UNIT-III

3.1 Regression modelling

Regression is a statistical technique to determine the linear relationship between two or more variables. Regression is primarily used for prediction and causal inference. Regression shows the relationship between one independent variable (X) and a dependent variable (Y), as in the formula below:

$$Y = \beta_0 + \beta_1 X + u$$

The magnitude and direction of that relation are given by the slope parameter (þ1), and the status of the dependent variable when the independent variable is absent is given by the intercept parameter (þ0). An error term (u) captures the amount of variation not predicted by the slope and intercept terms. The regression coefficient (R2) shows how well the values fit the data.

Regression thus shows us how variation in one variable co-occurs with variation in another. What regression cannot show is causation; causation is only demonstrated analytically, through substantive theory. For example, a regression with shoe size as an independent variable and foot size as a dependent variable would show a very high regression coefficient and highly significant parameter estimates, but we should not conclude that higher shoe size causes higher foot size. All that the mathematics can tell us is whether or not they are correlated, and if so, by how much.

It is important to recognize that regression analysis is fundamentally different from ascertaining the correlations among different variables. Correlation determines the strength of the relationship between variables, while regression attempts to describe that relationship between these variables in more detail. The linear regression model (LRM)

The simple (or bivariate) LRM model is designed to study the relationship between a pair of variables that appear in a data set. The multiple LRM is designed to study the relationship between one variable and several of other variables. In both cases, the sample is considered a random sample from some population. The two variables, X and Y, are two measured outcomes for each observation in the dataset.

This is written in any number of ways, but will specified as:

$$Y = \beta_1 + \beta_2 X + u$$

where

- Y is an observed random variable (also called the response variable or the lefthand side variable).
- X is an observed non-random or conditioning variable (also called the predictor or right-hand side variable).
- þ1 is an unknown population parameter, known as the constant or intercept term.
- þ2 is an unknown population parameter, known as the coefficient or slope parameter.
- u is an unobserved random variable, known as the error or disturbance term.

3.2 Multivariate Analysis

Many statistical techniques focus on just one or two variables

Multivariate analysis (MVA) techniques allow more than two variables to be analysed at once

Multiple regression is not typically included under this heading, but can be thought of as a multivariate analysis.

Multivariate analysis (MVA) is based on the statistical principle of multivariate statistics, which involves observation and analysis of more than one statistical outcome variable at a time. In design and analysis, the technique is used to perform trade studies across multiple dimensions while taking into account the effects of all variables on the responses of interest.

MVA can help summarise the data

E.g. factor analysis and segmentation based on agreement ratings on 20 attitude statements

• MVA can also reduce the chance of obtaining spurious results

Uses for multivariate analysis

• design for capability (also known as capability-based design)
• inverse design, where any variable can be treated as an independent variable
• Analysis of Alternatives (AoA), the selection of concepts to fulfil a customer need
• analysis of concepts with respect to changing scenarios
• identification of critical design-drivers and correlations across hierarchical levels.
• Multivariate analysis can be complicated by the desire to include physics-based analysis to calculate the effects of variables for a hierarchical "system-of-systems".

MVA Methods

Two general types of MVA technique

– Analysis of dependence
    • Where one (or more) variables are dependent variables, to be explained or predicted by others
        – E.g. Multiple regression, PLS, MDA
– Analysis of interdependence
    • No variables thought of as “dependent”
    • Look at the relationships among variables, objects or cases
        – E.g. cluster analysis, factor analysis

Principal Components

• Identify underlying dimensions or principal components of a distribution
• Helps understand the joint or common variation among a set of variables
• Probably the most commonly used method of deriving “factors” in factor analysis (before rotation)
• The first principal component is identified as the vector (or equivalently the linear combination of variables) on which the most data variation can be projected
• The 2nd principal component is a vector perpendicular to the first, chosen so that it contains as much of the remaining variation as possible
• And so on for the 3rd principal component, the 4th, the 5th etc

Multivariate Normal Distribution

• Generalisation of the univariate normal
• Determined by the mean (vector) and covariance matrix
    • E.g. Standard bivariate normal

Broader MVA Issues

– EDA is usually very worthwhile
    • Univariate summaries, e.g. histograms
    • Scatterplot matrix
    • Multivariate profiles, spider-web plots
– Missing data
    • Establish amount (by variable, and overall) and pattern (across individuals)
    • Think about reasons for missing data
    • Treat missing data appropriately
        – e.g. impute, or build into model fitting

3.3 SVM & Kernel Methods

Support vector machines

The SVM is a machine learning algorithm which solves classification problems. It uses a flexible representation of the class boundaries. It implements automatic complexity control to reduce over fitting. It has a single global minimum which can be found in polynomial time. It is popular because it can be easy to use. It often has good generalization performance and the same algorithm solves a variety of problems with little tuning.

Kernels

Kernel methods are a class of algorithms for pattern analysis, whose best known member is the support vector machine(SVM). The general task of pattern analysis is to find and study general types of relations (for example clusters, rankings, principal components, correlations, classifications) in datasets. For many algorithms that solve these tasks, the data in raw representation have to be explicitly transformed into feature vector representations via a user-specified feature map: in contrast, kernel methods require only a user-specified kernel, i.e., a similarity function over pairs of data points in raw representation.

Kernel methods owe their name to the use of kernel functions, which enable them to operate in a high-dimensional, implicit feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the feature space. This operation is often computationally cheaper than the explicit computation of the coordinates. This approach is called the "kernel trick". Kernel functions have been introduced for sequence data, graphs, text, images, as well as vectors.

Algorithms capable of operating with kernels include the kernel perceptron, support vector machines (SVM), Gaussian processes, principal components analysis (PCA), canonical correlation analysis, ridge regression, spectral clustering, linear adaptive filters and many others. Anylinear model can be turned into a non-linear model by applying the kernel trick to the model: replacing its features (predictors) by a kernel function. Most kernel algorithms are based on convex optimization or eigen problems and are statistically well-founded. Typically, their statistical properties are analyzed using statistical learning theory.

Types of Kernels:

- String kernel is a kernel function that operates on strings, i.e. finite sequences of symbols that need not be of the same length
- Path kernels
- Tree kernels
- Graph kernel is a kernel function that computes an inner product on graphs. Graph kernels can be intuitively understood as functions measuring the similarity of pairs of graphs.
- Fisher Kernels is a function that measures the similarity of two objects on the basis of sets of measurements for each object and a statistical model
- Polynomial Kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.

3.4 Rule Mining

The goal of association rule finding is to extract correlation relationships in the large datasets of items. Ideally, these relationships will be causative. Many businesses are interested in extracting interesting relationships in order to raise profit. Scientists, on the other hand, are interested in discovering previously unknown patterns in their field of research.

An illustrative example of association rule mining is so-called market basket analysis. This process is based on transactional data, which are huge amounts of records of individual purchases. For example, department stores such as Wal-Mart keep track of all transactions in a given period of time. One transaction could be t1 = {italian bread, 1% milk, energizer batteries}. It can contain an arbitrary

(technically, it is limited to what's in the store) number of items. All items purchased belong to a set of items Ω. In the computer science jargon, such a set is typically called a universe. Another notational detail is that transaction t1 can also be called an itemset, because it contains a set of purchased items.

An association rule could then be

computer ⇒ financial_management_software

which means that a purchase of a computer implies a purchase of financial management software. Naturally, this rule may not hold for all customers and every single purchase. Thus, we are going to associate two numbers with every such rule. These two numbers are called support and confidence.

3.5 Cluster Analysis

Cluster analysis is a data exploration (mining) tool for dividing a multivariate dataset into "natural" clusters (groups). We use the methods to explore whether previously undefined clusters (groups) may exist in the dataset. For instance, a marketing department may wish to use survey results to sort its customers into categories.

3.5.1 Types of Data in Cluster Analysis

Interval-Scaled Attributes

Continuous measurements on a roughly linear scale

Binary Attributes

A contingency table for binary data

Nominal Attributes

A generalization of the binary variable in that it can take more than 2 states, e.g., red, yellow, blue, green

Ordinal Attributes

An ordinal variable can be discrete or continuous

Order is important, e.g., rank

Can be treated like interval-scaled

Ratio-Scaled Attributes

A positive measurement on a nonlinear scale, approximately at exponential scale

Attributes of Mixed Type

A database may contain all the six types of variables

–symmetric binary, asymmetric binary, nominal, ordinal, interval and ratio

3.6 Clustering methods

•Partitioning approach:

–Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors

–Typical methods: k-means, k-medoids, CLARANS

•Hierarchical approach:

–Create a hierarchical decomposition of the set of data (or objects) using some criterion

–Typical methods: Diana, Agnes, BIRCH, ROCK, CAMELEON

•Density-based approach:

–Based on connectivity and density functions

Typical methods: DBSACN, OPTICS, DenClue

•Grid-based approach:

–based on a multiple-level granularity structure

–Typical methods: STING, WaveCluster, CLIQUE

•Model-based:

–A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other

–Typical methods: EM, SOM, COBWEB

3.6.1 Partitioning Method

• Given a k, find a partition of k clusters that optimizes the chosen partitioning criterion

–Global optimal: exhaustively enumerate all partitions

–Heuristic methods: k-means and k-medoids algorithms

–k-means :

Each cluster is represented by the center of the cluster

–k-medoids or PAM (Partition around medoids) :

Each cluster is represented by one of the objects in the cluster.

The K-Means Clustering Method

•Given k, the k-means algorithm is implemented in four steps:

–Partition objects into k nonempty subsets

–Compute seed points as the centroids of the clusters of the current partition (the centroid is the center, i.e., mean point, of the cluster)

–Assign each object to the cluster with the nearest seed point.

–Go back to Step 2, stop when no more new assignment

3.6.2 Hierarchical approach

In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

•Agglomerative:

- This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

•Divisive:

- This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram.

3.6.3 Density-based Clustering

Clusters are dense regions in the data space, separated by regions of lower object density

A cluster is defined as a maximal set of densityconnected points

Discovers clusters of arbitrary shape

Method-DBSCAN

3.6.4 Grid-based Clustering

The grid-based clustering approach differs from the conventional clustering algorithms in that it is concerned not with the data points but with the value space that surrounds the data points.

The great advantage of grid-based clustering is its significant reduction of the computational complexity, especially for clustering very large data sets.

In general, a typical grid-based clustering algorithm consists of the following five basic steps

1. Creating the grid structure, i.e., partitioning the data space into a finite number of cells.
2. Calculating the cell density for each cell.
3. Sorting of the cells according to their densities.
4. Identifying cluster centers.
5. Traversal of neighbor cells.

3.6.5 Model-based Clustering

Model-based clustering is a major approach to clustering analysis. This chapter introduces model-based clustering algorithms. First, we present an overview of model-based clustering. then, we introduce Gaussian mixture models, model-based agglomerative hierarchical clustering, and the

expectation-maximization (EM) algorithm. Finally, we introduce model-based clustering and two model-based clustering algorithms.

The word model is usually used to represent the type of constraints and geometric properties of the covariance matrices. In the family of model-based clustering algorithms, one uses certain models for clusters and tries to optimize the fit between the data and the models. In the model-based clustering approach, the data are viewed as coming from a mixture of probability distributions, each of which represents a different cluster. In other words, in model-based clustering, it is assumed that the data are generated by a mixture of probability distributions in which each component represents a different cluster.

Each component is described by a density function and has an associated probability or "weight" in the mixture.

In principle, we can adopt any probability model for the components, but typically we will assume that components are p-variate normal distributions.

Thus, the probability model for clustering will often be a mixture of multivariate normal distributions.

3.7 Clustering High Dimensional Data

Clustering high-dimensional data is the cluster analysis of data with anywhere from a few dozen to many thousands of dimensions. Such high-dimensional spaces of data are often encountered in areas such as medicine, where DNA microarray technology can produce a large number of measurements at once, and the clustering of text documents, where, if a word-frequency vector is used, the number of dimensions equals the size of the vocabulary.

Approaches

Subspace clustering
Projected clustering
Hybrid approaches
Correlation clustering

3.8 Predictive Analytics

Predictive analytics encompasses a variety of statistical techniques from predictive modelling, machine learning, and data mining that analyze current and historical facts to make predictions about future or otherwise unknown events.

In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions.

Predictive analytics is used in actuarial science, marketing, financial services, insurance, telecommunications, retail, travel, mobility, healthcare, child protection, pharmaceuticals, capacity planning and other fields.

Predictive analytics is an area of statistics that deals with extracting information from data and using it to predict trends and behavior patterns. Often the unknown event of interest is in the future, but predictive analytics can be applied to any type of unknown whether it be in the past, present or future. For example, identifying suspects after a crime has been committed, or credit card fraud as it occurs.

The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting them to predict the unknown outcome. It is important to note, however, that the accuracy and usability of results will depend greatly on the level of data analysis and the quality of assumptions.

Predictive analytics is often defined as predicting at a more detailed level of granularity, i.e., generating predictive scores (probabilities) for each individual organizational element. This distinguishes it from forecasting.

For example, "Predictive analytics—Technology that learns from experience (data) to predict

the future behavior of individuals in order to drive better decisions." In future industrial systems, the value of predictive analytics will be to predict and prevent potential issues to achieve near-zero break-down and further be integrated into prescriptive analytics for decision optimization.

Predictive Analytics Process

1. Define Project : Define the project outcomes, deliverable, scope of the effort, business objectives, identify the data sets that are going to be used.
2. Data Collection : Data mining for predictive analytics prepares data from multiple sources for analysis. This provides a complete view of customer interactions.
3. Data Analysis : Data Analysis is the process of inspecting, cleaning and modelling data with the objective of discovering useful information, arriving at conclusion
4. Statistics : Statistical Analysis enables to validate the assumptions, hypothesis and test them using standard statistical models.
5. Modelling : Predictive modelling provides the ability to automatically create accurate predictive models about future. There are also options to choose the best solution with multi-modal evaluation.
6. Deployment : Predictive model deployment provides the option to deploy the analytical results into everyday decision making process to get results, reports and output by automating the decisions based on the modelling.
7. Model Monitoring : Models are managed and monitored to review the model performance to ensure that it is providing the results expected.