Name: Zeenat **Class: TYIT** **Roll no: 578**

Assignment (practise set-1)

Q1. Explain about Classification and Prediction.

Ans:

Classification:

1. The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data.
2. Classification is about determining a (categorial) class (or label) for an element in a dataset
3. Classification, data is grouped into categories based on a training dataset.
4. Classification is categorization of the things or data that we already have with us. This categorization can be based on any kind of technique or algorithms
5. Classification is mostly based on our current or past assumptions
6. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, Yes or No, 0 or 1, Spam or Not Spam, cat or dog, etc. Classes can be called as targets/labels or categories.

Prediction:

1. Prediction through machine learning or deep learning can be done in a number of different ways, depending on the underlying algorithm that is used.
2. In prediction, a classification/regression model is built to predict the outcome(continuous value)
3. Prediction may be a kind of classification
4. They often rely on algorithms designed for classification, clustering, pattern recognition and image recognition.
5. As the name suggests, predictive models are designed to predict unknown values, properties or events.
6. Prediction is mostly based on our future assumptions
7. Prediction is like saying something which may going to be happened in future.

Classification Models

1. Naive Bayes:

Naive Bayes is a classification algorithm that assumes that predictors in a dataset are independent.

This means that it assumes the features are unrelated to every other.

2. K-Nearest Neighbors:

K-Nearest Neighbor is a classification and prediction algorithm that is used to divide data into classes based on the distance between the data points.

K-Nearest Neighbor assumes that data points which are close to one another must be similar and hence, the data point to be classified will be grouped with the closest cluster.

3. Decision Trees:

A Decision Tree is an algorithm that is used to visually represent decision-making.

A Decision Tree can be made by asking a yes/no question and splitting the answer to lead to another decision.

The question is at the node and it places the resulting decisions below at the leaves.

Example:
In a hospital, the grouping of patients based on their medical record or treatment outcome is considered *classification*, whereas, if you use a classification model to predict the treatment outcome for a new patient, it is considered a *prediction*.
The prediction of *numerical* (continuous) variables is called *regression*.

Q2. A data set is given to you about utilities fraud detection. You have built a classifier model and achieved a performance score of 98.5%.
Is this a good model? If yes, justify. If not, what can you do about it?

Ans:

Data set about utilities fraud detection is not balanced enough i.e. imbalanced. In such a data set, accuracy score cannot be the measure of performance as it may only be predict the majority class label correctly but in this case our point of interest is to predict the minority label. But often minorities are treated as noise and ignored. So, there is a high probability of misclassification of the minority label as compared to the majority label. For evaluating the model performance in case of imbalanced data sets, we should use Sensitivity (True Positive rate) or Specificity (True Negative rate) to determine class label wise performance of the classification model. If the minority class label's performance is not so good, we could do the following:

1. We can use under sampling or over sampling to balance the data.
2. We can change the prediction threshold value.
3. We can assign weights to labels such that the minority class labels get larger weights.
4. We could detect anomalies.

Q3. Mining Quantitative Association Rules

Ans:

1. An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database.
2. It determines the set of items that occurs together in the dataset.
3. Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam)
4. Association rule learning is one of the very important concepts of machine learning, and it is employed in Market Basket analysis, Web usage mining, continuous production, etc.
5. Market basket analysis is a technique used by the various big retailer to discover the associations between items.
6. Association rule learning works on the concept of If and Else Statement. If element is called antecedent, and then statement is called as Consequent
7. Measure the associations between thousands of data items, there are several metrics. These metrics are Support, Confidence, Lift .
8. Support is the frequency of A or how frequently an item appears in the dataset
9. Confidence indicates how often the rule has been found to be true
10. Lift It is the strength of any rule

Q4. The k-means partitioning algorithm elaborate

Ans:

k-means clustering algorithm One of the most used clustering algorithm. It is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. It allows to group the data according to the existing similarities among them in k clusters, given as input to the algorithm. A cluster refers to a collection of data points aggregated together because of certain similarities. K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

The k-means clustering algorithm mainly performs two tasks:

Determines the best value for K center points or centroids by an iterative process.
Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

How the K-means algorithm works:
To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

Fuzzy c-means(Logic) Fuzzy Logic is a form of Knowledge representation sutabil for notation that cannot be defined precisely but which depend upon their contex.

Q5. Rough set theory: definition – rule induction – feature selection- rough sets in data mining

Ans

Rough Set Theory (RST), proposed by Z Pawlak, is a new mathematical approach to vagueness and uncertainty. Tools based on RST are found to be useful in addressing data mining tasks such as classification, clustering and rule mining. In RST all computations are performed directly on the supplied data and works by making use of the granularity structure of the data.

Rule induction:

1. Rule induction is a technique that creates "if–else–then"-type rules from a set of input variables and an output variable.
2. Rule induction is an important technique of data mining or machine learning.
3. it more precise using fundamental definitions of rough set theory
4. A typical rule induction technique, such as Quinlan's C5, can be used to select variables because, as part of its processing, it applies information theory calculations in order to choose the input variables (and their values) that are most relevant to the values of the output variables.
5. Therefore, the least related input variables and values get pruned and disappear from the tree. Once the tree is generated, the variables chosen by the rule induction technique can be noted in the branches and used as a subset for further processing and analysis.
6. Tools based on RST are found to be useful in addressing data mining tasks such as classification, clustering and rule mining.

An application in agriculture. By applying the proposed algorithm, a set of significant rules are generated. These rules are expected to be helpful to the farmers of the state to design their farming plans, which will enable them to improve their coconut production.

Feature selection:

1. Feature selection aims to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features.
2. Statistical-based feature selection methods involve evaluating the relationship between each input variable and the target variable using statistics and selecting those input variables that have the strongest relationship with the target variable.
3. The main aim of feature selection (FS) is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features.

4. In many real world problems FS is a must due to the abundance of noisy, irrelevant or misleading features.
5. For instance, by removing these factors, learning from data techniques can benefit greatly.
6. A detailed review of feature selection techniques devised for classification tasks can be found in (Dash & Liu, 1997).

Rough sets in data mining:

1. Rough Set Theory (RST), proposed by Z Pawlak, is a new mathematical approach to vagueness and uncertainty.
2. The main goal of the rough set analysis is induction of approximations of concepts.
3. Rough sets constitutes a sound basis for KDD.
4. It offers mathematical tools to discover patterns hidden in data.
5. It can be used for feature selection, feature extraction, data reduction, decision rule generation, and pattern extraction (templates, association rules) etc.
6. Tools based on RST are found to be useful in addressing data mining tasks such as classification, clustering and rule mining.
7. In RST all computations are performed directly on the supplied data and works by making use of the granularity structure of the data.

Basic problems in data analysis solved by Rough Set:

1. Characterization of a set of objects in terms of attribute values.

2. Finding dependency between the attributes.

3. Reduction of superfluous attributes.

4. Finding the most significant attributes.

5. Decision rule generation.

Q6. Hidden Markov Models

Ans:

The Hidden Markov Model (HMM) is an analytical Model where the system being modeled is considered a Markov process with hidden or unobserved states. The first-order Markov process is often simply called the Markov process. If it is in a discrete space, it is called the Markov chain. Machine learning and pattern recognition applications, like gesture recognition & speech handwriting, are applications of the Hidden Markov Model. HMM provides solution of three problems : evaluation, decoding and learning to find most likelihood classification.

Example: Sunlight can be the variable and sun can be the only possible state. The structure of Hidden Markov model is restricted to the fact that basic algorithms can be implemented using matrix representations**.**

For instance, Hidden Markov Models are similar to Markov chains, but they have a few hidden states. Since they're hidden, you can't be see them directly in the chain, only through the observation of another process that depends on it.

Since, HMM is rich in mathematical structure it can be implemented for practical applications.
This can be achieved on two algorithms called as:
Forward Algorithm.
Backward Algorithm.

Applications of HMM:

i. Speech Recognition.
ii. Gesture Recognition.
iii. Language Recognition.
iv. Motion Sensing and Analysis.
v. Protein Folding**.**

Q7. Support Vector Machines

Ans:

1. Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms.

2. It is used for Classification as well as Regression problems.

3. The SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.

4. This best decision boundary is called a hyperplane.

5. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

6. SVM algorithm can be used for Face detection, image classification, text categorization, etc.

Types of SVM:

Linear SVM:

Linear SVM is used for linearly separable data. (which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier).

Non-linear SVM:

Non-Linear SVM is used for non-linearly separated data. (which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.)

Example: There is a cat that also has some features of dogs if we want a model that identify it is a cat or dog this type of model create using the SVM algorithm. First we train our model to images of cats and dogs so it can learn about different features of cats and dogs and after that we test it. So as support vector creates a decision boundary between cat and dog and choose extreme cases it will see the extreme case of cat and dog. On the basis of the support vectors, it will classify it as a cat.

Q8. Reinforcement Learning

Ans:

1. Reinforcement learning is a type of Machine Learning where an agent learns to behave in a environment by performing actions and seeing the results.

2. Reinforcement learning is one of three basic machine learning paradigms, alongside supervised learning and unsupervised learning.

3. Key areas of Interest is Environment , Action, Reward and State

4. The mathematical approach for mapping a solution in reinforcement learning is called Markov Decision Process (MDP)

Reinforcement Learning Definitions:

1. Agent: The RL algorithm that learns from trial and error

2. Environment: The world through which the agent moves

3. Action (A): All the possible steps that the agent can take

4. State (S): Current condition returned by the environment

5. Reward (R): An instant return from the environment to appraise the last action

6. Policy (π): The approach that the agent uses to determine the next action based on the current state

7. Value (V): The expected long-term return with discount, as opposed to the short-term reward R

8. Action-value (Q): This similar to Value, except, it takes an extra parameter, the current action (A)

Applications:

1. Robotics for industrial automation.

2. Machine learning and data processing

3. It helps you to create training systems that provide custom instruction and materials according to the requirement of students.

4. Business strategy planning

Q9. Statistical Learning

Ans:

1. Statistical Learning is a set of tools for understanding data.
2. These tools broadly come under two classes: supervised learning & unsupervised learning.
3. Statistical learning theory was introduced in the late 1960s but untill 1990s it was simply a problem of function estimation from a given collection of data.
4. Statistical Learning is Artificial_Intelligence is a set of tools for machine learning that uses statistics and functional analysis
5. Statistical learning is used to build predictive models based on the data.
6. Statistical Learning is math intensive which is based on the coefficient estimator and requires a good understanding of your data. On the other hand, Machine Learning identifies patterns from your dataset through the iterations which require a way less of human effort.
7. Statistical learning can be used to build applications for computer vision, text analytics, voice recognition, etc.

Q10. Regression and Classification with Linear Models

Ans

1. Regression and Classification algorithms are Supervised Learning algorithms. Both the algorithms are used for prediction in Machine learning and work with the labeled datasets. But the difference between both is how they are used for different machine learning problems.
2. The main difference between Regression and Classification algorithms that Regression algorithms are used to predict the continuous values such as price, salary, age, etc. and Classification algorithms are used to predict/Classify the discrete values such as Male or Female, True or False, Spam or Not Spam, etc.

   Classification:

Classification is a process of finding a function which helps in dividing the dataset into classes based on different parameters. In Classification, a computer program is trained on the training dataset and based on that training, it categorizes the data into different classes.

The task of the classification algorithm is to find the mapping function to map the input(x) to the discrete output(y).

Example: The best example to understand the Classification problem is Email Spam Detection. The model is trained on the basis of millions of emails on different parameters, and whenever it receives a new email, it identifies whether the email is spam or not. If the email is spam, then it is moved to the Spam folder.

Types of ML Classification Algorithms:

Classification Algorithms can be further divided into the following types:

Logistic Regression
K-Nearest Neighbours
Support Vector Machines
Kernel SVM
Naïve Bayes
Decision Tree Classification
Random Forest Classification

Regression:
Regression is a process of finding the correlations between dependent and independent variables. It helps in predicting the continuous variables such as prediction of Market Trends, prediction of House prices, etc.

The task of the Regression algorithm is to find the mapping function to map the input variable(x) to the continuous output variable(y).

Example: Suppose we want to do weather forecasting, so for this, we will use the Regression algorithm. In weather prediction, the model is trained on the past data, and once the training is completed, it can easily predict the weather for future days.

Types of Regression Algorithms:

1. Simple Linear Regression

2. Multiple Linear Regression
3. Polynomial Regression
4. Support Vector Regression
5. Decision Tree Regression
6. Random Forest Regression