UNIT 2

Data Analysis: Regression modelling, multivariate analysis, Bayesian modelling, inference and Bayesian networks, support vector and kernel methods, analysis of time series: linear systems analysis & nonlinear dynamics, rule induction, neural networks: learning and generalisation, competitive learning, principal component analysis and neural networks, fuzzy logic: extracting fuzzy models from data, fuzzy decision trees, stochastic search methods.

Regression modelling

Regression modeling is a statistical method used to predict a continuous outcome variable (also known as the dependent variable) based on one or more predictor variables (also known as independent variables). The goal of regression modeling is to find the relationship between the predictor variables and the dependent variable, and then use that relationship to make predictions about the dependent variable for new data points. This relationship is often expressed as an equation, with coefficients representing the strength and direction of the relationship between each predictor and the dependent variable.

There are several types of regression analysis, including:

1. **Linear regression:** models the relationship between the dependent variable and independent variables as a straight line.
2. **Logistic regression:** models the relationship between the independent variables and a binary dependent variable using a logistic function.
3. **Polynomial regression:** models the relationship between the dependent variable and the independent variables as an n-degree polynomial.
4. **Multiple regression:** models the relationship between the dependent variable and multiple independent variables.
5. **Non-linear regression:** models the relationship between the dependent variable and the independent variables using a non-linear function.
6. **Regularized regression:** models that incorporate a penalty term to reduce overfitting, such as ridge, lasso, and elastic net regression.

The choice of the appropriate regression technique depends on the nature of the dependent variable and the independent variables, and the desired outcome of the analysis. Regression analysis is widely used in various fields, such as economics, finance, psychology, and engineering, to make predictions and gain insights into relationships between variables.

Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable (also called the response variable) and one or more independent variables (also called predictor variables). The relationship is assumed to be linear, which means that the model is a straight line that can be described by an equation of the form:

$$y = \beta0 + \beta1x1 + \beta2x2 + ... + \betapxp + \varepsilon$$

where y is the dependent variable, x1, x2, ..., xp are the independent variables, β0, β1, β2, ..., βp are the coefficients that represent the intercept and slopes of the line, and ε is the random error that is assumed to be normally distributed and have a mean of zero.

The goal of linear regression is to find the values of the coefficients that best fit the data, by minimizing the sum of the squared differences between the actual and predicted values of the dependent variable. This method is called ordinary least squares (OLS) regression.

Linear regression is a widely used technique in many fields, including economics, finance, social sciences, and engineering. It is used to make predictions and to identify the key factors that affect the dependent variable. It can also be used to estimate the strength and direction of the relationships between variables, and to test hypotheses about those relationships.

Types of Regression:

1. **Simple**
2. **Multiple**
3. **Polynomial**
4. **Logistic**

Multivariate Analysis

Multivariate analysis is a statistical technique used to analyze and understand the relationships between multiple variables. It is a powerful tool for exploring and understanding complex data sets, and can be used for a wide range of purposes such as:

1. Exploring relationships between multiple variables and identifying patterns.
2. Predicting outcomes or relationships between variables.

3. Testing hypotheses about the relationships between variables.
4. Understanding the impact of multiple variables on a single outcome.
5. Comparing the strength and significance of multiple predictor variables.

There are several types of multivariate analysis methods, including:

1. Principal Component Analysis (PCA)
2. Factor Analysis
3. Discriminant Analysis
4. Multivariate Regression
5. Multivariate Analysis of Variance (MANOVA)
6. Canonical Correlation Analysis

The process of multivariate analysis typically involves several steps, including:

1. Data preparation: cleaning, transforming, and organizing the data into a suitable format for analysis.
2. Variable selection: choosing which variables to include in the analysis based on the research question and data structure.
3. Model building: fitting the chosen statistical model to the data.
4. Model assessment: evaluating the goodness of fit and validity of the model.
5. Interpretation: interpreting the results of the analysis and drawing meaningful conclusions about the relationships between variables.

It is important to have a strong understanding of statistics and a solid grasp of the underlying mathematical concepts before attempting multivariate analysis. Additionally, a thorough understanding of the data being analyzed and the research question being asked is essential for successful and meaningful results.

Bayesian Modeling

Bayesian modeling is a statistical approach that uses Bayes' theorem to model uncertainty in data, and involves the following steps:

1. **Specifying prior distributions:** Prior distributions are specified for the model parameters, representing our initial belief about the parameters before observing any data.
2. **Incorporating data through the likelihood function:** The likelihood function describes how well the model fits the data, and is proportional to the probability of observing the data given the model parameters.
3. **Updating prior distributions using Bayes' theorem**: Bayes' theorem is used to update the prior distributions into posterior distributions, which represent the updated belief about the parameters given the data.
4. **Making inferences and predictions:** The posterior distributions can be used to make inferences about the model parameters and predictions about future data.
5. **Modeling uncertainty:** Bayesian modeling provides a framework for modeling uncertainty and making probabilistic predictions.
6. **Software implementation:** There are several popular software packages for implementing Bayesian models, including R, BUGS, Stan, and JAGS. These packages allow for efficient computation of the posterior distributions.

Bayesian modeling can be applied to a wide range of problems, including regression, classification, time series analysis, and more. It provides a flexible and powerful framework for making inferences about uncertain quantities, and is an important tool for understanding and making predictions about complex data.

Naive Bayes Algorithm

Naive Bayes is a probabilistic machine learning algorithm based on Bayes' theorem. It operates by calculating class probabilities and feature probabilities, and using Bayes' theorem to combine these probabilities and make predictions for new data points. Here is a step-by-step technical description of the naive Bayes algorithm:

1. Calculate class probabilities: Given a training set of n data points, the class probabilities are estimated as follows:

   P(C_k) = (number of data points in class C_k) / n

where C_k represents one of the K classes in the data, and P(C_k) is the estimated probability of class C_k.

2. Calculate feature probabilities: Given a training set of n data points with d features, the probabilities of each feature for each class are calculated as follows:

   P(X_j | C_k) = (number of data points in class C_k with feature X_j) / (number of data points in class C_k)

   where X_j is one of the d features, and P(X_j | C_k) is the estimated probability of feature X_j given class C_k.

3. Combine probabilities using Bayes' theorem: Given a new data point with features X, the posterior probabilities of each class C_k are calculated as follows:

   P(C_k | X) = P(X | C_k) * P(C_k) / P(X)

   where P(C_k | X) is the estimated probability of class C_k given the features X, and P(X) is a normalizing constant equal to the sum of the probabilities of each class given X.

4. Predict the class with the highest posterior probability: The class with the highest estimated posterior probability is predicted as the class of the new data point.
5. Estimate accuracy: The accuracy of the algorithm can be estimated using a test set or by cross-validation.

Naive Bayes algorithms can be further specialized into Gaussian naive Bayes, Bernoulli naive Bayes, and Multinomial naive Bayes, depending on the distributional assumption made for the feature probabilities. These variations are well-suited for different types of data and applications, and can be selected based on the characteristics of the data and the problem being solved.

Inference and Bayesian networks

Bayesian networks are a type of probabilistic graphical model that represents the probabilistic relationships between variables. In Bayesian networks, the variables are represented as nodes in a directed acyclic graph (DAG), with directed edges representing the relationships between the variables.

Inference in Bayesian networks refers to the process of making predictions or drawing inferences about the values of variables based on observed data or evidence. This is done by using the network structure and the probabilities of the variables to calculate the probabilities of the variables given the observed data.

There are two main types of inference algorithms in Bayesian networks: exact inference and approximate inference. Exact inference algorithms calculate the exact probabilities of variables given the observed data and the network structure. These algorithms are generally fast and efficient, but can become computationally intractable for large networks. Approximate inference algorithms are used to estimate the probabilities of variables in large or intractable networks. These algorithms trade off accuracy for computational efficiency and can be used to make predictions or draw inferences about the variables.

Bayesian networks are used in a variety of applications, including predictive modeling, causal reasoning, and decision making. In predictive modeling, Bayesian networks are used to make predictions about future events based on historical data. In causal reasoning, Bayesian networks are used to make inferences about the effect of changes in one variable on another variable. In decision making, Bayesian networks are used to evaluate the uncertainty and trade-offs associated with different decisions.

In summary, Bayesian networks are a type of probabilistic graphical model that represents the probabilistic relationships between variables. Inference in Bayesian networks refers to the process of making predictions or drawing inferences about the values of variables based on observed data or evidence. Bayesian networks are used in a variety of applications, including predictive modeling, causal reasoning, and decision making.

Support Vector Machine

Support Vector Machines (SVMs) are a type of supervised machine learning algorithm used for classification and regression analysis. The main idea behind SVMs is to find a hyperplane that separates the data into different classes with the maximum margin, which is the distance between the closest data points from either class and the hyperplane.

Here are some key technical points about SVMs:

1. **Maximal Margin Classifier:** SVMs are based on the concept of a maximal margin classifier, which is a classifier that maximizes the margin between the two classes. The margin is defined as the distance between the closest data points from either class and the hyperplane.
2. **Linear and Non-Linear Classification:** SVMs can be used for both linear and non-linear classification. In linear classification, the hyperplane is a linear boundary that separates the data into different classes. In non-linear classification, the data is transformed into a high-dimensional space where it becomes linearly separable.
3. **Dual Formulation:** The optimization problem in SVMs is formulated as a convex quadratic programming problem in the dual form, which is then solved to find the support vectors and the coefficients of the hyperplane.
4. **Kernels:** SVMs use a technique called kernel trick, which allows them to find non-linear decision boundaries by transforming the data into a high-dimensional space. The kernel function is used to calculate the inner products of the transformed data points. Common kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid kernels.
5. **Overfitting:** One of the challenges in using SVMs is avoiding overfitting, which is when the model fits the training data too well and is not able to generalize to new data. To avoid overfitting, techniques such as cross-validation and regularization can be used.
6. **Applications:** SVMs have a wide range of applications, including text classification, image classification, and bioinformatics. They are also used for time-series analysis and forecasting, and are popular in the fields of computer vision and pattern recognition.

In conclusion, Support Vector Machines (SVMs) are a powerful and versatile machine learning algorithm for classification and regression analysis. They are based on the concept of a maximal margin classifier and can handle both linear and non-linear data. SVMs use the kernel trick to transform the data into a high-dimensional space and avoid overfitting. They have many applications in a variety of fields.

There are two main types of support vector machine (SVM) models:

1. **Linear SVM:** This type of SVM separates data into classes using a straight line (or hyperplane) with maximum margin. It is useful for datasets that are linearly separable.
2. **Non-Linear SVM:** This type of SVM is used for datasets that are not linearly separable. It separates data into classes using a non-linear boundary. This can be achieved using

various techniques, such as the use of polynomial, radial basis function (RBF), or sigmoid kernels.

Additionally, there are two variations of SVM based on the type of output:

1. **Binary Classification SVM:** This type of SVM is used for datasets with two classes, where the goal is to separate the classes into two distinct groups.
2. **Multi-Class Classification SVM:** This type of SVM is used for datasets with more than two classes, where the goal is to separate the classes into multiple groups.

The hard-margin and soft-margin support vector machine (SVM) models are two variations of linear SVM models.

1. **Hard-Margin SVM:** This type of SVM is used when the training data is linearly separable, meaning that there is a clear boundary between the classes. In hard-margin SVM, no training samples are allowed to be on the wrong side of the boundary or inside the margin, resulting in a "hard" boundary. However, in practice, it is often the case that the training data is not perfectly separable, leading to overfitting.
2. **Soft-Margin SVM:** This type of SVM allows for a certain number of training samples to be on the wrong side of the boundary or inside the margin, resulting in a "soft" boundary. This allows the model to handle non-separable data, which is more representative of real-world data. In soft-margin SVM, a trade-off is made between the width of the margin and the number of training samples on the wrong side of the boundary. The goal is to find a balance between the two, to minimize the total classification error.

Kernel Methods

Kernel methods play a crucial role in support vector machines (SVMs), a type of machine learning algorithm used for classification and regression. The main idea behind kernel methods in SVMs is to map the input data into a higher dimensional feature space, where it is possible to find a hyperplane that separates the classes with maximum margin.

The mapping from the original input space to the feature space is performed using a kernel function, which is a mathematical operation that calculates the dot product of two inputs in the feature space, even though the inputs are in the original input space. This is known as the "kernel trick" and enables SVMs to perform non-linear classification and regression, without having to explicitly compute the higher-dimensional feature representation.

There are several types of kernel functions that can be used in SVMs, including:

1. **Linear kernel:** The dot product of the input vectors in the original input space.
2. **Polynomial kernel:** A polynomial function of the dot product of the input vectors in the original input space.
3. **Radial basis function (RBF) kernel:** A Gaussian function that measures the similarity between two inputs based on the Euclidean distance between them.
4. **Sigmoid kernel:** A S-shaped function that maps the input features into the range of 0 to 1.

The choice of kernel function depends on the characteristics of the data and the desired complexity of the model. In general, radial basis function (RBF) and polynomial kernels are commonly used, as they are able to capture non-linear relationships between the input and output variables.

Kernel methods are a powerful tool in SVMs, as they allow for the modeling of complex relationships between the input and output variables, and provide a robust and effective method for non-linear classification and regression.

Analysis of time series: linear systems analysis & nonlinear dynamics

A time series is nothing but a sequence of various data points that occurred in a successive order for a given period.

To perform the time series analysis, we must follow the following steps:

- Collecting the data and cleaning it
- Preparing Visualization with respect to time vs key feature
- Observing the stationarity of the series
- Developing charts to understand its nature.

- Model building – AR, MA, ARMA and ARIMA
- Extracting insights from prediction

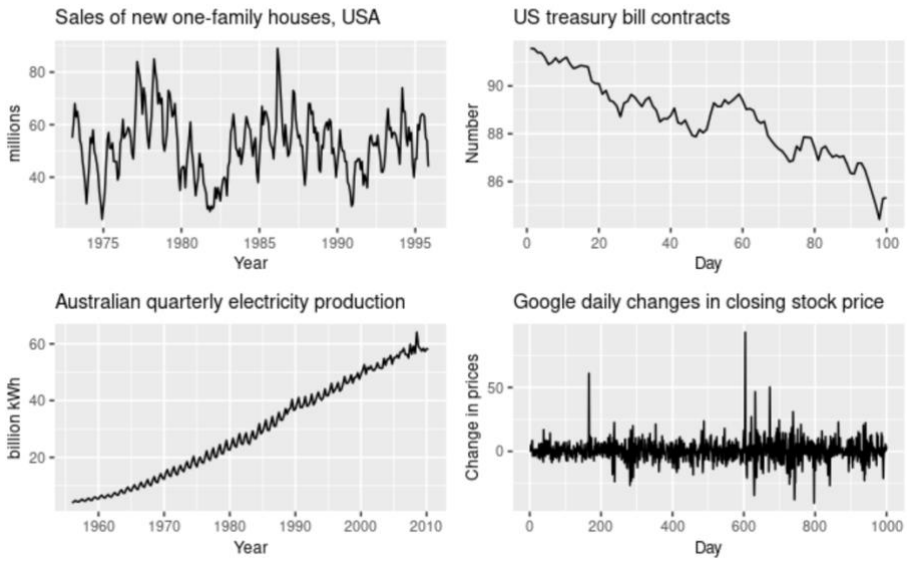
Time series analysis has following components:

Trend: In which there is no fixed interval and any divergence within the given dataset is a continuous timeline. The trend would be Negative or Positive or Null Trend

Seasonality: In which regular or fixed interval shifts within the dataset in a continuous timeline. Would be bell curve or saw tooth

Cyclical: In which there is no fixed interval, uncertainty in movement and its pattern

Irregularity: Unexpected situations/events/scenarios and spikes in a short time span.



The monthly housing sales (top left) show strong seasonality within each year, as well as some strong cyclic behaviour with a period of about 6–10 years. There is no apparent trend in the data over this period.

The US treasury bill contracts (top right) show results from the Chicago market for 100 consecutive trading days in 1981. Here there is no seasonality, but an obvious downward trend. Possibly, if we had a much longer series, we would see that this downward trend is actually part of a long cycle, but when viewed over only 100 days it appears to be a trend.

The Australian quarterly electricity production (bottom left) shows a strong increasing trend, with strong seasonality. There is no evidence of any cyclic behaviour here.

The daily change in the Google closing stock price (bottom right) has no trend, seasonality or cyclic behaviour. There are random fluctuations which do not appear to be very predictable, and no strong patterns that would help with developing a forecasting model.

Methods to Check Stationarity

During the TSA model preparation workflow, we must assess whether the dataset is stationary or not. This is done using Statistical Tests. There are two tests available to test if the dataset is stationary:
Augmented Dickey-Fuller (ADF) Test
Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

Null and Alternate Hypothesis:

In time series analysis, the null hypothesis and alternative hypothesis are statements that describe the relationship between variables being tested. The null hypothesis is a statement that assumes there is no significant relationship between variables or no difference in the values being tested. The alternative hypothesis, on the other hand, is a statement that assumes there is a significant relationship between variables or a difference in the values being tested.

In simpler words, the null hypothesis represents the status quo or the assumption that nothing interesting is happening, while the alternative hypothesis represents the opposite idea, that something interesting or significant is happening.

For example, in the context of the ADF test, the null hypothesis is that the time series has a unit root, indicating that it is non-stationary. The alternative hypothesis is that the time series does not have a unit root, indicating that it is stationary. So, if the test statistic is smaller than a critical value, we reject the null hypothesis and accept the alternative hypothesis, concluding that the time series is stationary.

In summary, the null hypothesis is a statement that there is no significant relationship between variables, while the alternative hypothesis is a statement that there is a significant relationship between variables. The choice of null and alternative hypotheses is important in time series

analysis because it determines the direction of the statistical test and the interpretation of the results.

Augmented Dickey-Fuller (ADF) Test

In the ADF (Augmented Dickey-Fuller) Test, the null hypothesis assumes that the time series has a unit root, indicating that it is non-stationary. The alternative hypothesis assumes that the time series does not have a unit root, indicating that it is stationary.

To check the null and alternative hypotheses in the ADF test, we need to perform the following steps:

- **Calculate the test statistic:** The ADF test calculates a test statistic, which is a measure of how much the time series deviates from the null hypothesis of having a unit root.
- **Compare the test statistic with critical values:** The ADF test provides critical values for the test statistic at various levels of significance (e.g., 1%, 5%, and 10%). We compare the calculated test statistic with the critical values to determine whether the null hypothesis can be rejected or not.
- **Interpret the results:** If the calculated test statistic is smaller than the critical value, we can reject the null hypothesis and conclude that the time series is stationary (i.e., the alternative hypothesis is true). Conversely, if the calculated test statistic is larger than the critical value, we cannot reject the null hypothesis and must conclude that the time series is non-stationary.

In simple words, we use the ADF test statistic and critical values to determine whether the null hypothesis (time series has a unit root) can be rejected in favor of the alternative hypothesis (time series does not have a unit root). If the test statistic is smaller than the critical value, we reject the null hypothesis and conclude that the time series is stationary. If the test statistic is larger than the critical value, we cannot reject the null hypothesis and must conclude that the time series is non-stationary.

Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test is a statistical test used to determine whether a time series is stationary or non-stationary. Unlike the Augmented Dickey-Fuller (ADF) test, which tests for the presence of a unit root (i.e., non-stationarity), the KPSS test tests for the absence of a unit root (i.e., stationarity).
The KPSS test involves computing a test statistic and a p-value based on the data. The null hypothesis of the KPSS test is that the time series is stationary, while the alternate hypothesis is that the time series is non-stationary. The test compares the test statistic to critical values based on the significance level and the number of observations in the time series.

If the test statistic is greater than the critical value, the null hypothesis can be rejected, and the time series can be considered non-stationary. On the other hand, if the test statistic is less than the critical value, the null hypothesis cannot be rejected, and the time series can be considered stationary.
The KPSS test is often used in conjunction with the ADF test to determine the stationarity of a time series. If both tests agree that the time series is stationary, then it can be used in statistical models and analyses that assume stationarity.

Overall, the KPSS test is a useful tool for identifying the presence or absence of unit roots in a time series and determining its stationarity.

Here are some reasons why we check for stationarity:

Ease of modeling: Stationary time series are easier to model than non-stationary time series. With a stationary time series, we can assume that the statistical properties of the series remain the same over time, and we can use simpler models to capture the underlying patterns. This simplifies the modeling process and makes it easier to interpret the results.

Accurate predictions: Stationary time series are more predictable than non-stationary time series. If the statistical properties of a time series change over time, then it becomes more difficult to make accurate predictions. By checking for stationarity, we can ensure that the time series can be modeled accurately and that our predictions are more reliable.

Statistical tests: Many statistical tests used in time series analysis, such as autocorrelation and unit root tests, require the time series to be stationary. These tests are used to check the

assumptions of time series models and to determine if a particular model is appropriate for the data.

Linear Systems Analysis:

1. Deals with the study of linear relationships between variables in a time series and the behavior of the system over time.
2. Based on the assumption that changes in one variable can be expressed as a linear combination of previous values of the same variable and/or other related variables.
3. Used for modeling and prediction of stationary time series data, where the statistical properties such as mean and variance remain constant over time.
4. Coefficients in the models are estimated using techniques such as least squares and maximum likelihood.

Linear Time Series Analysis Methods:

- Autoregression (AR)
- Moving Average (MA)
- Autoregressive Integrated Moving Average (ARIMA)
- Vector Autoregression (VAR)
- Vector Error Correction Model (VECM)
- Kalman Filter
- Transfer Function Models

Nonlinear Dynamics:

1. Focuses on the study of nonlinear relationships between variables and the complex behavior that arises from these relationships.
2. Used to study systems that exhibit chaotic behavior, where small changes in initial conditions can result in large changes in the outcome over time.
3. Chaos theory deals with the study of complex and unpredictable behavior in deterministic nonlinear systems.
4. Fractal analysis is used to study the self-similar patterns in time series data and to estimate the fractal dimension.

5. Neural networks are used to model the nonlinear relationships in the time series data and make predictions based on those relationships.

Nonlinear Time Series Analysis Methods:

- Chaos Theory
- Fractal Analysis
- Neural Networks
- Support Vector Machines (SVM)
- Genetic Algorithms
- Hidden Markov Models (HMM)
- Nonlinear Autoregression with exogenous inputs (NARX)
- Recurrent Neural Networks (RNN)
- Extreme Learning Machines (ELM)

In conclusion, linear systems analysis is suitable for stationary time series data, while nonlinear dynamics is useful for studying complex and nonstationary time series data. The choice between the two depends on the characteristics of the time series being analyzed and the goals of the analysis.