression and classification.
* High accuracy – you do not need to compare with better-supervised learning models.

15. What are some of the kNN algorithm’s drawbacks?

* Accuracy depends on the quality of the data.
* With large data, the prediction stage might be slow.
* Sensitive to the scale of the data and irrelevant features.
* Require high memory – need to store all of the training data.
* Given that it stores all of the training, it can be computationally expensive.

16. Explain the decision tree algorithm in a few words.

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

17. What is the difference between a node and a leaf in a decision tree?

Root node: The base of the decision tree. Splitting: The process of dividing a node into multiple sub-nodes. Decision node: When a sub-node is further split into additional sub-nodes. Leaf node: When a sub-node does not further split into additional sub-nodes; represents possible outcomes.

18. What is a decision tree’s entropy?

In the context of Decision Trees, entropy is a measure of disorder or impurity in a node. Thus, a node with more variable composition, such as 2Pass and 2 Fail would be considered to have higher Entropy than a node which has only pass or only fail

19. In a decision tree, define knowledge gain.

The information gained in the decision tree can be defined as the amount of information improved in the nodes before splitting them for making further decisions. To understand the information gain let's take an example of three nodes

20. Choose three advantages of the decision tree approach and write them down.

* Simple to understand and to interpret
* Requires little data preparation.
* The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.
* Able to handle both numerical and categorical data.
* Able to handle multi-output problems.

21. Make a list of three flaws in the decision tree process.

* Not good for Regression.
* Overfitting Problem
* Expensive

22. Briefly describe the random forest model.

Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.