UNIT-IV

Object Segmentation: Regression Vs Segmentation – Supervised and Unsupervised Learning, Tree Building – Regression, Classification, Overfitting, Pruning and Complexity, Multiple Decision Trees etc. Time Series Methods: Arima, Measures of Forecast Accuracy, STL approach, Extract features from generated model as Height, Average Energy etc and Analyze for prediction

Regression Vs Segmentation:

Regression:

Regression analysis is a form of predictive modelling technique which investigates the relationship between a **dependent** (target) and **independent variable (s)** (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

Regression refers to average relationship between two or more variables. One of these variables is called the dependent or the explained variable and the other variable is independent or the explaining variable. Regression is one of the statistical method that is used to estimate the unknown of one variable from the known value of the related variable.

For Example regression helps finance and investment professionals and other business professionals. It helps in predicting company sales depending upon weather, previous sales, GDP growth etc., Regression considers a set of random variables to find the mathematical relationship between them. The relationship is mostly in the form of straight line which approximates the separate data points. There are two types of regression namely simple linear regression and multiple linear regression.

Simple linear regression represents the relationship between two variables where one of them is independent variable X and other variable is dependent variable Y.

Multiple linear regression model that contains multiple independent variables is referred to as multiple regression model. An equation

Segmentation is the process of segmenting the data according to the company's requirement to refine the analyses based on the defined context through certain tools. The Main objective is to understand the customers better and then obtain actionable data to improve the outcome. It allows to filter the analyses depending upon some elements.

Object Segmentation is the process of segmenting the moving objects. That means locating the objects and its boundaries from background. The frames that are captured contain noises that are extracted from captured frames. The objects after filtering are passed to optical flow vectors for thresholding operation. The unwanted objects are then removed. In filtering process, the holes are created. They are closed through a morphological operation.

Segmentation:

Segmentation is the Process of segmenting the data according to the company's requirement to refine the analyses based on the defined context through certain tools.

Objective segmentation is the process of segmenting the moving objects. Locating the objects and its boundaries from background. The frames that are captured contain noise that are extracted from captured frames. The Objects after filtering are passed to optical flow vectors for thresholding operation. The unwanted objects are then removed. In the filtering process the holes are created. They are closed through morphological operation.

Supervised Learning:

- Supervised learning allows you to collect data or produce a data output from the previous experience.
- Helps you to optimize performance criteria using experience
- Supervised machine learning helps you to solve various types of real-world computation problems

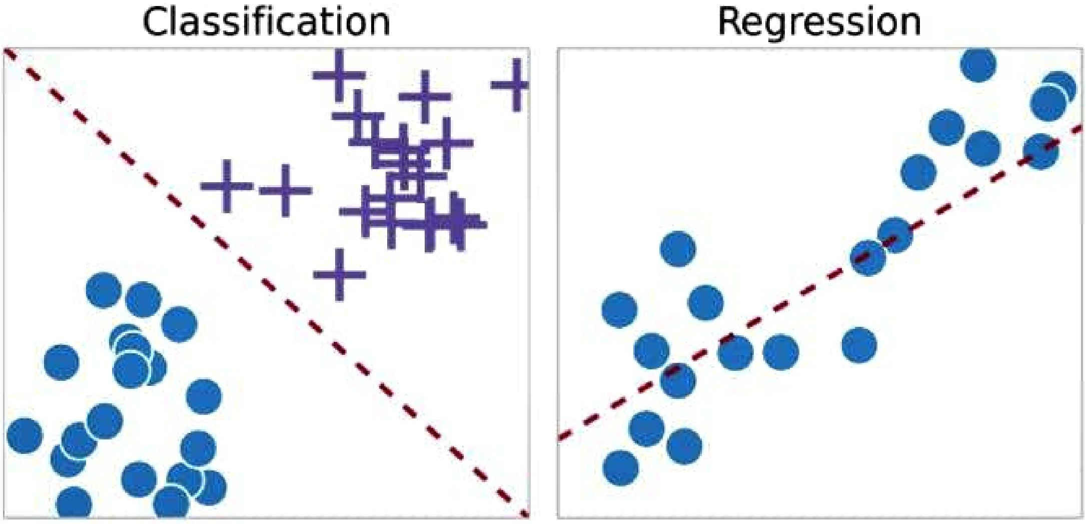
Example:

For example, you want to train a machine to help you predict how long it will take you to drive home from your workplace. Here, you start by creating a set of labeled data. This data includes

- Weather conditions
- Time of the day
- Holidays

All these details are your inputs. The output is the amount of time it took to drive back home on that specific day.

Types of Supervised Machine Learning Techniques



Regression:

Regression technique predicts a single output value using training data.

Example: You can use regression to predict the house price from training data. The input variables will be locality, size of a house, etc.

Classification:

Classification means to group the output inside a class. If the algorithm tries to label input into two distinct classes, it is called binary classification. Selecting between more than two classes is referred to as multiclass classification.

Example: Determining whether or not someone will be a defaulter of the loan.

Strengths: Outputs always have a probabilistic interpretation, and the algorithm can be regularized to avoid overfitting.

Weaknesses: Logistic regression may underperform when there are multiple or non-linear decision boundaries. This method is not flexible, so it does not capture more complex relationships.

Unsupervised Learning

- Unsupervised machine learning finds all kind of unknown patterns in data.
- Unsupervised methods help you to find features which can be useful for categorization.
- It is taken place in real time, so all the input data to be analyzed and labeled in the presence of learners.
- It is easier to get unlabeled data from a computer than labeled data, which needs manual intervention.

Example:

She knows and identifies this dog. A few weeks later a family friend brings along a dog and tries to play with the baby.

Baby has not seen this dog earlier. But it recognizes many features (2 ears, eyes, walking on 4 legs) are like her pet dog. She identifies a new animal like a dog. This is unsupervised learning, where you are not taught but you learn from the data (in this case data about a dog.) Had this been supervised learning, the family friend would have told the baby that it's a dog.

Unsupervised learning problems further grouped into clustering and association problems.

sample

Cluster/group

Clustering

Clustering is an important concept when it comes to unsupervised learning. It mainly deals with finding a structure or pattern in a collection of uncategorized data. Clustering algorithms will process your data and find natural clusters (groups) if they exist in the data. You can also modify how many clusters your algorithms should identify. It allows you to adjust the granularity of these groups.

Association

Association rules allow you to establish associations amongst data objects inside large databases. This unsupervised technique is about discovering exciting relationships between variables in large databases. For example, people that buy a new home most likely to buy new furniture.

Other Examples:

- A subgroup of cancer patients grouped by their gene expression measurements
- Groups of shopper based on their browsing and purchasing histories
- Movie group by the rating given by movies viewers

Supervised Learning	Unsupervised Learning
Supervised learning algorithms are trained using labeled data.	Unsupervised learning algorithms are trained using unlabeled data.
Supervised learning model takes direct feedback to check if it is predicting correct output or not.	Unsupervised learning model does not take any feedback.
Supervised learning model predicts the output.	Unsupervised learning model finds the hidden patterns in data.
In supervised learning, input data is provided to the model along with the output.	In unsupervised learning, only input data is provided to the model.
The goal of supervised learning is to train the model so that it can predict the output when it is given new data.	The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset.
Supervised learning needs supervision to train the model.	Unsupervised learning does not need any supervision to train the model.
Supervised learning can be categorized in Classification and Regression problems.	Unsupervised Learning can be classified in Clustering and Associations problems .
Supervised learning can be used for those cases where we know the input as well as corresponding outputs.	Unsupervised learning can be used for those cases where we have only input data and no corresponding output data.
Supervised learning model produces an accurate result.	Unsupervised learning model may give less accurate result as compared to supervised learning.

Supervised learning is not close to true Artificial intelligence as in this, we first train the model for each data, and then only it can predict the correct output.	Unsupervised learning is more close to the true Artificial Intelligence as it learns similarly as a child learns daily routine things by his experiences.
It includes various algorithms such as Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decision tree, Bayesian Logic, etc.	It includes various algorithms such as Clustering, KNN, and Apriori algorithm.

Supervised and Unsupervised Learning:

Parameters	**Supervised Learning**	**Unsupervised Learning**
Process	Input and Output Variables are Provided	Only Input Data is Provided
Input Data	The Algorithms are trained through labelled data.	The Algorithms are trained through unlabelled data.
Number of Classes	Number of classes is known already	Number of classes is not known.
Computational Complexity	It is a Simple Method	It is Complex
Use of Data	It uses Training data to learn a link in between Input and Outputs	It does not need any output data.
Algorithms Used	Support Vector Machines, Random Forest, Neural Network, Classification Trees, Linear and Logistics Regression	Unsupervised algorithms are divided into categories like k-means, cluster algorithms, hierarchal clsutering
Real Time Learning	Learning methods works offline	Learning methods works in real time
Accuracy of Results	It is Accurate and Trusty worthy method	It is Less accurate and trusty worthy method.
Main Drawback	Classification of Bigdata is a	Because data used in

	challenging task	unsupervised learning is labelled and unknown it is difficult to obtain precise information regarding the data sorting.

TREE BUILDING

Decision Tree is a diagrammatic representation of alternative courses of action and sequence of states of nature.

The Various steps involved in decision tree analysis are

1. Determine the number of decisions to be taken and the alternative strategies available for each decision in sequential manner.
2. Determine the outcome(or event) which may occur from each alternative strategy.
3. Construct a tree diagram representing the order in which decisions are taken and outcomes are occurring. The decision tree diagram begins from left side and move towards right side.
4. Determine the Probabilities of occurrences of each state of nature.
5. Determine the pay off values for each pair (or combination) of state of nature and course of action.
6. Calculate expected pay off value for each course of action starting from right side of the decision trees.
7. Select the course of action (or alternative strategy) with the best expected pay off value.
8. With backwards from last decision point to first decision point and at each decision point repeat the steps from step 4 to step7.

Time Series Methods:

The time series data is ordered sequence of observations on quantative variable that is measured over equally spaced time interval. The time series are used in statistics, signal Processing, econometrics, pattern recognition, weather forecasting.

The time series analysis is used to analyse the time series data and to forecast the future value of variable under consideration. The data in time series analysis contain set of identifiable components and random errors that make the pattern difficult to identify.

ARIMA is one of the time series methods. ARIMA stands for Autoregressive Integrated Moving Average. In time series analyses it is generalization. This model is fitted to time series data to understand the data or to predict future points in the series. They are applied in certain cases where the data shows evidence of non stationary and where the initial differencing step is applied to minimize the non stationary.

ARIMA is well known stochastic method that is used to analyze the time series data sets. It contains three time series components such AR (Auto-Regressive), I (Integrated), MA (Moving Average). Every component minimizes the final residuals indicated as p, d, q respectively. Integrated (I) is initial step in ARIMA to extract trend data. This is possible by differencing the data from earlier values. The first step is indicated by (0,1,0) and the second step is indicated by (0,2,0). This continues until data becomes trend less. The time series will become trend less after differencing. The second step in ARIMA is auto regression to analyze the time series data. Once this data becomes stationary the AR component gets activated. The auto regressive step will extract the previous value from present value. This is obtained through simple linear regression by considering independent or predictor variables as it time lagged values.

$Y=C+Y+Y+Y.....+,ib$

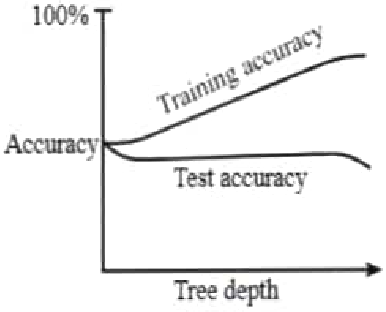
OVERFITTING

The Decision Trees will be at high risk to overfit the training data to a high degree when they are not pruned.

Over fitting might occurs when models select the noise or errors in training data set. Therefore overfitting will be seen viewed as performance gap between training and test data.

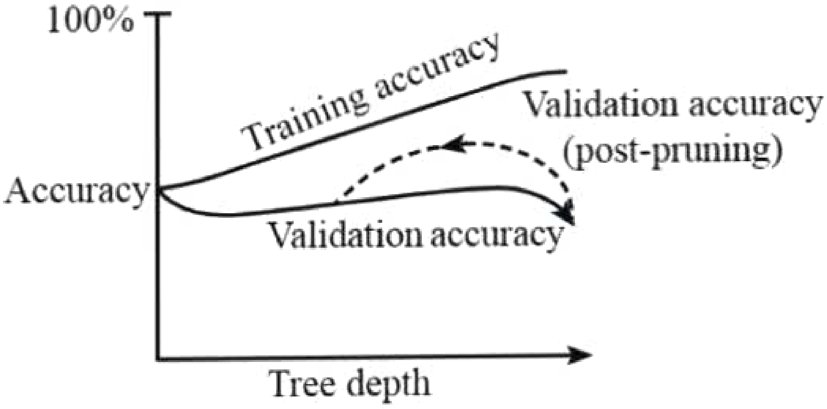
A General method to reduce overfitting in decision trees is decision tree pruning. There are two methods namely

1. Post Pruning
2. Pre Pruning



The above figure shows relationship between Tree Depth and Overfitting

The following figure shows the Reduced Error Pruning



Dr.G.Naga Satish

STL APPROACH:

STL Stands for Seasonal and Trend Decomposition using Loess. Loess is a method used to estimate the non linear relationships. STL is versatile, robust ad stastical method of decomposing the time series data into three components as Trend, Seasonality and residual.

The purpose of development of STL is to build a decomposition procedure and a computer implementation that can satisfy the below criteria.

1. A Simple design and a straight forward use.

2. Easy computer implementation and a fast computation for long time series.

3. Specification of observations for each cycle of seasonal component to an integer greater than 1.

4. Flexible to specify the variation in trend and seasonal components.

5. Robust trend and seasonal components that cannot be distorted by transient and aberrant behaviour of data.

6. It should be able to decompose the series with missing values.

STL contains a set of smoothing operations that employ same smoother locally weighted regression or loess with only one exception.

STL has various parameters that must be selected by the data analysts.

$n_{(p)}$= Number of observations in every cycle of seasonal component.

$n_{(i)=}$ Number of Passes through inner loop.

$n_{(o)}$= Number of robustness iterations of outer loop.

$n_{(l)}$= Smoothing parameters for low pass filter

$n_{(t)}$= Smoothing parameter for trend component.

$n_{(x)}$= Smoothing parameter for seasonal component.

STL can be implemented in any environments for the purpose of graphics and data analysis.

Pruning:

Pruning is a technique that is used to reduce the size of decision tress by removing parts of the tree which provide less power for classifying instances. It reduces the complexity of final classifier and even improves the accuracy by reduction of over fitting.

Pruning occurs either in top-down or bottom-up. The top-down pruning traverses the nodes and trims the sub tree starting at root. The bottom up pruning begins at leaf nodes.

Pruning when applied on decision trees will remove one or more sub trees from it. There are various methods for decision tree pruning .They replaces the sub tree with leaf if classification accuracy is not reduced over pruning data set. Pruning increases number of classification errors on training set but improves classification accuracy on unseen data The Pruning techniques can be divided into two groups. The methods in first group will compute the probability of sub tree misclassification and then make the pruning decision through an independent test set called pruning data set. In Second group the iterative grow and prune method is used while creating a tree. These pruning techniques are as follows.

1. Cost Complexity Pruning
2. Reduced Error Pruning
3. Critical Value Pruning
4. Minimum Error Pruning
5. Pessimistic Error Pruning
6. Error Based Pruning
7. Optimal Pruning
8. Minimum Description Length Pruning

Extract features from generated model as Height, Average Energy etc and Analyze for prediction

Feature extraction is the process of extracting useful characteristics from data. It calculates the values from input images. A feature is also called as descriptor. It is defined as a function of one or more measurements by specifying certain quantifiable property of complete image or sub image or of single object. Three methods of feature extraction are performed that lead to three different results for every classification. They are compared each other. The best feature extraction method for very two classifications is used for the part and complete system together.

For example consider an image model that is generated. The feature extraction focuses around the measurement of geometric properties and surface characteristics of regions. The features extracted from this image model are

Type	Feature	Description
Y	(i, j)	Center of Gravity
I, R ,G, B	(G)	Mean Gray Value
L	L*a*b	Color Components of region
Y	h,w, A, L,R	Height, width, area, roundness, perimeter
Y	a	Average
Y	E	Energy
I, R, G,B	K	Deviation Contrasts

The features provide relevant information for classification. To reduce the computational time required, in pattern recognition process it is necessary to select features suitable for classification.

The extracted features are later on analysed for prediction. A Predictor is used for this purpose. The model is trained on labelled examples can be good and bad quality to retrieve quality classification model. A model is trained on labelled examples of salient and non-salient images for retrieving content classification model. While using a model to classify the data the output of image is a class label and a score matrix.

The predictor is selected due to its advantages. The advantage does not have the problem of Overfitting. The Overfitting occurs when model has multiple features related to the number of observations and results in poor predictive performance. This problem is relevant to consider while working with machine learning on images because number of features extracted from an image are large.

Measures of Forecast Accuracy:

Forecast accuracy is a method of deviation of prediction or forecast from the actual outcome. It is generally defined as

Error=Actual Demand-Forecast

Or

e=A-F

Forecast accuracy can be measured by using two methods

1. Mean Forecast Accuracy(MFE)
2. Mean Absolute Deviation (MAD)

Mean Forecast Accuracy (MFE):

Mean forecast error represents the deviation of forecast from actual demand. It is the mean of differences for every period in between number of period forecasts and actual demand for related periods. This error is mostly used as bias for following and adjusting forecasts. If it is positive then the forecasts are low compared to actual demand. And if it is negative then the forecasts will be high. Mean forecast deviation is defined as follows

MFE=$\frac{\sum_{i=1}^{n}(ei)}{n}$

If the value of MFE is be greater than zero then model tends to under forecast and if the value of MFE is less than zero then model tends to over forecast.

Mean Absolute Deviation (MAD)

Mean absolute deviation deviates the forecasted demand from actual demand. It is the mean deviation for every period absolute terms and respective period demand.

MAD=$\sum_{i=1}^{n}|ei|$

Uses of Forecast Error:

1. It is used to compare alternative forecasting models
2. It is used to forecast the model bias.
3. It is used to generate absolute size of forecast errors.
4. It is used to identify the forecast models which needed to be adjusted.