**Assignment\_14**

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

**Ans: Supervised 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.

3. Give three supervised learning examples.

**Ans:** **Spam detection , Customer sentiment analysis , Image- and object-recognition**

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

**Ans: Classification uses an algorithm to accurately assign test data into specific categories. It recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be labeled or defined. Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest.**

**Regression is used to understand the relationship between dependent and independent variables. It is commonly used to make projections, such as for sales revenue for a given business**

5. Give some popular classification algorithms as examples.

**Ans:**

**Logistic Regression.**

**Naive Bayes.**

**K-Nearest Neighbors.**

**Decision Tree.**

**Support Vector Machines**.

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.**

**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.**

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: A kernel is a function used in SVM for helping to solve problems. They provide shortcuts to avoid complex calculations. The amazing thing about kernel is that we can go to higher dimensions and perform smooth calculations with the help of it. We can go up to an infinite number of dimensions using kernels.**

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

**Ans: Effective in high dimensional cases. Its memory efficient as it uses a subset of training points in the decision function called support vectors. Different kernel functions can be specified for the decision functions and its possible to specify custom kernels.**

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.

**Ans :In other words, for kNN, there is no training step because there is no model to build. Template matching & interpolation is all that is going on in kNN. Neither is there a validation step. Validation measures model accuracy against the training data as a function of iteration count (training progress**).

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

**Ans: In KNN, K is the number of nearest neighbors. The number of neighbors is the core deciding factor. K is generally an odd number if the number of classes is 2. When K=1, then the algorithm is known as the nearest neighbor algorithm**.

3. A decision tree with inductive bias

**Ans: In the case of decision trees, the depth of the tress is the inductive bias. If the depth of the tree is too low, then there is too much generalisation in the model.**

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

**Ans:**

* **Simple to implement and intuitive to understand.**
* **Can learn non-linear decision boundaries when used for classfication and regression.**
* **No Training Time for classification/regression :**

**The KNN algorithm has no explicit training step and all the work happens during prediction.**

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

**Ans:**

**1.It is advised to use the KNN algorithm for multiclass classification if the number of samples of the data is less than 50,000. Another limitation is the feature importance is not possible for the KNN algorithm.**

**2  Does not work well with large dataset: In large datasets, the cost of calculating the distance between the new point and each existing points is huge which degrades the performance of the algorithm.  
  
3. Does not work well with high dimensions: The KNN algorithm doesn't work well with high dimensional data because with large number of dimensions, it becomes difficult for the algorithm to calculate the distance in each dimension.  
  
4. Need feature scaling: We need to do feature scaling (standardization and normalization) before applying KNN algorithm to any dataset. If we don't do so, KNN may generate wrong predictions.  
  
5 Sensitive to noisy data, missing values and outliers: KNN is sensitive to noise in the dataset. We need to manually impute missing values and remove outliers.**

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: The root node is just the topmost decision node. In other words, it is where you start traversing the classification tree. The leaf nodes (green), also called terminal nodes, are nodes that don't split into more nodes. Leaf nodes are where classes are assigned by majority vote.**

18. What is a decision tree's entropy?

**Ans: Entropy is an information theory metric that measures the impurity or uncertainty in a group of observations. It determines how a decision tree chooses to split data.**

19. In a decision tree, define knowledge gain.

**Ans: 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.**

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

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

**Ans: They are unstable, meaning that a small change in the data can lead to a large change in the structure of the optimal decision tree. They are often relatively inaccurate. Many other predictors perform better with similar data.**

22. Briefly describe the random forest model.

**Ans: Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.**

**Random Forest is used for both classification and regression—for example, classifying whether an email is “spam” or “not spam”**