Assignment 15 Solutions

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

Ans: The differences between supervised, semi-supervised, and unsupervised learning are:

- **Supervised learning** aims to learn a function that, given a sample of data and desired outputs, approximates a function that maps inputs to outputs.
- **Semi-supervised learning** aims to label unlabeled data points using knowledge learned from a small number of labeled data points.
- **Unsupervised learning** does not have (or need) any labeled outputs, so its goal is to infer the natural structure present within a set of data points.

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

Ans: 5 Examples of Classification Problems :

- **Logistic regression:** Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. ... It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.
- **Decision trees:** Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves.
- **Random forest:** 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. It performs better results for classification problems.
- **XGBoost:** XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.Why XGBoost must be a part of your machine learning toolkit.
- **Light GBM:** Light Gradient Boosted Machine, or LightGBM for short, is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm.

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

Ans: Process of classification consists of two phases :

1. Construction of the classifier
2. Usage of the classifier.

A classifier in machine learning is an algorithm that automatically orders or categorizes data into one or more of a set of “classes.” One of the most common examples is an email classifier that scans emails to filter them by class label: Spam or Not Spam.

The main goal of the Classification algorithm is to identify the category of a given dataset, and these algorithms are mainly used to predict the output for the categorical data. Multi-class Classifier: If a classification problem has more than two outcomes, then it is called as Multi-class Classifier.

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

Ans: Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges.

Support Vectors are simply the coordinates of individual observation. The SVM classifier is a frontier that best segregates the two classes (hyper-plane/ line).

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.

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

Ans: The following are some of the benefits and drawbacks of SVM:

- **Benefits:**
  - 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.
- **Drawbacks:**
  - 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.

6. Go over the kNN model in depth ?

Ans: kNN is the simplest machine learning algorithm to understand and also to explain. It is a versatile algorithm i.e. useful for both classification and regression. It has one big advantage is that kNN ha no pre assumption about the data. **Let the data speak for itself**.

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'.

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

Ans: Training error here is the error you'll have when you input your training set to your KNN as test set. Since your test sample is in the training dataset, it'll choose itself as the closest and never make mistake. For this reason, the training error will be zero when K = 1, irrespective of the dataset.

kNN produces predictions by looking at the k nearest neighbours of a case x to predict its y, so that's fine. In particular, the kNN model basically consists of its training cases - but that's the cross validation procedure doesn't care about at all.

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

Ans: KNN can be used for classification — the output is a class membership (predicts a class — a discrete value). An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

KNN classifier does not have any specialized training phase as it uses all the training samples for classification and simply stores the results in memory. KNN is a non-parametric algorithm because it does not assume anything about the training data. This makes it useful for problems having non-linear data.

9. Create the kNN algorithm ?

Ans: The k-nearest neighbor algorithm is imported from the scikit-learn package.

- Create feature and target variables.
- Split data into training and test data.
- Generate a k-NN model using neighbors value.
- Train or fit the data into the model.
- Predict the future.

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

Ans: A decision tree is a tree-like model that acts as a decision support tool, visually displaying decisions and their potential outcomes, consequences, and costs. Drawing a decision tree diagram starts from left to right and consists of **burst** nodes that split into different paths.

There are three different types of nodes: chance nodes, decision nodes, and end nodes. A chance node, represented by a circle, shows the probabilities of certain results. A decision node, represented by a square, shows a decision to be made, and an end node shows the final outcome of a decision path.

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

Ans: In a decision tree analysis, the decision-maker has usually to proceed through the following six steps:

- Define the problem in structured terms.
- Model the decision process.
- Apply the appropriate probability values and financial data.
- Solve the decision tree.
- Perform sensitivity analysis.

12. Describe in depth the decision tree algorithm ?

Ans: Decision Tree algorithm belongs to the family of supervised learning algorithms. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).

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

Ans: The inductive bias (also known as learning bias) of a learning algorithm is the set of assumptions that the learner uses to predict outputs of given inputs that it has not encountered. The kind of necessary assumptions about the nature of the target function are subsumed in the phrase inductive bias.

Overfitting makes the model relevant to its data set only, and irrelevant to any other data sets. Some of the methods used to prevent overfitting include ensembling, data augmentation, data simplification, and cross-validation.

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

Ans: Advantages and Disadvantages of Decision Trees in Machine Learning. 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.

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

Ans: Appropriate Problems for Decision Tree Learning are: Instances are represented by attribute-value pairs. The target function has discrete output values.

- Disjunctive descriptions may be required.
- The training data may contain errors.
- The training data may contain missing attribute values.

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

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. The fundamental difference is that in Random forests, only a subset of features are selected at random out of the total and the best split feature from the subset is used to split each node in a tree, unlike in bagging where all features are considered for splitting a node.

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. There are two measures of importance given for each variable in the random forest. The first measure is based on how much the accuracy decreases when the variable is excluded.The second measure is based on the decrease of Gini impurity when a variable is chosen to split a node.