Assignment 14 Solutions ¶

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

Ans: 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 ?

Ans: Supervised learning uses a training set to teach models to yield the desired output. This training dataset includes inputs and correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized.

3. Give three supervised learning examples ?

Ans: Example of Supervised Learning Algorithms are:

- Linear Regression.
- Nearest Neighbor.
- Gaussian Naive Bayes.
- Decision Trees.
- Support Vector Machine (SVM)
- Random Forest.

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

Ans: Fundamentally, classification is about predicting a label and regression is about predicting a quantity. That classification is the problem of predicting a discrete class label output for an example. That regression is the problem of predicting a continuous quantity output for an example.

5. Give some popular classification algorithms as examples ?

Ans: Popular algorithms that can be used for multi-class classification include:

- k-Nearest Neighbors
- Decision Trees.
- Naive Bayes

6. Briefly describe the SVM model ?

Ans: A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text.

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

Ans: In cost-sensitive learning instead of each instance being either correctly or incorrectly classified, each class (or instance) is given a misclassification cost. Thus, instead of trying to optimize the accuracy, the problem is then to minimize the total misclassification cost.

8. In the SVM model, define Support Vectors ?

Ans: 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 ?

Ans: SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form. These functions can be different types. For example linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid.

10. What are the factors that influence SVM's effectiveness ?

Ans: SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.

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

Ans: SVM works relatively well when there is a clear margin of separation between classes. SVM is more effective in high dimensional spaces. SVM is effective in cases where the number of dimensions is greater than the number of samples.SVM is relatively memory efficient.

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

Ans: 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.
2. In the kNN algorithm, the k value is chosen.
3. A decision tree with inductive bias

Ans: The Short notes on below topics is:

- **The kNN algorithm has a validation flaw.** The relatively low accuracy of kNN is caused by several factors. One of them is that every characteristic of the method has the same result on calculating distance. The solution of this problem is to give weight to each data characteristic.
- **In the kNN algorithm, the k value is chosen.** The optimal K value usually found is the square root of N, where N is the total number of samples. Use an error plot or accuracy plot to find the most favorable K value. KNN performs well with multi-label classes, but you must be aware of the outliers.
- **A decision tree with inductive bias** Shorter trees are preferred over longer ones. 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 ?

Ans: Some Advantages of KNN are:

- 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 ?

Ans: Some Disadvantages of KNN are:

- 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 ?

Ans: A decision tree is a graphical representation of all the possible solutions to a decision based on certain conditions. Tree models where the target variable can take a finite set of values are called classification trees and target variable can take continuous values (numbers) are called regression trees.

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

Ans: A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

18. What is a decision tree's entropy ?

Ans: Entropy helps us to build an appropriate decision tree for selecting the best splitter. Entropy can be defined as a measure of the purity of the sub split. Entropy always lies between 0 to 1. The entropy of any split can be calculate by this formula.

19. In a decision tree, define knowledge gain ?

Ans: Information gain is the reduction in entropy or surprise by transforming a dataset and is often used in training decision trees.Information gain is calculated by comparing the entropy of the dataset before and after a transformation.

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

Ans: Advantages of Decision Trees :

- Easy to read and interpret. One of the advantages of decision trees is that their outputs are easy to read and interpret without requiring statistical knowledge.
- Easy to prepare.
- Less data cleaning required.

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

Ans: Issues in Decision Tree Learning :

- Overfitting the data.
- Guarding against bad attribute choices.
- Handling continuous valued attributes.
- Handling missing attribute values.
- Handling attributes with differing costs.

22. Briefly describe the random forest model ?

Ans: The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.