Assignment no.- 5

Aim - Demonstrate performing Regression on data sets.

Course outcomes -

CO2: Demonstrate the classification, clustering & etc. in large data sets.

CO3: Ability to add mining algorithm as a component to the exiting tools.

CO4: To learn physical design, logical design & enabling technologies of Internet of things.

Software & Hardware Requirements -

Theory -

Weka has a large number of regression algorithms. The large number of machine learning algorithms supported by Weka is one of the biggest benefits of using the platform.

Regression Algorithm

- Regression is a supervised machine learning technique which is used to predict continuous values.
- The ultimate goal of the regression algorithm is to plot a best-fit line or a curve between the data.
- The three main metrics that are used for evaluating the trained regression model are variance, bias & error

Types of regression algorithm

- Linear Regression
- k-Nearest Neighbors
- Decision Tree
- Support Vector Machines
- Multi-Layer Perceptron

1) Linear Regression

- Linear regression only supports regression type problems.
- It works by estimating coefficients for a line or hyperplane that best fits the training data.
- It is a very simple regression algorithm, fast to train & can have great performance if the output variable for your data is linear combination of your inputs.
- The performance of linear regression can be reduced if your training data has

input attributes that are highly correlated.

- Weka can detect & remove highly correlated input attributes automatically by setting eliminateColinearAttributes to True, which is the default.
- Weka can automatically perform feature select^n to only select those relevant attributes by setting the attributeSelectionMethod. This is enabled by default & can be disabled.
- Weka implementation uses a ridge regularizat^n technique in order to reduce the complexity of learned model.
- It does this by minimizing the square of the absolute sum of the learned coefficients, which will prevent any specific coefficient from becoming too large.

K-Nearest Neighbors

- The k-nearest neighbors algorithm supports both classification & regression. It is also called KNN for short.
- It works by storing the entire training dataset & querying it to locate the k most similar training patterns when making a prediction.
- It is simple algorithm, but one that does not assume very much about the problem other than that the distance between data instances is meaningful in making predictions. As such, it often

achieves very good performance.

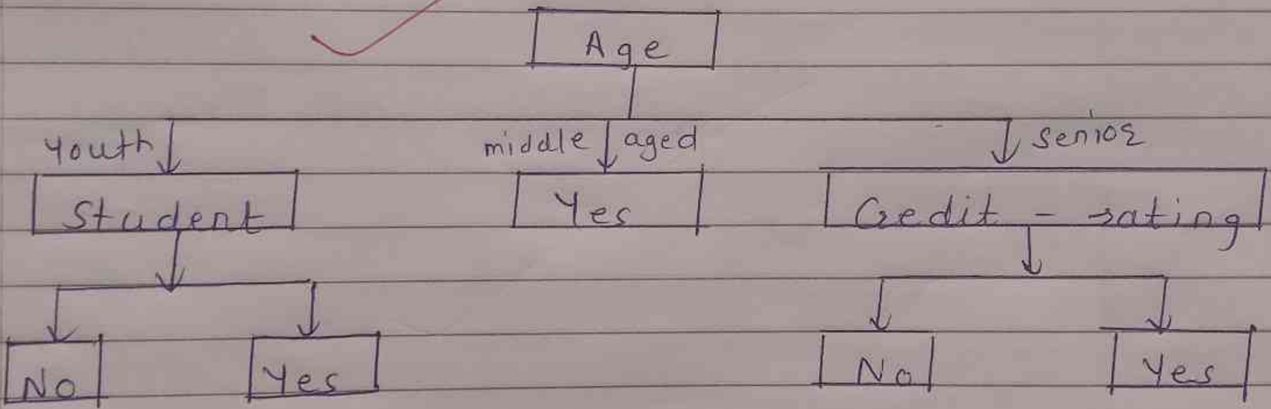
- When making predictions on regression problems, kNN will take the mean of the k most similar instances in the training dataset.
- In Weka kNN is called IBk which stand for Instance Based k.
- The size of the neighborhood is controlled by the k parameter. eg -
  if set to 1, then predictions are made using the single most similar training instance to a given new pattern for which a prediction is requested.
- Common values for k are 3, 7, 11 & 21 larger for larger dataset size.
- Weka can automatically discover a good value for k using cross validation inside the algorithm by setting the crossValidate parameter to True.

Decision Tree

- Decision trees can support classification & regression problems.
- Decision trees are most recently refered to as classification And Regression Trees or CART.
- They work by creating a tree to evaluate an instance of data, start at the root of the tree & moving down to the leaves until a prediction

can be made.

- The process of creating a decision tree works by greedily selecting the best split point in order to make predictions & repeating the process until the tree is a fixed depth.
- After the tree is construct, it is pruned in order to improve the model's ability to generalize to new data.



Support Vector Regression

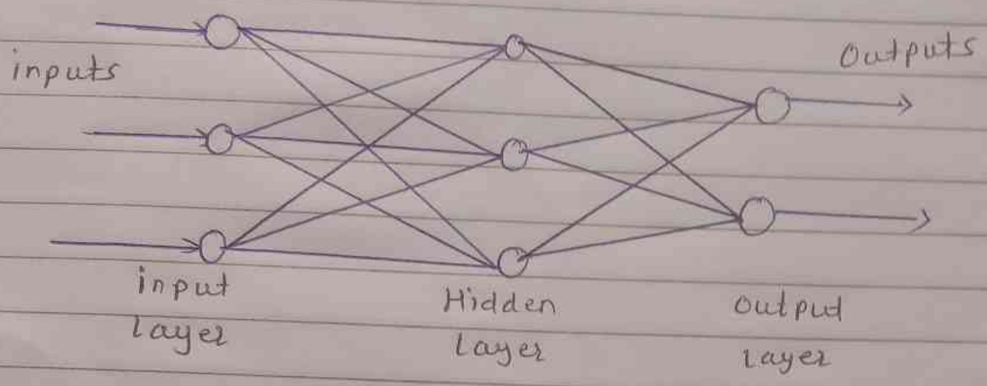
- support vector machines were developed for binary classification problems, although extensions to the technique have been made to support multi-class classification & regression problems.
- The adaptation of SVM for regression is called Support Vector Regression or SVR for short.
- SVM was developed for numerical input variables, although will automatically

convert nominal values to numerical values

- Input data is also normalized before being used.
- Unlike SVM that finds a line that best separates the training data into classes, SVR works by finding a line of best fit that minimizes the error of cost function. This is done using an optimizat^n process that only considers those data instances in the training dataset that are closest to the line with the minimum cost. These instances are called support vectors.
- SVM used application face detection, email classification, gene classification, intrusion detection.

Multi-Layer Perceptron

- The multi-layer perceptron algorithms support both regression & classification problems.
- It is also called artificial neural network or simply neural network.
- It is an algorithm inspired by a model of biological neural networks in the brain where small processing units called neurons are organized into layers that if configured well are capable of approximating any function.
-



- In the multi-layer perceptron diagram above, we can see that there are three inputs & thus three input nodes & the hidden layer has three nodes The output layer gives two outputs, therefore there are two output nodes.
- The nodes in the input layer take input & forward it for further process in the diagram above the nodes in the input layer forwards their output to each of the three nodes in the hidden layer, & in the same way the hidden layer processes the informat^{n} & passes it to the output layers.

Conclusion -

In this practical I have studied about the regression algorithm & the types of regression algorithm.