**Assignment\_15**

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

**Ans: supervised learning : uses labels to define the output.**

**Semi – supervised : semi- supervised models uses both labelled and unlabeled data for training.**

**Unsupervised : It uses unlabeled data learns itself for the output data.**

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

* **DNA Expression Microarray.**
* **Email Spam**
* **Handwritten Digit Recognition**
* **Image segmentation**
* **Speech Recognition**

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

Ans: **(1) data acquisition and segmentation**

**(2) data preprocessing**

**(3) feature extraction/dimension reduction**

**(4) recognition and classification**.

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

**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.**

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

#### Ans: SVM Advantages

* **SVM’s are very good when we have no idea on the data.**
* **Works well with even unstructured and semi structured data like text, Images and trees.**
* **The kernel trick is real strength of SVM. With an appropriate kernel function, we can solve any complex problem.**
* **Unlike in neural networks, SVM is not solved for local optima.**
* **It scales relatively well to high dimensional data.**
* **SVM models have generalization in practice, the risk of over-fitting is less in SVM.**
* **SVM is always compared with ANN. When compared to ANN models, SVMs give better results.**

**SVM Disadvantages**

* **Choosing a “good” kernel function is not easy.**
* **Long training time for large datasets.**
* **Difficult to understand and interpret the final model, variable weights and individual impact.**
* **Since the final model is not so easy to see, we can not do small calibrations to the model hence its tough to incorporate our business logic.**
* **The SVM hyper parameters are Cost -C and gamma. It is not that easy to fine-tune these hyper-parameters. It is hard to visualize their impact**

6. Go over the kNN model in depth.

**Ans: The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.**

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

Ans:

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

**Ans: 1c. KNN (K=1)**

1. **KNN model. Pick a value for K. ...**
2. **This would always have 100% accuracy, because we are testing on the exact same data, it would always make correct predictions.**
3. **KNN would search for one nearest observation and find that exact same observation. KNN has memorized the training set.**

9. Create the kNN algorithm.

**Ans:**

**The abbreviation KNN stands for “K-Nearest Neighbour”. It is a supervised machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements. The number of nearest neighbours to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'.**

**Building out the KNN Framework**

**Use the distance function to get the distance between a test point and all known data points. Sort distance measurements to find the points closest to the test point (i.e., find the nearest neighbors) Use majority class labels of those closest points to predict the label of the test point**

10. What is a decision tree, exactly? What are the various kinds of nodes? Explain all in depth.

**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).**

12. Describe in depth the decision tree algorithm.

**Ans: Tree depth is a measure of how many splits a tree can make before coming to a prediction. This process could be continued further with more splitting until the tree is as pure as possible.**

**Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that the purity of the node increases with respect to the target variable**.

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

**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.**

**Handling overfitting**

**1.Reduce the network's capacity by removing layers or reducing the number of elements in the hidden layers.**

**2. Apply regularization , which comes down to adding a cost to the loss function for large weights.**

**3 .Use Dropout layers, which will randomly remove certain features by setting them to zero.**

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

**Ans: Decision Tree is used to solve both classification and regression problems. But the main drawback of Decision Tree is that it generally leads to overfitting of the data.**

**Some advantages of decision trees are:**

* **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.**

**Disadvantages of decision trees:**

**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.**

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: Random forest model**

**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.**

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.**