Data Analytics (IT-3006)

Kalinga Institute of Industrial Technology
Deemed to be University
Bhubaneswar-751024

School of Computer Engineering

3 Credit

Lecture Note – Unit 2

Course Contents

Sr #	Major and Detailed Coverage Area	Hrs
2	**Data Analysis**	12
	IIntroduction, Types of Data Analytics, Importance of Data Analytics, Data Analytics Applications, Regression Modelling Techniques: Linear Regression, Multiple Linear Regression, Non Linear Regression, Logistic Regression, Time Series Analysis, Performance analysis (RMSE, MAPE).	

Introduction

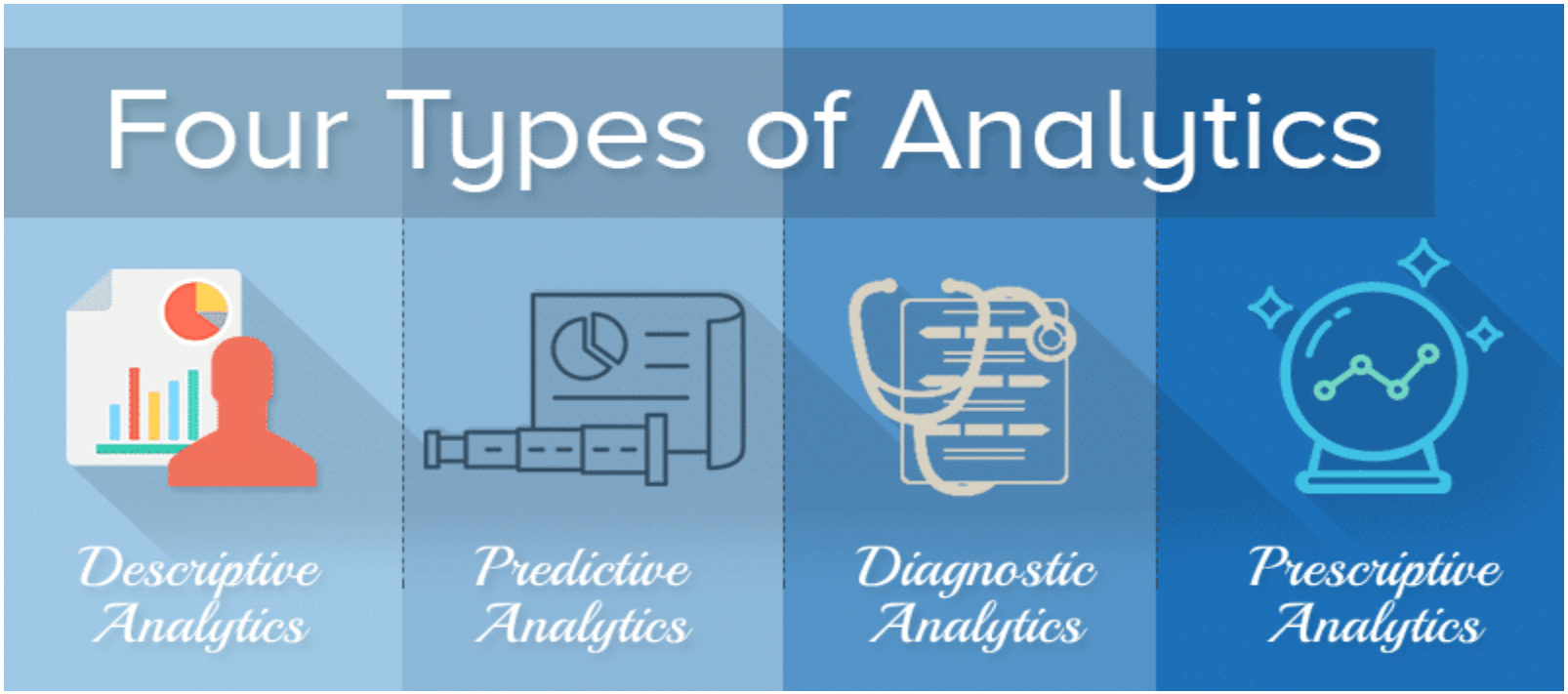
- Rapid advances in computing, data storage, networks etc have dramatically increased the ability to access, store, and process huge amount of data.
- The field of scientific research and business are challenged with the need to extract relevant information from the huge amounts of data from heterogeneous data sources such as sensors, text achieves, images, videos, audio etc.
- In such voluminous data, general patterns, structures, regularities go undetected. In many cases, such patterns can be exploited to increase the productivity of an enterprise.

Data Analysis

- Data Analysis is a process of inspecting, cleaning, transforming and modeling data with the goal of discovering useful information, suggesting conclusions and supporting decision-making.
- Intelligent data analysis (IDA) uses the concept from artificial intelligence (AI), information retrieval (IR), machine learning (ML), pattern reorganization, visualization, distributed programming and a host of other computer science concepts to automate the task of extracting unknown, valuable information /knowledge from the large amount of data.
- IDA process demands a combination of processes like extraction, analysis, conversion, classification, organization, and reasoning.
- The IDA process consists of 3 stages namely:
  - Data preparation
  - Data mining and rule finding
  - Result validation and interpretation.

Data Analytics

- Data analytics refers to the process of examining datasets to draw conclusions about the information they contain.
- Data analytic techniques enable to take raw data and uncover patterns to extract valuable insights from it.



Analytics Approach – What is the data telling?

Approach	Explanation
Descriptive	What's happening in my business? • Comprehensive, accurate and historical data • Effective Visualisation
Diagnostic	Why is it happening? • Ability to drill-down to the root-cause • Ability to isolate all confounding information
Predictive	What's likely to happen? • Decisions are automated using algorithms and technology • Historical patterns are being used to predict specific outcomes using algorithms
Prescriptive	What do I need to do? • Recommended actions and strategies based on champion/challenger strategy outcomes • Applying advanced analytical algorithm to make specific recommendations

Data Analysis vs. Data Analytics

Basis for Comparison	Data Analytics	Data Analysis
Form	Data analytics is 'general' form of analytics which is used in businesses to make decisions from data which are data-driven.	Data analysis is a specialized form of data analytics used in businesses to analyze data and take some insights of it.
Structure	Data analytics consist of data collection and inspect in general and has one or more users.	Data analysis consisted of defining a data, investigation, cleaning, transforming the data to give a meaningful outcome.
Tools	R, Tableau Public, Python, SAS, Apache Spark, Excel are used.	OpenRefine, KNIME, RapidMiner, Google Fusion Tables, Tableau Public, NodeXL, WolframAlpha are used.

Data Analysis vs. Data Analytics cont...

Basis for Comparison	Data Analytics	Data Analysis
Sequence	The life cycle consist of Business Case Evaluation, Data Identification, Data Acquisition & Filtering, Data Extraction, Data Validation & Cleansing, Data Aggregation & Representation, Data Analysis, Data Visualization, Utilization of Analysis Results.	The sequence followed are data gathering, data scrubbing, analysis of data and interpret the data precisely so that you can understand what data want to convey.
Usage	Find masked patterns, anonymous correlations, customer preferences, market trends and other necessary information that can help to make more notify decisions for business purpose.	Descriptive analysis, exploratory analysis, inferential analysis, predictive analysis and take useful insights from the data.

Data Analysis vs. Data Analytics cont...

Basis for Comparison	Data Analytics	Data Analysis
Example	Suppose, 1gb customer purchase related data of past 1 year is available, now one has to find that what the customers next possible purchases.	Suppose, 1gb customer purchase related data of past 1 year is available, now one has to find what happened so far.

Summary

- Both data analytics and data analysis are used to uncover patterns, trends, and anomalies lying within data, and thereby deliver the insights businesses need to enable evidence-based decision making.
- Where they differ, data analysis looks at the past, while data analytics tries to predict the future.
- Analysis is the detailed examination of the elements or structure of something. Analytics is the systematic computational analysis of data.

Regression (Meaning)

Basic idea:

- Use data to identify relationships among variables and use these relationships to make predictions.

Regression (Meaning)

- “To move backwards”.
- “Return to an earlier time or stage”
- Re occurrences of trends

Regression Modelling Techniques

- One of the fundamental task in data analysis is to find **how different variables are related to each other** and one of the **central tool for learning about such relationships is regression.**
- Lets take a simple example: Suppose your manager asked you to **predict annual sales.**
  - There can be **factors (drivers)** that affects sales such as competitive **pricing, product quality, shipping time & cost, online reviews, easy return policy, loyalty rewards, word of mouth recommendations, ease of checkout** etc.
  - In this case, **sales** is your **dependent variable.** Factors affecting sales are **independent variables.**
- In simple words, regression analysis is used to **model the relationship between a dependent variable and one or more independent (predictors) variables** and then use the relationships to make predictions about the future.

Regression Modelling Techniques cont…

- The regression analysis allows to model the dependent variable as a function of its predictors i.e. $\mathbf{Y = f(X_i, \beta) + e_i}$ where **Y is dependent variable, f is the function, X_i is the independent variable, β is the unknown parameters, e_i is the error term, and i varies from 1 to n.**
- *Terminologies*
  - **Multicollinearity:** When the **predictors are highly correlated to each other then the variables** are said to be multicollinear. Many types of regression techniques assumes multicollinearity should not be present in the dataset. It is because it causes problems in ranking variables based on its importance or it makes job difficult in selecting the most important independent variable (factor).
  - **Heteroscedasticity: When dependent variable's variability is not equal across values of an independent variable**, it is called **heteroscedasticity**. Example -As one's income increases, the variability of food consumption will increase. A poorer person will spend a rather constant amount by always eating inexpensive food; a wealthier person may occasionally buy inexpensive food and at other times eat expensive meals. Those with higher incomes display a greater variability of food consumption.

Terminologies cont...

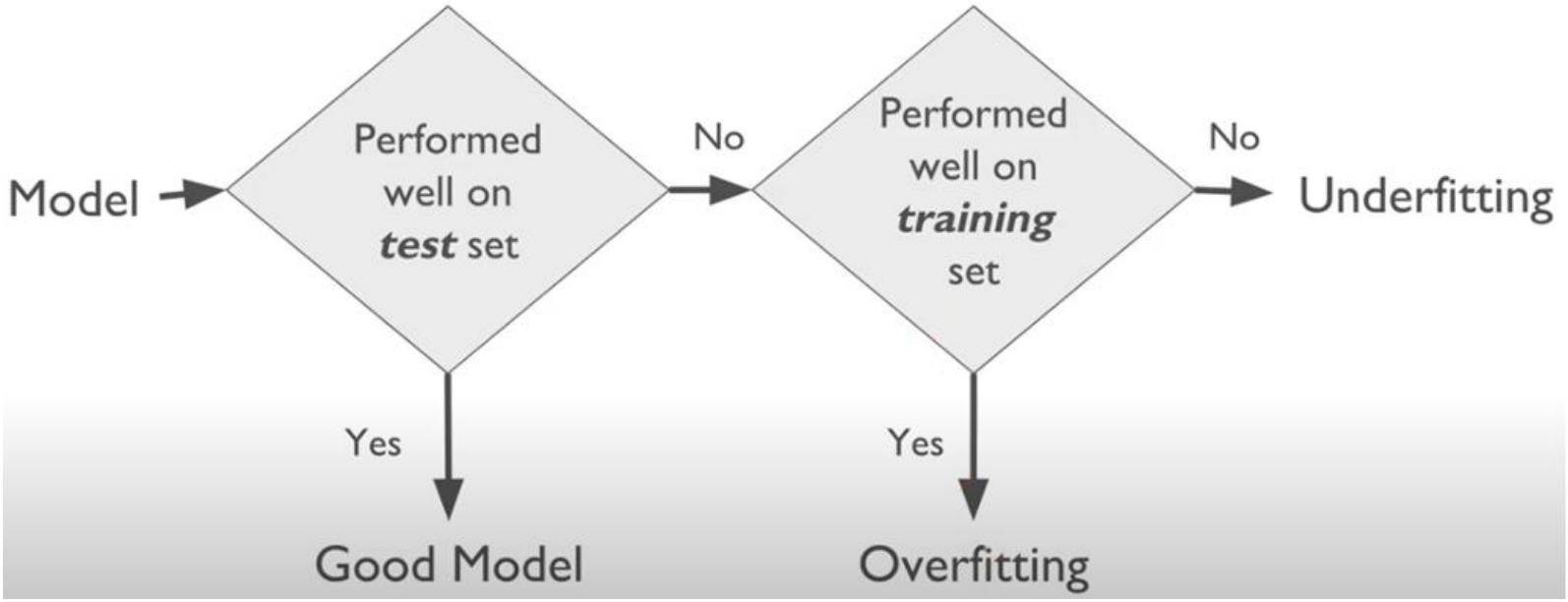
- **Training and Test dataset:** In a dataset, a **training set is implemented to build up a model,** while a **test (or validation) set is to validate the model built.** So, training data is used to fit the model and testing data to test it.

- **Overfitting:** It means that **model works well on the training dataset** but is **unable to perform better on the test datasets**. It is also known as problem of **high variance**. Variance indicates how much the estimate of the model will alter if different training data were used.

- **Underfitting**: When the **model works so poorly that it is unable to fit even training set** well then it is said to be underfitting the data. It is also known as problem of **high bias**. A bias is the amount that a model's prediction differs from the target value.

Terminologies cont...

- *Overfitting and Underfitting*



Terminologies cont...

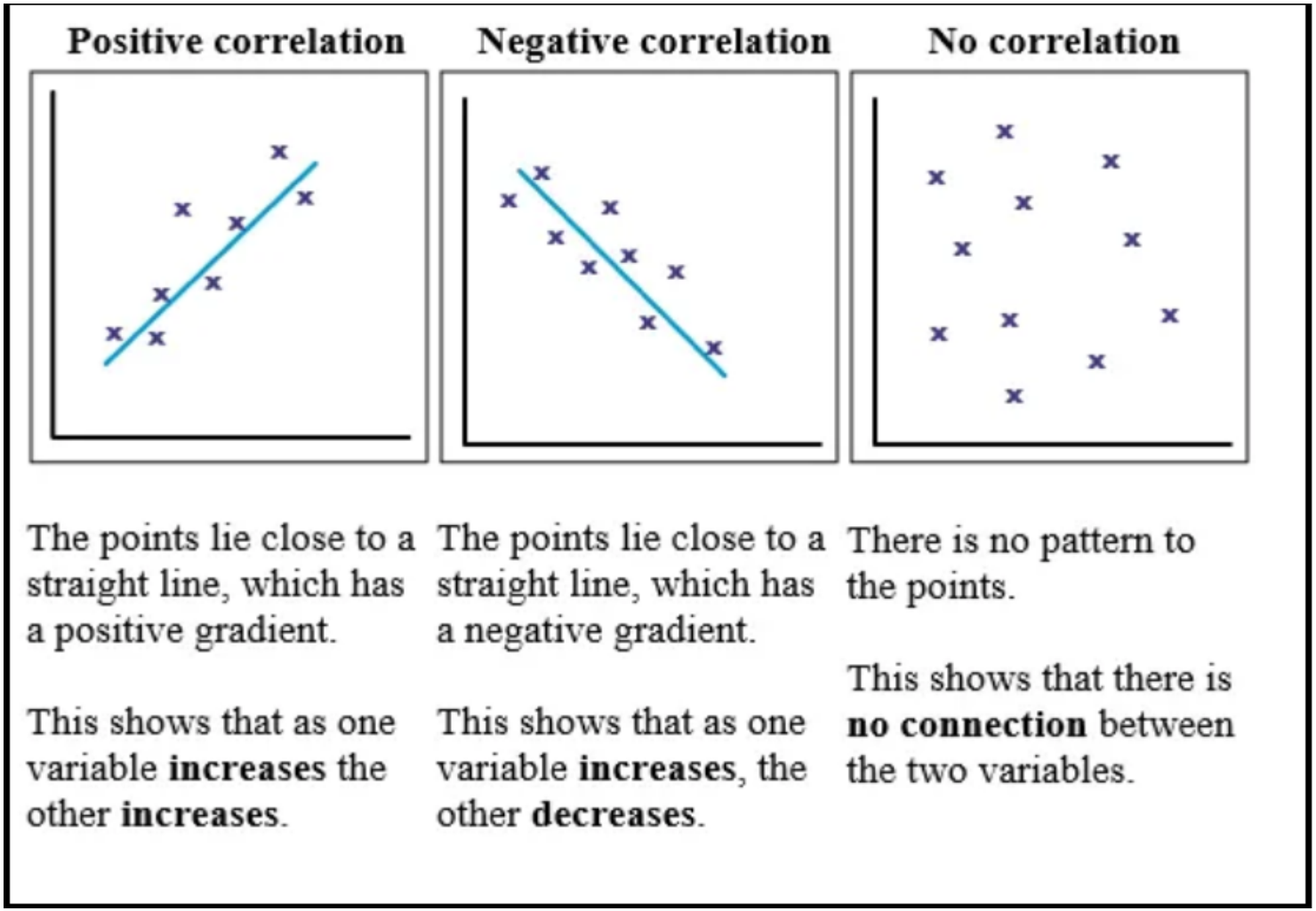
Correlation: Correlation means **association** - more precisely it is a **measure of the extent to which two variables are related.** There are three possible results of a correlational study: a positive correlation, a negative correlation, and no correlation.

- **A positive correlation** is a relationship between **two variables in which both variables move in the same direction.** Therefore, when one variable increases as the other variable increases, or one variable decreases while the other decreases. An example of positive correlation would be **height and weight**. Taller people tend to be heavier.
- **A negative correlation** is a relationship between two variables in which **an increase in one variable is associated with a decrease in the other**. An example of negative correlation would be height above sea level and temperature. As you climb the mountain (increase in height) it gets colder (decrease in temperature).
- **A zero correlation** exists when there is **no relationship between two variables.** For example there is no relationship between the amount of **tea drunk and level of intelligence.**

Terminologies cont... Scattergrams

- A correlation can be expressed visually. This is done by drawing a **scattergram** (also known as a scatterplot, scatter graph, scatter chart, or scatter diagram).
- A scattergram is a graphical display that shows the relationships or associations between two numerical variables (or co-variables), which are represented as points (or dots) for each pair of score.
- A scattergraph indicates the strength and direction of the correlation between the co-variables.
- When you draw a scattergram it doesn't matter which variable goes on the x-axis and which goes on the y-axis.
- Correlations always deals with paired scores, so the values of the 2 variables should be taken together and used to make the diagram.

Terminologies cont... Scattergrams



Terminologies cont... Correlation Coefficients: Determining Correlation Strength

- Instead of drawing a scattergram a correlation can be expressed numerically as a coefficient, ranging from -1 to +1.

- The **correlation coefficient (r)** indicates the extent to which the pairs of numbers for these two variables lie on a straight line. **Values over zero indicate a positive correlation, while values under zero indicate a negative correlation.**

- A correlation of –1 indicates a perfect negative correlation, meaning that as one variable goes up, the other goes down. A correlation of +1 indicates a perfect positive correlation, meaning that as one variable goes up, the other goes up.

Terminologies cont... Correlation Coefficients

- The following formula is normally used to estimate the **correlation coefficients** between two variables X and Y.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

- x is the independent variable and y is the dependent variable.
- n is the number of observations
- r, the computed value is known as the correlation coefficients .

Correlation Coefficients Calculation

Company	Sales in 1000s (Y)	Number of agents in 100s (X)
A	25	8
B	35	12
C	29	11
D	24	5
E	38	14
F	12	3
G	18	6
H	27	8
I	17	4
J	30	9

- $n = 10, \sum X = 80, \sum Y = 255, \sum XY = 2289$
- $\sum X^2 = 756, \sum Y^2 = 7097, (\sum X)^2 = 6400, (\sum Y)^2 = 65025, r = 0.95$

Class Exercise

Find the correlation coefficients of the below sample.

Subject	Age (x)	Glucose Level (y)
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

Types of Regression

- Every regression technique has some assumptions attached to it which need to meet before running analysis. These techniques differ in terms of type of dependent and independent variables and distribution. The types of regression algorithms are:
- Linear Regression
- Multiple Linear Regression
- Non Linear Regression
- Logistic Regression
- Polynomial Regression
- Quantile Regression
- Ridge Regression
- Lasso Regression
- Elastic Net Regression
- Principal Components Regression (PCR)
- Partial Least Squares (PLS) Regression
- Support Vector Regression
- Ordinal Regression
- Poisson Regression
- Negative Binomial Regression
- Quasi Poisson Regression
- Cox Regression

Linear Regression

- Linear regression attempts to model the **relationship between two variables by fitting a linear equation to observed data**. One variable is considered to be an **explanatory (independent) variable**, and the other is considered to be a **dependent variable**. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

- Before attempting to **fit a linear model to observed data**, **a modeler should first determine whether or not there is a relationship between the variables of interest.**

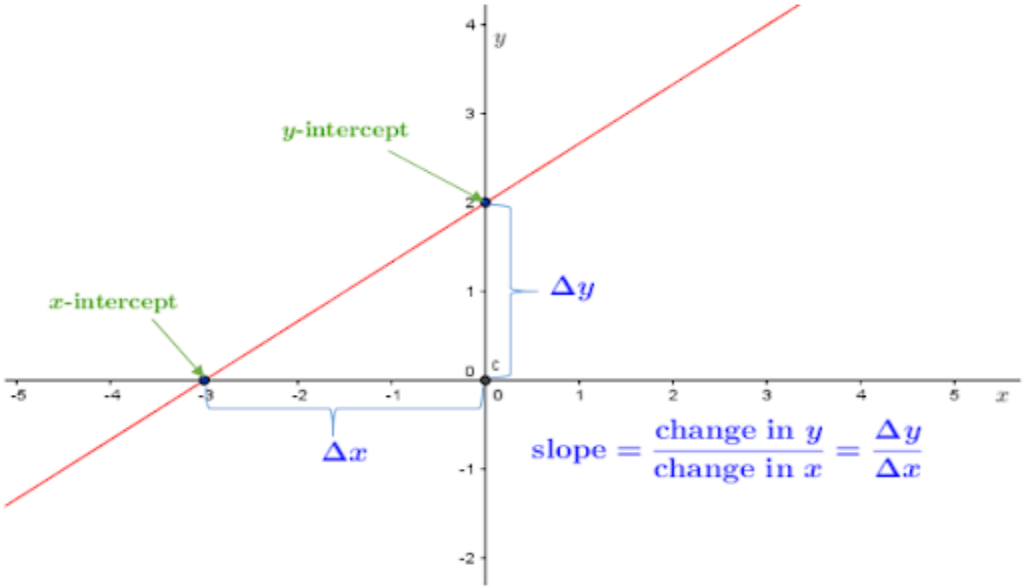
Linear Regression

- A scatterplot can be a helpful tool in determining the strength of the relationship between two variables.

- A valuable numerical measure of association between two variables is the correlation coefficient, which is a value **between -1 and 1 indicating the strength of the association** of the observed data for the two variables.

Linear Regression cont…

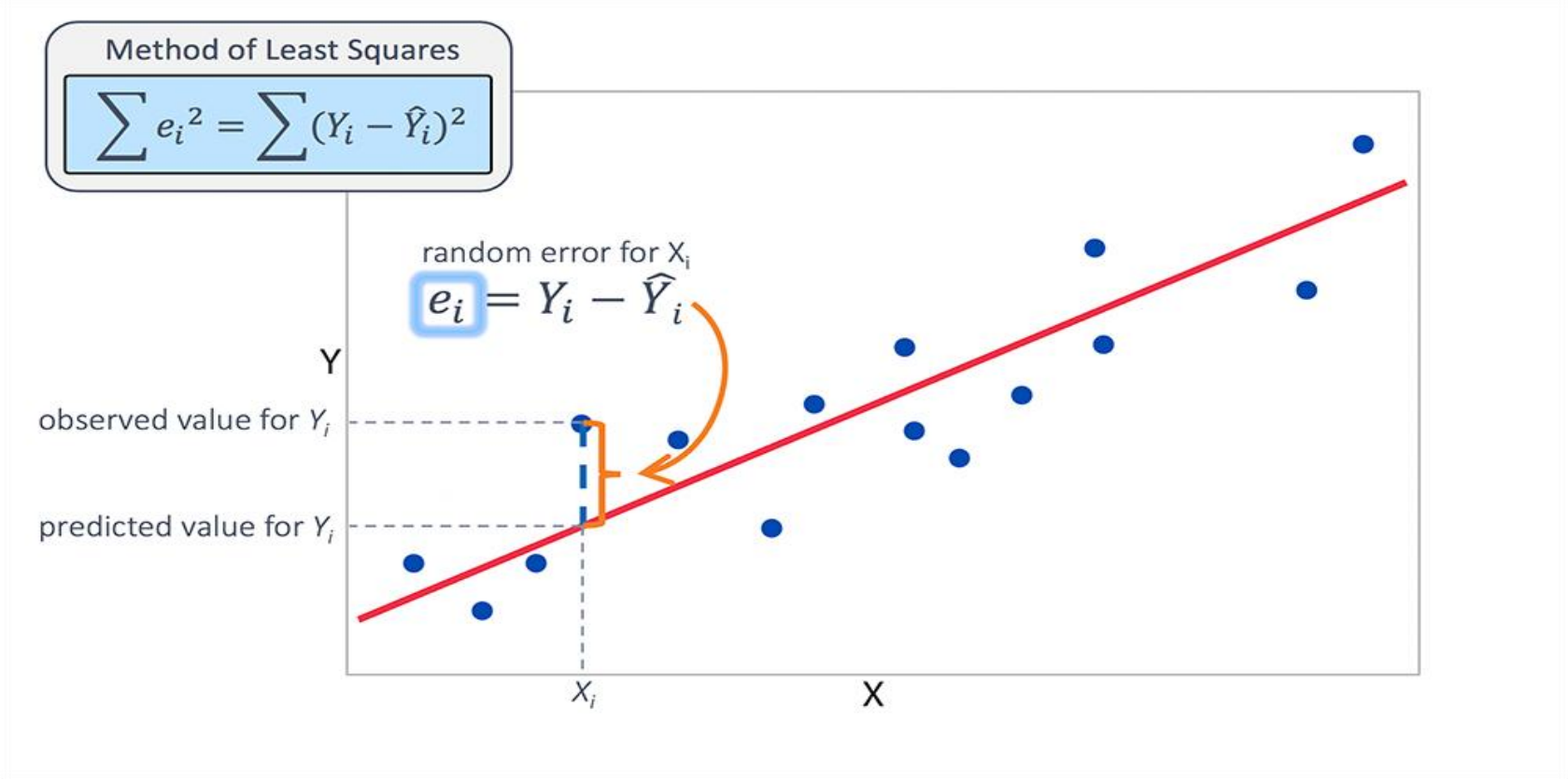
- A linear regression line has an equation of the form **Y = a + bX + e**, where **X is the explanatory variable and Y is the dependent variable.** The **slope of the line is b, a is the intercept (the value of y when x = 0), and e is the random error.**
- The slope and intercept is as follows in the following linear equation of line:



- In the above snap, a is considered as the y-intercept.

Linear Regression cont…

- The random error in the following linear equation of line:



- To fit the regression line, **a statistical approach known as least squares method.**

Linear Regression cont…

- The calculation of b and a is as follows:

$$b = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} \qquad a = \frac{\Sigma Y}{n} - b\,\frac{\Sigma X}{n}$$

- If b > 0, then x(predictor) and y(target) have a positive relationship. That is increase in x will increase y.
- If b < 0, then x(predictor) and y(target) have a negative relationship. That is increase in x will decrease y.

- If sum of squared error is taken as a metric to evaluate the model, then goal to obtain a line that best reduces the error.

$$\text{Error} = \sum_{i=1}^{n} (actual\_output - predicted\_output) ** 2$$

Linear Regression cont...

Company	Sales in 1000s (Y)	Number of agents in 100s (X)
A	25	8
B	35	12
C	29	11
D	24	5
E	38	14
F	12	3
G	18	6
H	27	8
I	17	4
J	30	9

$$b = \frac{10 \times 2289 - (80 \times 255)}{[10 \times 756 - (80)^2]} = 2.1466; \qquad a = \frac{255}{10} - 2.1466\frac{80}{10} = 8.3272$$

Linear Regression cont…

- The **linear regression** will thus be Predicted (Y) = 8.3272 + 2.1466 X

The **value of a is 8.3272,** which means that the regression line cuts the vertical axis of the graph at that point. Similarly, the **value of b** is **2.1466** indicating that the value of Y will increase by 2.1466 every time that the value of X increases by 1.

- The above equation can be used to predict the volume of sales for an insurance company given its agent number. Thus if a company has 1000 agents (10 hundreds) the predicted value of sales will be around ?

- In summary, linear regression consists of the following steps:
  - Collection of sample of independent and dependent variable.
  - Compute b and a.
  - Use these values to formulate the linear regression equation.
  - Given the new values for X predict the value of Y.

Multiple Linear Regression

- Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. It is sometimes known simply as multiple regression, and it is an extension of linear regression.
- Example:
  - Do age and intelligence quotient (IQ) scores predict grade point average (GPA)?
  - Do weight, height, and age explain the variance in cholesterol levels?
  - Do height, weight, age, and hours of exercise per week predict blood pressure?
- The formula for a multiple linear regression is:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots\ldots\ldots\ldots\ldots\ldots + \beta_n x_n + e$

where, y = the predicted value of the dependent variable.

β_0 = the y-intercept (value of y when all other parameters are set to 0)

$\beta_1 x_1$= the regression coefficient (β_1) of the first independent variable (x_1)

$\beta_n x_n$= the regression coefficient (β_n) of the last independent variable (x_n)

e = model error

Multiple Linear Regression with Two Independent Variables

- The formula for a multiple linear regression is:
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$
  where, y = the predicted value of the dependent variable.
  β_0 = the y-intercept (value of y when all other parameters are set to 0)
  $\beta_1 x_1$= the regression coefficient (β_1) of the first independent variable (x_1)
  $\beta_2 x_2$= the regression coefficient (β_n) of the second independent variable (x_2)
  e = model error
- β_1 and β_2 is calculated as follows:

$$\beta_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\beta_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

- β_0 is calculated as follows:

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2$$

where $\bar{y} = \frac{\Sigma Y}{n}$ and $\bar{x}_i = \frac{\Sigma X_i}{n}$

Non-Linear Regression

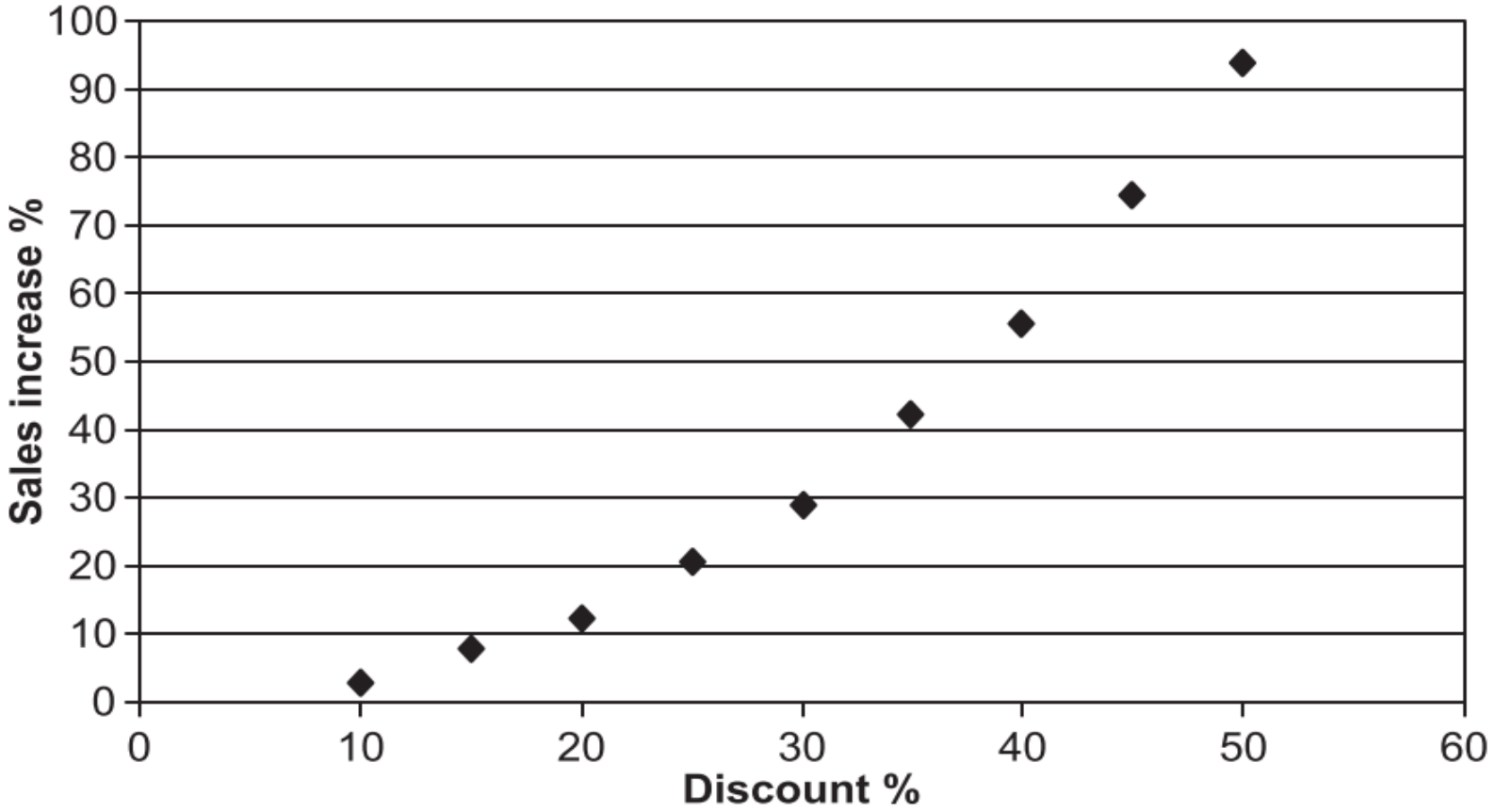
- In the case of linear and multiple linear regression, the dependent variable is linearly dependent on the independent variable(s). But, in several situations, the situation is no simple where the two variables might be related in a non-linear way.
- This may be the case where the results from the correlation analysis show no linear relationship but these variables might still be closely related.
- If the result of the data analysis show that there is a non-linear (also known as curvilinear) association between the two variables, then the need is to develop a non-linear regression model.
- The non-linear data can be handled in 2 ways:
  - Use of polynomial rather than linear regression model
  - Transform the data and then use linear regression model.

Non-Linear Regression cont…

Product	Increase in sale in% (Y)	Discount in %(X)
A	3.05	10
B	7.62	15
C	12.19	20
D	20.42	25
E	28.65	30
F	42.06	35
G	55.47	40
H	74.68	45
I	93.88	50

Non-Linear Regression cont…

The scatter diagram of sales increase for various discount percentage looks as follows:



The value of r is 0.97 which indicates a very strong, almost perfect, positive correlation, and the data value appears to form a slight curve.

Non-Linear Regression cont…

Polynomials are the equations that involve powers of the independent variables. A second degree (quadratic), third degree (cubic), and n degree polynomial functions:

- Second degree: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$
- Third degree: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + e$
- n degree: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots \ldots + \beta_n x^n + e$

Where:

- β_0 is the intercept of the regression model
- $\beta_1, \beta_2, \beta_3$ are the coefficient of the predictors.

How to find the right degree of the equation?

As we increase the degree in the model, it tends to increase the performance of the model. However, increasing the degrees of the model also increases the risk of over-fitting and under-fitting the data. So, one of the approach can be adopted:

- **Forward Selection:** This method increases the degree until it is significant enough to define the best possible model.
- **Backward Elimination:** This method decreases the degree until it is significant enough to define the best possible model.

Non-Linear Regression cont…

- The techniques of fitting of the polynomial model in one variable can be extended to the fitting of polynomial models in two or more independent variables.
- A second-order polynomial is more used in practice, and its model with two independent variables is specified by: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + e$
- This is also termed as **response surface**. The methodology of response surface is used to fit such models and helps in designing an experiment. This type is generally covered in the topics in the design of experiment.

Class work

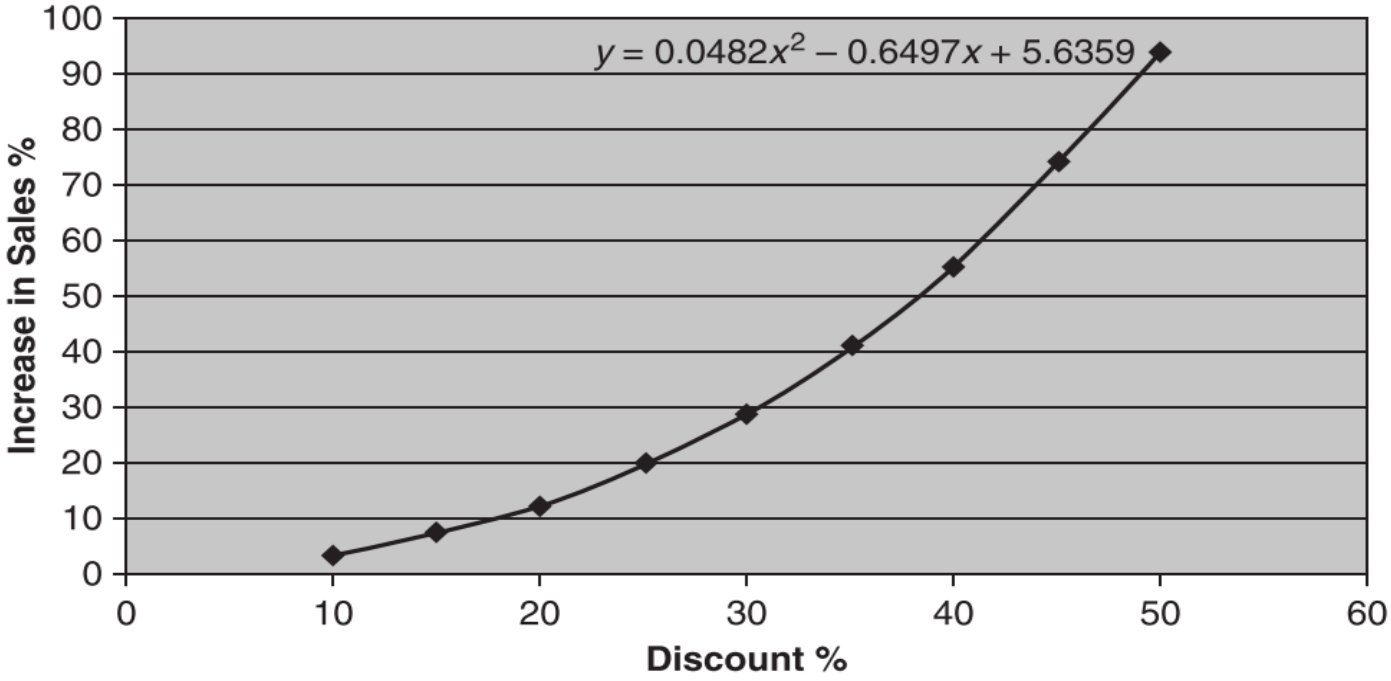
- Define the second-order polynomial model with two independent variables.
- Define the second-order polynomial model with three independent variables.
- Define the third-order polynomial model with two independent variables.

Non-Linear Regression cont…

A polynomial regression is regression that involves multiple powers of predictor(s). So, regression tools and diagnostics can be applied to polynomial regression.

Non-Linear Regression cont…

- The tools exists in software such as SAS, Excel or the language such as Python, R can estimate the value of coefficients of predictor such as β_0, $\beta 1$ etc and to fit a curve in a non-linear fashion for the given data.
- Following figure depicts the graph of increase in sale vs. discount.

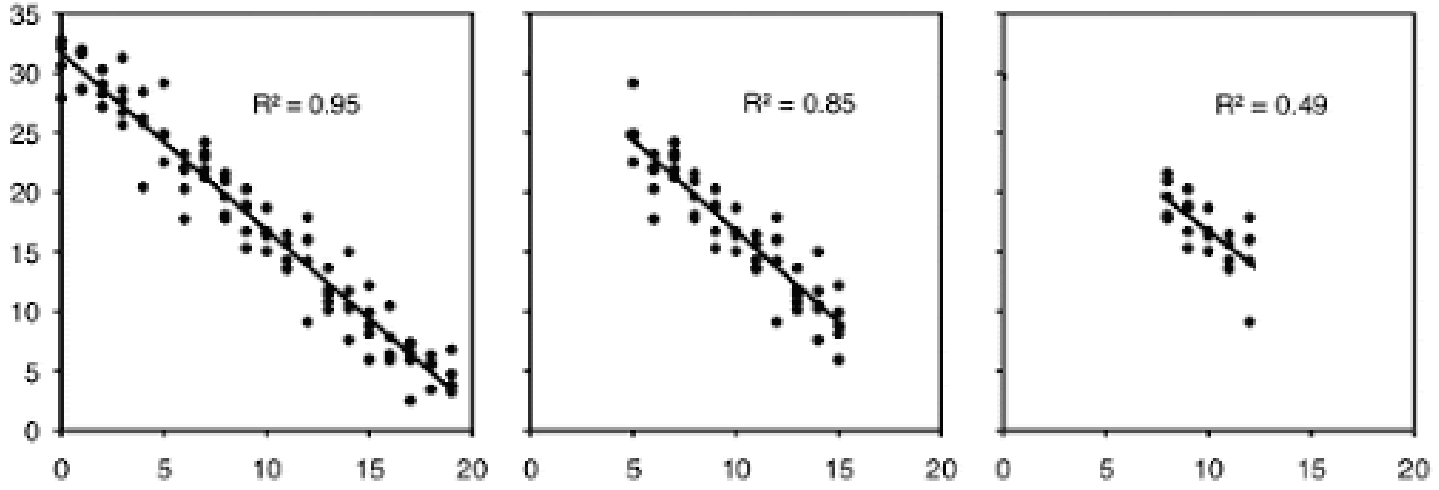


The predicted model is $Y = 5.6359 - 0.6497\ x + 0.0482\ x^2$

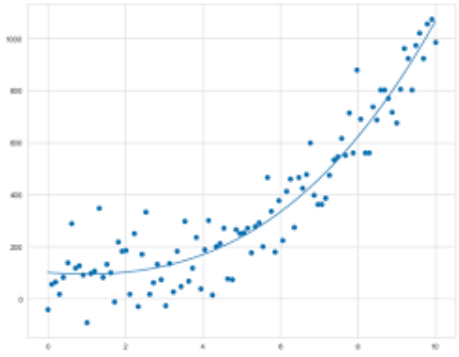
Non-Linear Regression cont…

R^2 is a known as coefficient of determination and it's a number that indicates how well the data fits into the develop model i.e. a line or curve.

Linear



Curve



Non-Linear Regression cont…

- An R^2 of 1 indicates that the regression model perfectly fits the data while an R^2 of 0 indicate that model does not fit the data at all.
- An R^2 is calculated as follows:

$$\text{Sum of Squares Regression (SSR)} = \Sigma(\hat{Y}_i - \bar{Y}_i)^2$$

$$\text{Sum of Squares Error (SSE)} = \Sigma(\hat{Y}_i - Y_i)^2$$

$$\text{Sum of Squares Total (SST)} = \Sigma(Y_i - \bar{Y})^2$$

$$\text{SST} = \text{SSR} + \text{SSE}$$

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}}$$

where

$\bar{Y}$ is the mean of the actual values of Y

$\hat{Y}_i$ is predicted values of Y_i.

Non-Linear Regression cont…

- In the example, a value of 0.99 for R^2 indicates that a quadratic model is good fit for the data.
- Another preferable way to perform non-linear regression is to try to transform the data in order to make the relationship between the two variables more linear and then use a regression model rather than a polynomial one. Transformations aim to make a non-linear relationship between two variables more linear so that it can be described by a linear regression model.
- Three most popular transformations are the:
  - Square root ($\sqrt{X}$)
  - Logarithm (log X)
  - Negative reciprocal (- 1/ X)

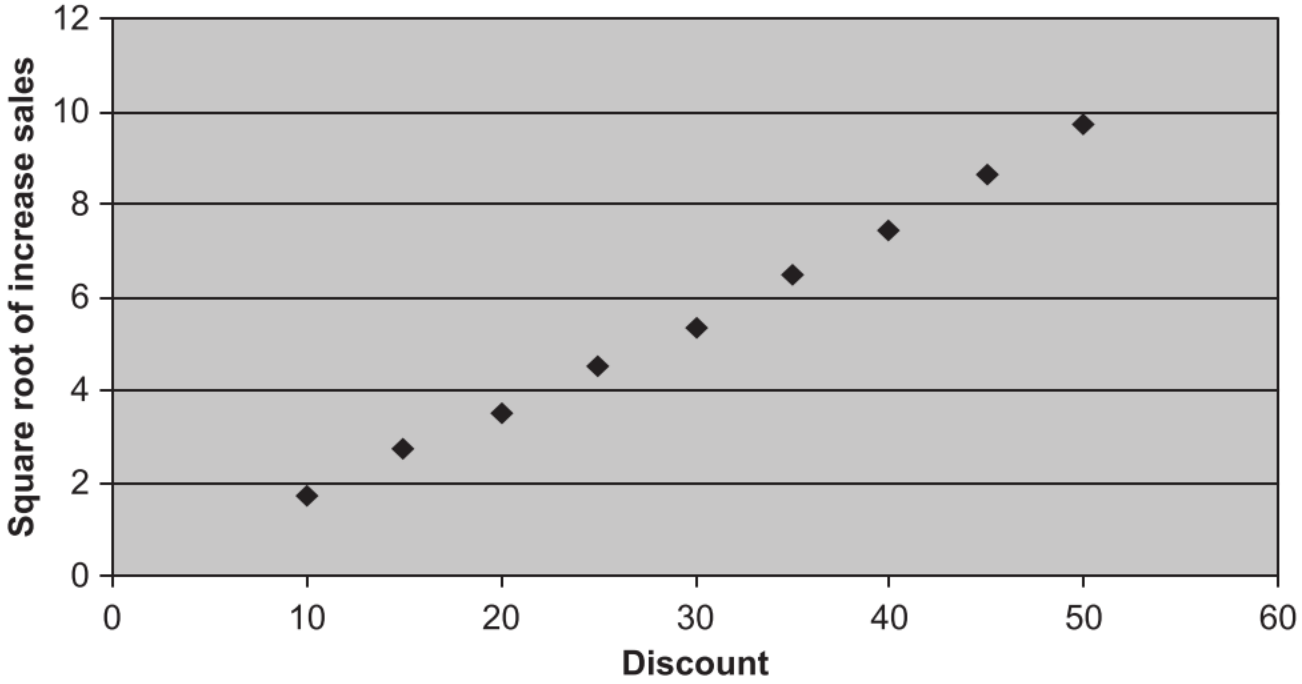
Non-Linear Regression cont…

Application of Square root (√Y)

Product	Discount in %(X)	Increase in sale in% (Y)	SQRT (Y)
A	10	3.05	√3.05 = 1.75
B	15	7.62	√7.62 = 2.76
C	20	12.19	√12.19 = 3.49
D	25	20.42	√20.42 = 4.52
E	30	28.65	√28.65 = 5.35
F	35	42.06	√42.06 = 6.49
G	40	55.47	√55.47 = 7.45
H	45	74.68	√74.68 = 8.64
I	50	93.88	√93.88 = 9.69

Non-Linear Regression cont…

Square root of transformation



In the similar fashion, Logarithm and negative reciprocal techniques can be applied to the dependent variable followed up by the application of linear regression model.

Logistic Regression

- Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications.
- Logistic Regression is used when the dependent variable (target) is categorical. For example:
  - To predict whether an email is spam (1) or not (0). If the model infers a value of 0.932 on a particular email message, it implies a 93.2% probability that the email message is spam. The model predicts the email message is spam 93.2% of the time and the remaining 6.8% will not.
  - Whether the tumor is malignant (1) or not (0)
- There are 3 types of Logistic Regression
  - **Binary Logistic Regression:** The categorical response has only two 2 possible outcomes. Example: Spam or Not.
  - **Multinomial Logistic Regression:** Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan)
  - **Ordinal Logistic Regression:** Three or more categories with ordering. Example: Movie rating from 1 to 5.

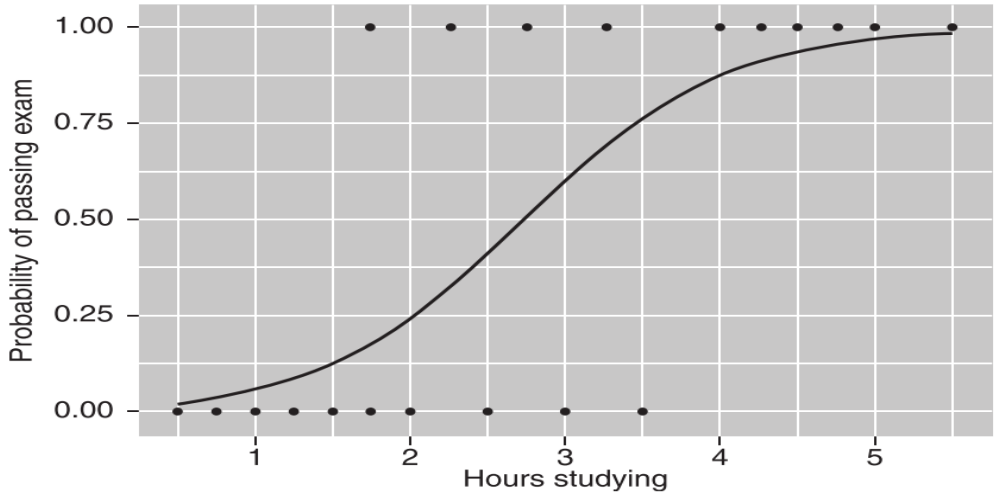
Logistic Regression cont…

- It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function which is the cumulative logistic distribution.
- Since the predicted values are probabilities and therefore are restricted to (0, 1), a logistic regression model **only predicts the probability of particular outcome given the values of the existing data**.
- **Example:** A group of 20 students spends between 0 and 6 hours studying for an exam. How does the number of hours spent studying affect the probability of the student passing the exam? The reason for using logistic regression for this problem is that the values of the dependent variable, pass and fail, while represented by "1" and "0", are not cardinal numbers. If the problem was changed so that pass/fail was replaced with the grade 0–100 (cardinal numbers), then simple regression analysis could be used. The table shows the number of hours each student spent studying, and whether they passed (1) or failed (0).

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

Logistic Regression cont…

- In logistic regression, we don't directly fit a straight line to the data like in linear regression. Instead, we fit a S shaped curve, called **sigmoid** or **logistic regression curve**. A logistic regression curve showing probability of passing an exam versus hours studying is shown below.
- Y-axis goes from 0 to 1. This is because the sigmoid function always takes as maximum (i.e. 1) and minimum (i.e. 0), and this fits very well to the goal of classifying samples in two different categories (fail or pass).
- The sigmoid function is sigmoid(x) = 1 / (1 + e^{-x}) where x is the weighted sum of independent variable i.e. x = $\beta_0 + \beta_1 x_i$ where i is the individual independent variable instance.



Logistic Regression cont...

Consider a model with one predictor X_1, and one binary response variable Y, which we denote p = P(Y = 1 | X_1 = x), where p is the probability of success. p should meet criteria: (i) it must always be positive, (ii) it must always be less than equals to 1.

We assume a linear relationship between the independent variable and the logit of the event i.e. Y = 1. In statistics, the logit is the logarithm of the **odds** i.e. p / (1-p). This linear relationship can be written in the following mathematical form (where ℓ is the logit, b is the base of the logarithm, and β is the parameter of the model.

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1$$

The odds can be recovered by exponentiation of the logit:

$$\frac{p}{1-p} = b^{\beta_0+\beta_1 x_1} \rightarrow p = \frac{b^{\beta_0+\beta_1 x_1}}{b^{\beta_0+\beta_1 x_1}+1} = \frac{1}{1+b^{-(\beta_0+\beta_1 x_1)}} = S_b(\beta_0+\beta_1 x_1)$$

Where S_b is the sigmoid function with base b. However in some cases it can be easier to communicate results by working in base 2, base 10, or exponential constant e.
In reference to the students example, solving the equation with software tool and considering base as e, the coefficient is β_0 = -4.0777 and β_1= 1.5046

Logistic Regression cont...

- For example, for a student who studies 2 hours, entering the value Hours = 2 in the equation gives the estimated probability of passing the exam of 0.26.

$$\text{Probability of passing exam} = \frac{1}{1 + \exp(-(1.5046 \cdot 2 - 4.0777))} = 0.26$$

- Similarly, for a student who studies 4 hours, the estimated probability of passing the exam is 0.87:

$$\text{Probability of passing exam} = \frac{1}{1 + \exp(-(1.5046 \cdot 4 - 4.0777))} = 0.87$$

- Following table shows the probability of passing the exam for several values of hours studying.

Hours of study	Probability of passing the exam
1	0.07
2	0.26
3	0.61
5	0.97

Bayesian Modelling

- Bayesian data analysis deals with set of practical methods for making inferences from the available data. The methods used probability models to model the given data and also predict future values.
- Thus, essentially it is a statistical paradigm that answers research questions about unknown parameters using probability statements. For example, what is the probability that the average male height is between 70 and 80 inches or that the average female height is between 60 and 70 inches?
- Bayesian data analysis consists of 3 important steps:
  - **Setting up the prior distribution:** Using domain expertise or prior knowledge to develop a joint probability distribution for all independent variable of the data under consideration and also the dependent variable. This terms as prior distribution.
  - **Setting up the posterior distribution:** After taking into the account the observed data, calculate and interpret the appropriate posterior distribution. This is estimating the conditional probability distribution of the data parameters, given the observed data.
  - **Evaluating the fit of the model:** This is to seek answer for the questions: how well does the developed model fit the data? Are the conclusion reasonable? How sensitive are the results to the modelling assumptions (as per step 1)? the In response, the model can be altered or expanded with the three steps.

Joint Probability

- A statistical measure that calculates the likelihood of two events occurring together and at the same point in time is called Joint probability.
- Let A and B be the two events, joint probability is the probability of event B occurring at the same time that event A occurs.
- The formula P (A∩B) represents the joint probability of events with intersection, where, A and B are the two events. The symbol "∩" in a joint probability is called an intersection. The probability of event A and event B happening is the same thing as the point where A and B intersect. Hence, the joint probability is also called the intersection of two or more events.
- **Example:** Find the probability that the number three will occur twice when two dice are rolled at the same time.

  Solution: Number of possible outcomes when a dice is rolled = 6 i.e. {1, 2, 3, 4, 5, 6}. Let A be the event of occurring 3 on first dice and B be the event of occurring 3 on the second dice. Both the dice have six possible outcomes, the probability of a three occurring on each die is 1/6. P(A) =1/6, P(B)=1/6 and P(A ∩B) = 1/6 x 1/6 = 1/36.

Conditional Probability

- Conditional probability is the probability of one thing being true given that another thing is true. This is distinct from joint probability, which is the probability that both things are true without knowing that one of them must be true.
- For example, one joint probability is "the probability that your left and right socks are both black," whereas a conditional probability is "the probability that your left sock is black if you know that your right sock is black,“
- Event A is that it is raining outside, and it has a 0.3 (30%) chance of raining today. Event B is that you will need to go outside, and that has a probability of 0.5 (50%). A conditional probability would look at these two events in relationship with one another, such as the probability that it is both raining and you will need to go outside. The formula for conditional probability is: $P(B|A) = P(A \cap B) / P(A)$
- **Example**: In a group of 100 sports car buyers, 40 bought alarm systems, 30 purchased bucket seats, and 20 purchased an alarm system and bucket seats. If a car buyer chosen at random bought an alarm system, what is the probability they also bought bucket seats?

Bayesian Interface

- Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available.
- It is the process of fitting a probability model to a set of data which results in a probability distribution on the parameters of the model.
- The model is then extended to new unobserved data and thus makes predictions for new observations.
- Axioms of probability:
  - **Sum Rule:** The sum rule is P(A + B) = P(A) + P(B) where A and B are each events that could occur, but cannot occur at the same time. Example: The probability that the next person walking into class will be a student and the probability that the next person will be a teacher. If the probability of the person being a student is 0.8 and the probability of the person being a teacher is 0.1, then the probability of the person being either a teacher or student is 0.8 + 0.1 = 0.9.
  - **Product Rule:** The product rule is P(E*F) = P(E) * P(F) where E and F are events that are independent. Example: When picking cards from a deck of 52 cards, the probability of getting an ace is 4/52 = 1/13, because there are 4 aces among the 52 cards. The probability of picking a heart is 13/52 = ¼, because there are 13 hearts among the 52 cards. . The probability of picking the ace of hearts is 1/4*1/13 =1/52.
  - **Not Rule**: The not rule is $P(\bar{A}) = 1 - P(A)$ where A is an event.

Bayesian Modelling cont…

Bayesian methods are based on three important theories in probability:

- Bayes Theorem
- Law of total probability
- Normalization

Bayes Theorem

Mathematically Bayes' theorem is defined as:

$P(A \mid B) = (P(B \mid A) * P(A)) / P(B)$

where A and B are events, P(A|B) is the conditional probability that event A occurs given that event B has already occurred. P(B|A) has the same meaning but with the roles of A and B reversed, and P(A) and P(B) are the marginal probabilities of event A and event B occurring respectively.

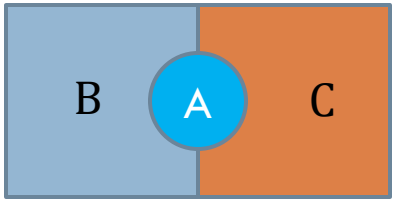
Example: There are 52 cards in the pack, 26 of them are red and 26 are black. What is the probability of the card being a 4 given that we know the card is red? Event A is the event that the card picked is a 4 and event B is the card being red. Hence, P(A|B) in the equation above is P(4|red).

P(A) = P(4) = 4/52 = 1/13, P(B) = P(red) = 26/52 = 1/2, P(B|A) = P(red|4) = 1/2

P(4|red) = (1/2 * 1/13) / (1/2) = 1/13

Law of total probability

The rule states that if the probability of an event is unknown, it can be calculated using the known probabilities of several distinct events. Consider the image:



There are three events: A, B, and C. Events B and C are distinct from each other while event A intersects with both events. We do not know the probability of event A. However, we know the probability of event A under condition B and the probability of event A under condition C. The total probability rule states that by using the two conditional probabilities, we can find the probability of event A. Mathematically, the total probability rule can be written in the following equation where n is the number of events and B_n is the distinct event.

$$P(A) = \sum_{n} P(A \cap B_n) \quad \text{where, } P(A \cap B) = P(A|B) \times P(B)$$

Law of total probability cont...

As per the diagram, the total probability of event A from the situation can be found using the equation is : $P(A) = P(A \cap B) + P(A \cap C)$.

Example: You are a stock analyst following ABC Corp. You discovered that the company is planning to launch a new project that is likely to affect the company's stock price. You have identified the following probabilities:

- There is a 60% probability of launching a new project.
- If a company launches the project, there is a 75% probability that its stock price will increase.
- If a company does not launch the project, there is a 30% probability that its stock price will increase.

You want to find the probability that the company's stock price will increase.

Solution:

P(Launch a project | Stock price increases) = 0.6 × 0.75 = 0.45

P(Do not launch | Stock price increases) = 0.4 × 0.3 = 0.12

P(Stock price increases) = P(Launch a project | Stock price increases) + P(Do not launch | Stock price increases) = 0.45 + 0.12 = 0.57. Thus, there is a 57% probability that the company's share price will increase.

Law of total probability cont…

Class Exercise 1: A person has undertaken a mining job. The probabilities of completion of job on time with and without rain are 0.42 and 0.90 respectively. If the probability that it will rain is 0.45, then determine the probability that the mining job will be completed on time.

Solution: ?

Further study: The Total Probability Rule and Decision Tree @ https://corporatefinanceinstitute.com/resources/knowledge/other/total-probability-rule/

Class Exercise 2: An airport screens bags for forbidden items, and an alarm is supposed to be triggered when a forbidden item is detected. Suppose that 5 percent of bags contain forbidden items. If a bag contains a forbidden item, there is a 98 percent chance that it triggers the alarm. If a bag doesn't contain a forbidden item, there is an 8 percent chance that it triggers the alarm. Given a randomly chosen bag triggers the alarm, what is the probability that it contains a forbidden item? Draw the decision tree.

Solution: ?

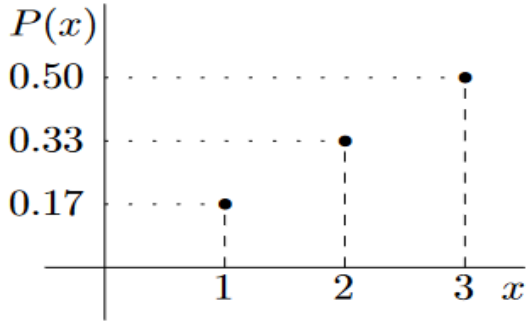
Normalization

A probability distribution function is said to be "normalized" if the sum of all its possible results is equal to **one.**

Example: Let's assume that we have a dice with 6 sides. One side is marked 1, two sides are marked 2, and three sides are marked 3. Since there are six total sides, this means that the probability of rolling each number is as shown below:

Number Rolled	Probability
1	$1/6 \approx 0.17$
2	$2/6 = 1/3 \approx 0.33$
3	$3/6 = 1/2 = 0.50$

It can also be represented as a graph of values versus probabilities



Class Exercise

If you roll this dice 25 times, about how many times will you expect to get each value (1, 2, and 3)?

Bayesian Interface cont...

Bayesian interface computes the posterior probability according to Bayes' theorem:

$$P(H \mid E) = \frac{P(H) * P(E \mid H)}{P(E)}$$

where:

- H is the hypothesis whose probability is affected by data.
- E is the evidence i.e. the unseen data which was not used in computing the prior probability
- P(H) is the prior probability i.e. it is the probability of H before E is observed
- P(H | E) is the posterior probability i.e. the probability of H given E and after E is observed.
- P(E | H) is the probability of observing E given H. It indicates the compatibility of the evidence with the given hypothesis.
- P(E) is the marginal likelihood or model evidence.

Bayesian Interface cont...

Bayes' theorem also can be written as: $P(H \mid E) = (P(H) * P (E \mid H)) * \lambda$, where $\lambda = 1/ P(E)$ and is the normalizing constant ensuring that $P (H \mid E)$ sums to 1 for each state of E.

Class Exercise

Consider the use of online dating sites by age group:

	18-29	30-49	50-64	65+	Total
Used online dating site	60	86	58	21	225
Did not use online dating site	255	426	450	382	1513
Total	315	512	508	403	1738

Based on above table,

1. What is the probability that an 18-29 year old uses online dating sites?
2. What is the probability that 65+ year old do not uses online dating sites?
3. What is the probability that an 18-29 year and 30-49 year old uses online dating sites?
4. What is the probability that an 18-29 year and 30-49 and 50-64 year old uses online dating sites?

Bayesian Model – Naïve Bayes Classifier

- **Naive Bayes classifiers (NBC)** are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. *every pair of features being classified is independent of each other.*
- Consider the problem of playing golf, and the dataset is shown on right.
- The need is to classify whether the day is suitable for playing golf, given the features of the day. The columns represent these features and the rows represent individual entries.

	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY GOLF
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

Naïve Bayes Classifier cont…

- If we take the first row of the dataset, we can observe that is not suitable for playing golf if the outlook is rainy, temperature is hot, humidity is high and it is not windy.
- We make two assumptions. The first, these predictors are independent i.e. if the temperature is hot, it does not necessarily mean that the humidity is high. Another assumption is that all the predictors have an equal effect on the outcome i.e. the day being windy does not have more importance in deciding to play golf or not.
- The Bayes theorem can be written as P(y|X) = (P(X|y) * P(y)) / P(X) where the variable y is the dependent variable(play golf), which represents if it is suitable to play golf or not given the conditions. X represent the independent variables (outlook, temperature, humidity and windy).
- X is given as X = (x_1, x_2, …., x_n) where represent the feature and mapped to outlook, temperature, humidity and windy.
- By substituting for X and expanding using the chain rule we get:

$$P(y|x_1,\ldots,x_n) = \frac{P(x_1|y)P(x_2|y)\ldots P(x_n|y)P(y)}{P(x_1)P(x_2)\ldots P(x_n)}$$

- Now, the values for each can be obtained by looking at the dataset and substitute them into the equation. For all entries in the dataset, the denominator does not change, it remain static. Therefore, the denominator can be removed and a proportionality can be introduced.

Naïve Bayes Classifier cont…

$$P(y|x_1,\ldots,x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

- In the example, the class variable(y) has only two outcomes, yes or no. There could be cases where the classification could be multivariate. Therefore, the need is to find the class y with maximum probability.

$$y = \text{argmax}_y P(y)\prod_{i=1}^{n} P(x_i|y)$$

- Using the above function, we can obtain the class, given the predictors.
- P(Y) = 9/ 14 and P(N) = 5/14 where Y stands for Yes and N stands for No.
- The outlook probability is: P(sunny | Y) = 2/9, P(overcast | Y) = 4/9, P(rain | Y) = 3/9, P(sunny | N) = 3/5, P(overcast | N) = 0, P(rain | N) = 2/5
- The temperature probability is: P(hot | Y) = 2/9, P(mild | Y) = 4/9, P(cool | Y) = 3/9, P(hot | N) = 2/5, P(mild | N) = 2/5, P(cool | N) = 1/5

Naïve Bayes Classifier cont…

- The humidity probability is: P(high | Y) = 3/9, P(normal | Y) = 6/9, P(high | N) = 4/5, P(normal | N) = 2/5.
- The windy probability is: P(true | Y) = 3/9, P(false | Y) = 6/9, P(true | N) = 3/5, P(false | N) = 2/5
- Now we want to predict "Enjoy Sport" on a day with the conditions: <outlook = sunny; temperature = cool; humidity = high; windy = strong>
- P(Y) P(sunny | Y) P(cool | Y) P(high | Y) P(strong | Y) = .005 and P(N) P(sunny | N) P(cool | N) P(high | N) P(strong | N) = .021
- Since, the probability of No is the larger, we can predict "Enjoy Sport" to be No on that day.

Types of Naive Bayes Classifier

- **Multinomial Naive Bayes:** This is mostly used for document classification problem, i.e. whether a document belongs to the category of sports, politics, technology etc. The features/predictors used by the classifier are the frequency of the words present in the document.
- **Bernoulli Naive Bayes:** This is similar to the multinomial naive bayes but the predictors are boolean variables. The parameters that we use to predict the class variable take up only values yes or no, for example if a word occurs in the text or not.
- **Gaussian Naive Bayes:** The predictors take up a continuous value and are not discrete.

Naïve Bayes Classifier cont…

Pros

- It is easy and fast to predict class of test data set. It also perform well in multi class prediction.
- When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.
- It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

Cons

- The assumption of independent predictors. In real life, it is almost impossible to get a set of predictors which are completely independent.
- If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as "Zero Frequency". To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.

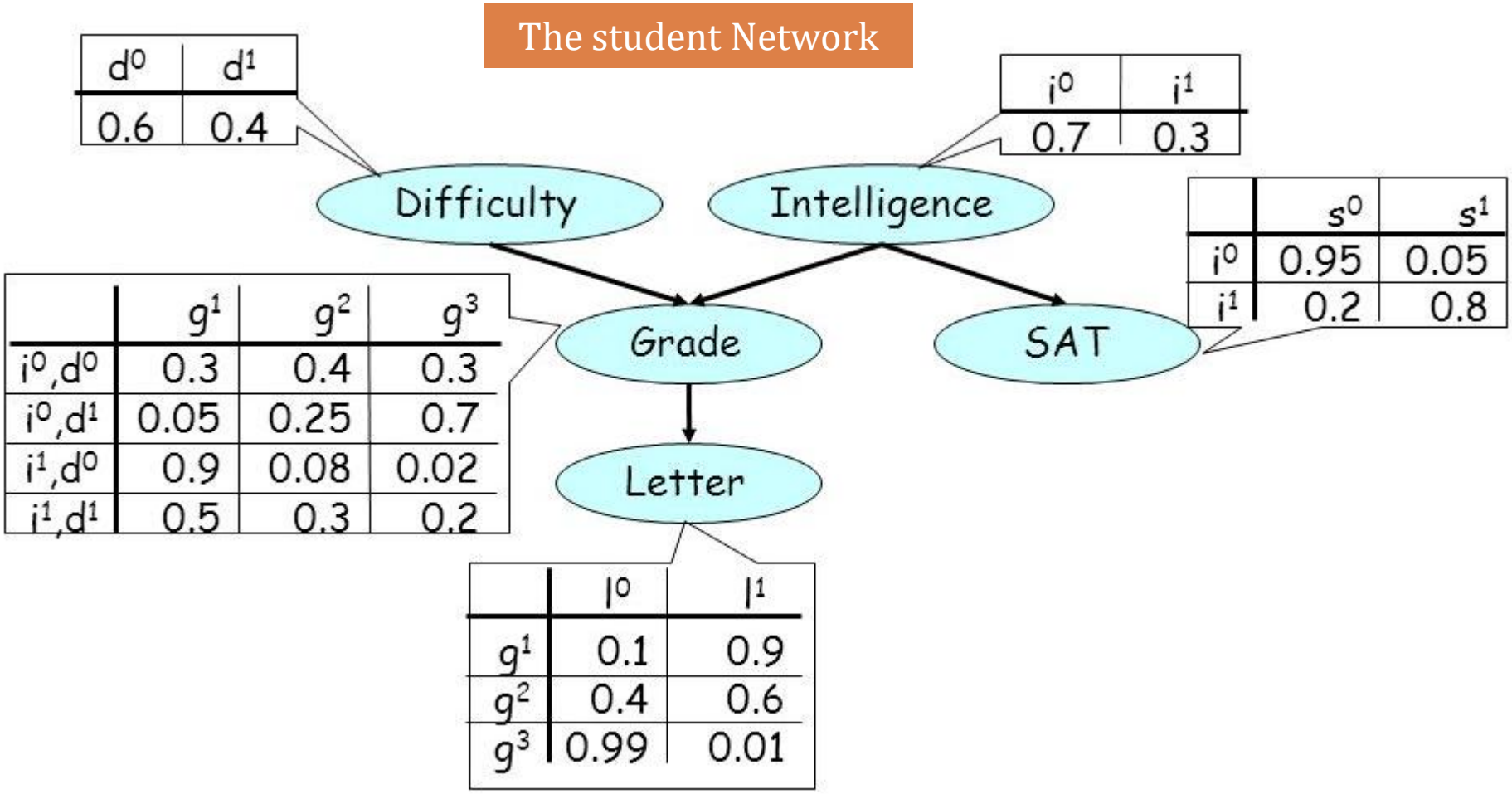
Bayesian Networks

- Probabilistic models can be challenging to design and use. Most often, the problem is the lack of information about the domain required to fully specify the conditional dependence between variables. If available, calculating the full conditional probability for an event can be impractical.
- A common approach to addressing this challenge is to add some simplifying assumptions, such as assuming that all random variables in the model are conditionally independent. This is a drastic assumption, although it proves useful in practice, providing the basis for the Naive Bayes classification algorithm.
- An alternative approach is to develop a probabilistic model of a problem with some conditional independence assumptions. This provides an intermediate approach between a fully conditional model and a fully conditionally independent model.
- Bayesian belief networks (BNN) are one example of a probabilistic model where some variables are conditionally independent.
- Thus, BNN provide an intermediate approach that is less constraining than the global assumption of conditional independence made by the Naive Bayes classifier, but more tractable than avoiding conditional independence assumptions altogether.
- A Bayesian belief network is a type of probabilistic graphical model.

Probabilistic Graphical Models

- Formally, a probabilistic graphical model (or graphical model for short) consists of a graph structure.
- Each node of the graph is associated with a random variable, and the edges in the graph are used to encode relations between the random variables.
- Depending on whether the graph is directed or undirected, we classify graphical modes into two categories — **Bayesian networks** and Markov networks.
- A canonical example of Bayesian networks is the so-called student network which is shown in the next. This graph describes a setting for a student enrolled in a university class. There are 5 random variables in the graph:
  - Difficulty (of the class): Takes values 0 (low difficulty) and 1 (high difficulty).
  - Intelligence (of the student): Takes values 0 (not intelligent) and 1 (intelligent)
  - Grade (the student gets in the class): Takes values 1 (good grade), 2 (average grade), and 3 (bad grade)
  - SAT (student's score in the SAT exam): Takes values 0 (low score) and 1 (high score)
  - Letter (quality of recommendation letter the student gets from the professor after completing the course): Takes values 0 (not a good letter) and 1 (a good letter)

Probabilistic Graphical Models cont...



Probabilistic Graphical Models cont...

The edges in the graph encode dependencies in the graph.

- The "Grade" of the student depends on the "Difficulty" of the class and the "Intelligence" of the student.
- The "Grade," in turn, determines whether the student gets a good letter of recommendation from the professor.
- The "Intelligence" of the student influences their "SAT" score, in addition to influencing the "Grade."
- The direction of arrows depicts cause-effect relationships — "Intelligence" affects the "SAT" score, but the "SAT" score does not influence the "Intelligence."

Finally, let's look at the tables associated with each of the nodes. Formally, these are called **conditional probability distributions** (**CPD**s).

- The CPDs for "Difficulty" and "Intelligence" are fairly simple, because these variables do not depend on any of the other variables. The tables basically encode the probabilities of these variables, taking 0 or 1 as values. As you might have noticed, the values in each of the tables must sum to 1.
- Next, let's look at the CPD for "SAT." Each row corresponds to the values that its parent ("Intelligence") can take, and each column corresponds to the values that "SAT" can take. Each cell has the conditional probability p(SAT=s | Intelligence=i), that is, given that the value of "Intelligence" is i, what is the probability of the value of "SAT" being s.

Probabilistic Graphical Models cont...

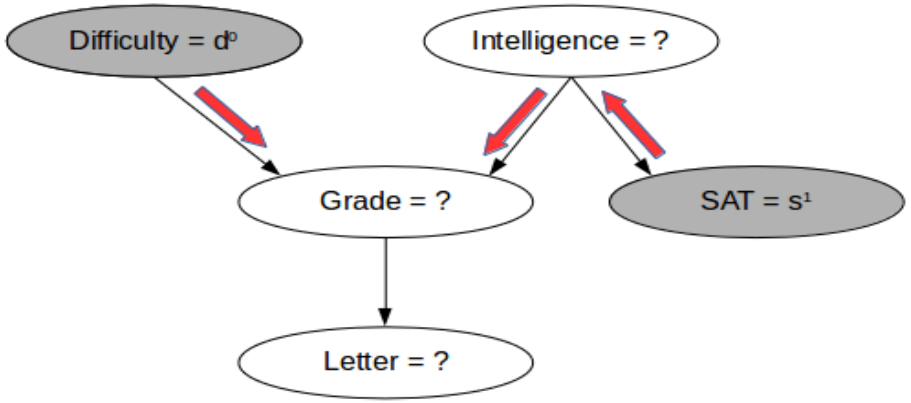
- For instance, we can see that p(SAT=s^1 | Intelligence = i^1) is 0.8, that is, if the intelligence of the student is high, then the probability of the SAT score being high as well is 0.8. On the other hand, p(SAT=s^0 | Intelligence = i^1), which encodes the fact that if the intelligence of the student is high, then the probability of the SAT score being low is 0.2.
- Note that the sum of values in each row is 1. That makes sense because given that Intelligence=i^1, the SAT score can be either s^0 or s^1, so the two probabilities must add up to 1. Similarly, the CPD for "Letter" encodes the conditional probabilities p(Letter=l | Grade=g). Because "Grade" can take three values, we have three rows in this table.
- The CPD for "Grade" is easy to understand with the above knowledge. Because it has two parents, the conditional probabilities will be of the form p(Grade=g | Difficulty=d, SAT=s), that is, what is the probability of "Grade" being g, given that the value of "Difficulty" is d and that of "SAT" is s. Each row now corresponds to a pair of values of "Difficulty" and "Intelligence." Again, the row values add up to 1.

An essential requirement for Bayesian networks is that the graph must be a directed acyclic graph (DAG).

Probabilistic Graphical Models cont…

The problem Setting:

- We have data about each student in each of the courses — their intelligence, what their SAT score was, what grade they got, and whether they got a good letter from the professor. From this data, we can estimate the parameters of the CPDs.
- For instance, the data might show that students with high intelligence often get good SAT scores, and we might be able to learn from it that $p(SAT=s^1 \mid Intelligence = i^1)$is high. This is the learning phase.
- Now, for a new data point, we will observe some of the variables, but not all. For example, in the graph below, we know about the difficulty of a course and a student's SAT score, and want to estimate the probability of the student getting a good grade.



Probabilistic Graphical Models cont...

- While we don't have a CPD that gives us that information directly, we can see that a high SAT score from the student would suggest that the student is likely intelligent, and consequently, the probability of a good grade is high if the difficulty of the course is low, as shown using the red arrows in the previous image. We may also want to estimate the probability of multiple variables simultaneously, like what is the probability of the student getting a good grade and a good letter?
- The variables with known values are called "observed variables," while those whose values are unobserved are called "hidden variables" or "latent variables." Conventionally, observed variables are denoted using grey nodes, while latent variables are denoted using white nodes, as in the previous image. We may be interested in finding the values of some or all of the latent variables.
- The graph structures that we've been talking about so far actually capture important information about the variables. Specifically, they define a set of **conditional independences** between the variables, that is, statements of the form — "If A is observed, then B is independent of C." Let's look at some examples.

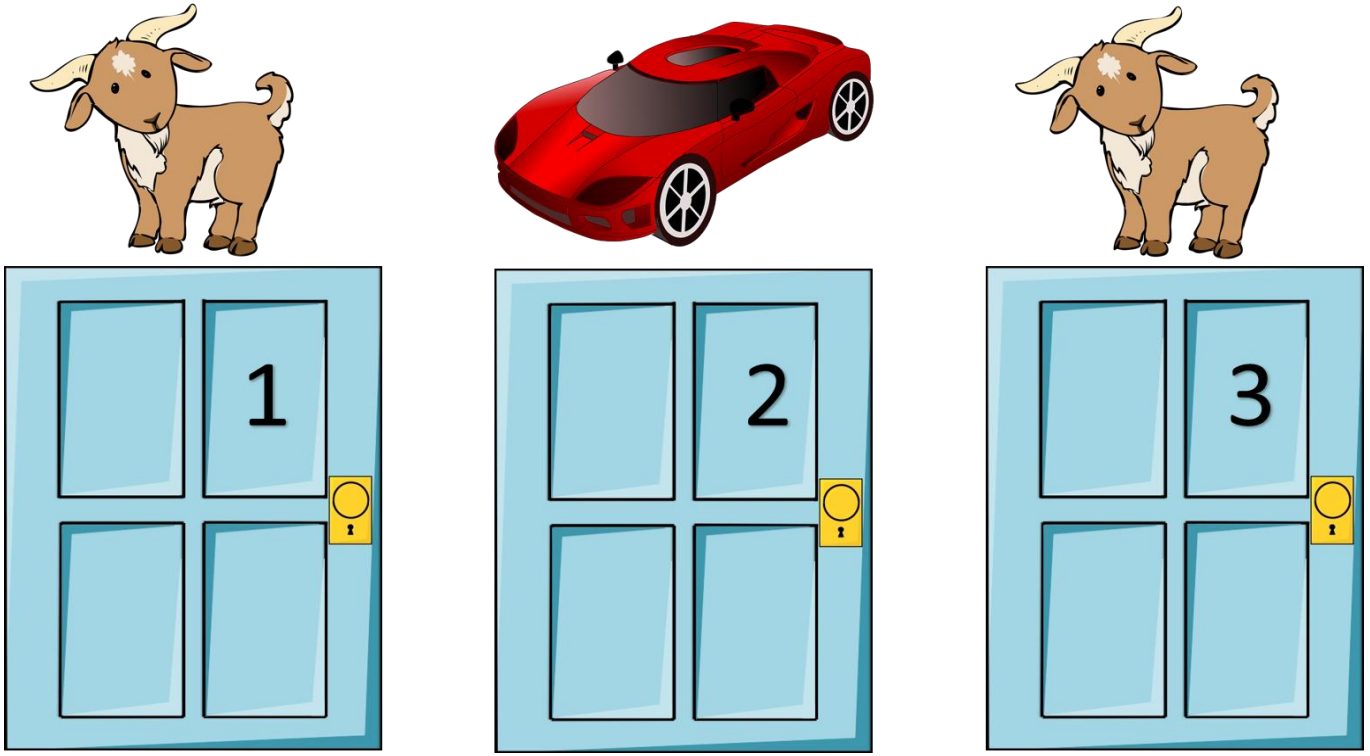
Probabilistic Graphical Models cont...

- In the student network, let's say you know that a student had a high SAT score. What can you say about her grade? As we saw earlier, a high SAT score suggests that the student is intelligent, and therefore, you would expect a good grade. What if the student has a low SAT score? In this case, you would not expect a good grade.
- Now, let's say that you also know that the student is intelligent, in addition to her SAT score. If the SAT score was high, then you would expect a good grade. What if the SAT score was low? You would still expect a good grade because you know that the student is intelligent, and you would assume that she just didn't perform well enough on the SAT. Therefore, knowing the SAT score doesn't tell us anything if we see the intelligence of the student. To put this as a conditional independence statement, we would say — "If Intelligence is observed, then SAT and Grade are independent."
- We got this conditional independence information from the way these nodes were connected in the graph. If they were connected differently, we would get different conditional independence information.

Probabilistic Graphical Models cont…

- Let's see this with another example. Let's say you know that the student is intelligent. What can you say about the difficulty of the course? Nothing, right? Now, what if I tell you that the student got a bad grade on the course? This would suggest that the course was hard because we know that an intelligent student got a bad grade. Therefore we can write our conditional independence statement as follows — "If Grade is unobserved, then Intelligence and Difficulty are independent."
- Because these statements capture an independence between two nodes subject to a condition, they are called conditional independences. Note that the two examples have opposite semantics — in the first one, the independence holds if the connecting node is observed; in the second one, the independence holds if the connecting node is unobserved. This difference is because of the way the nodes are connected, that is, the directions of arrows.

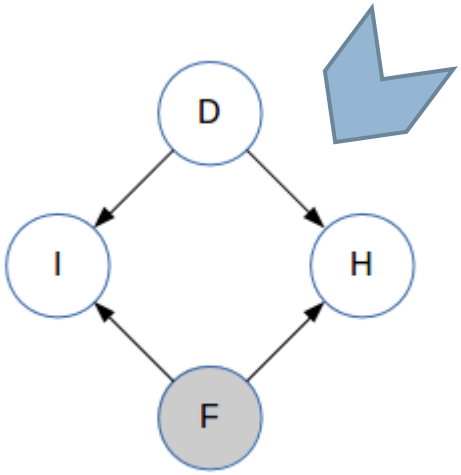
Application: Monty Hall Problem



The host of the game shows you three closed doors, with a car behind one of the doors and something invaluable behind the others. You get to pick a door. Then, the host opens one of the remaining doors, and shows that it does not contain the car. Now, you have an option to switch the door, from the one you picked initially to the one that the host left unopened. Do you switch?

Application: Monty Hall Problem cont...

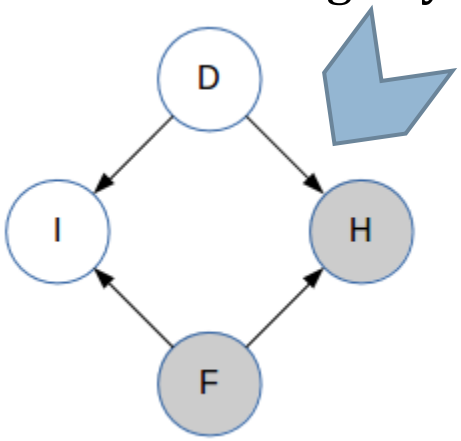
Let's start by defining some variables:
D : The **d**oor with the car.
F : Your **f**irst choice.
H : The door opened by the **h**ost.
I : **I**s F = D?
D, F, and H take values 1, 2, or 3 and I takes values 0 or 1. D and I are unobserved, while F is observed. Until the host opens one of the doors, H is unobserved. Therefore, we get the following Bayesian network for our problem:



Note *the directions of arrows — D and F are independent, I clearly depends on D and F, and the door picked by the host H also depends on D and F. So far, we don't know anything about D. (This is similar to the structure in the student network, where knowing the intelligence of the student does not tell you anything about the difficulty of the course.)*

Application: Monty Hall Problem cont...

Now, the host picks a door H and opens it. So, H is now observed. Therefore, we get the following Bayesian network for our problem:



Observing H does not tell us anything about I, that is, whether we have picked the right door. That is what our intuition tells us. However, it does tell us something about D (Again, drawing analogy with the student network, if you know that the student is intelligent, and the grade is low, it tells you something about the difficulty of the course.)

Let's see this using numbers. The CPD tables for the variables are as follows (This is when no variables have been observed.):

p(D)

1	2	3
1/3	1/3	1/3

p(F)

1	2	3
1/3	1/3	1/3

Application: Monty Hall Problem cont...

p(I | D, F)

	0	1
D=1, F=1	0	1
D=1, F=2	1	0
D=1, F=3	1	0
D=2, F=1	1	0
D=2, F=2	0	1
D=2, F=3	1	0
D=3, F=1	1	0
D=3, F=2	1	0
D=3, F=3	0	1

I=1 when D and F are identical, and I=0 when D and F are different.

p(H | D, F)

	1	2	3
D=1, F=1	0	1/2	1/2
D=1, F=2	0	0	1
D=1, F=3	0	1	0
D=2, F=1	0	0	1
D=2, F=2	1/2	0	1/2
D=2, F=3	1	0	0
D=3, F=1	0	1	0
D=3, F=2	1	0	0
D=3, F=3	1/2	1/2	0

If D and F are equal, then the host picks one door from the other two with equal probability, while if D and F are different, then the host picks the third door.

Application: Monty Hall Problem cont…

Now, let's assume that we have picked a door, that is, F is now observed, say F=1. What are the conditional probabilities of I and D, given F?

$$p(I|F=1) = \frac{p(I, F=1)}{p(F=1)} = \frac{\sum_D p(I|F=1, D)p(D)}{p(F=1)}$$

$$p(D|F=1) = \frac{p(D, F=1)}{p(F=1)} = \frac{p(D)p(F=1)}{p(F=1)}$$

Using these equations, we get the following probabilities:

p(I | F=1)	
0	1
2/3	1/3

P(D | F=1)		
1	2	3
1/3	1/3	1/3

So far, the probability that we have picked the correct door is 1/3 and the car could still be behind any door with equal probability.

Application: Monty Hall Problem cont...

Now, the host opens one of the doors other than F, so we observe H. Assume H=2. Let's compute the new conditional probabilities of I and D given both F and H.

$$p(D|F=1,H=2) = \frac{p(D,F=1,H=2)}{p(F=1,H=2)} = \frac{p(D)p(F=1)p(H=2|D,F=1)}{\sum_D p(D)p(F=1)p(H=2|D,F=1)}$$

$$p(I|F=1,H=2) = \frac{p(I,F=1,H=2)}{p(F=1,H=2)} = \frac{\sum_D p(I|D,F=1)p(D)p(F=1)p(H=2|D,F=1)}{\sum_D p(D)p(F=1)p(H=2|D,F=1)}$$

Using these equations, we get the following probabilities:

p(I | H=2)

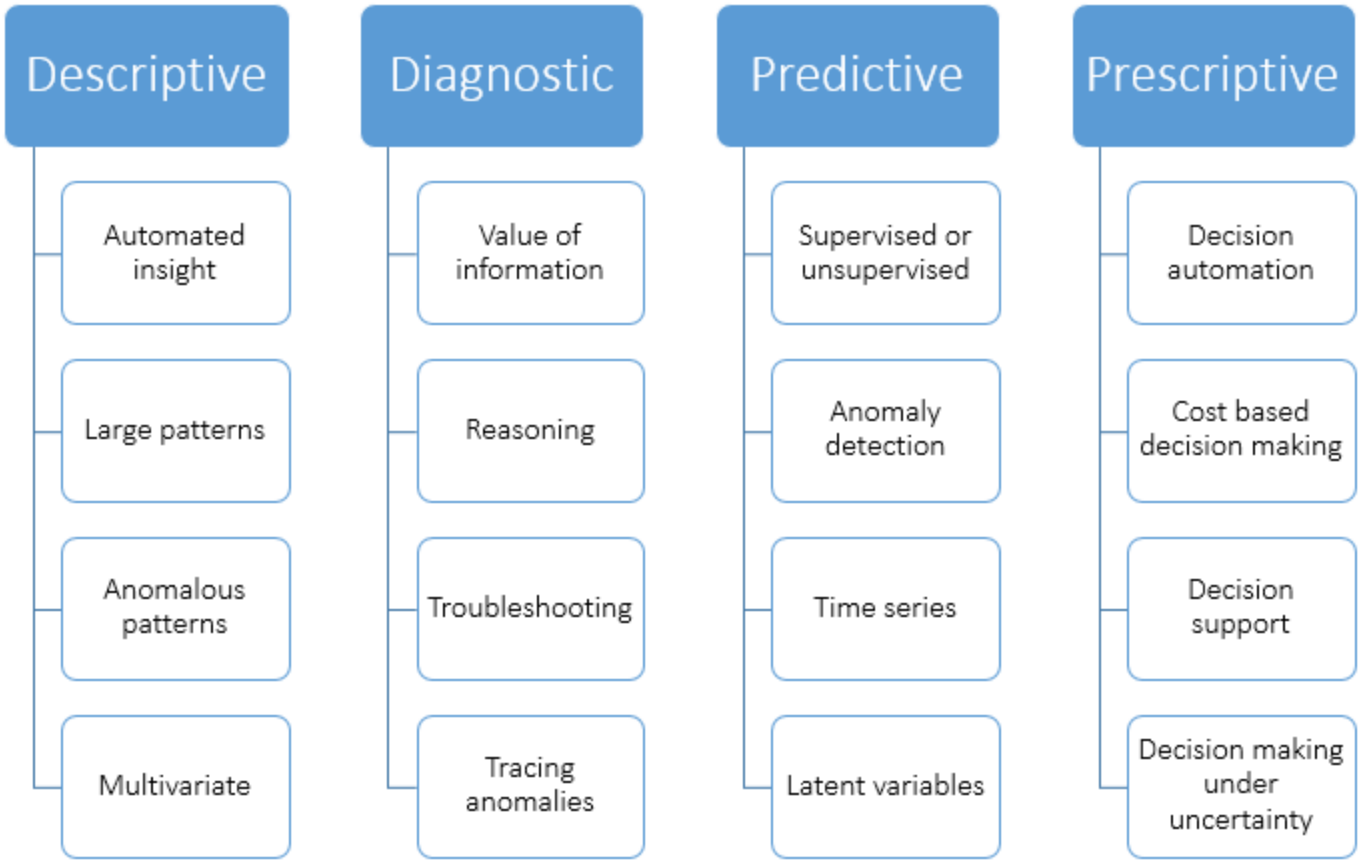
0	1
2/3	1/3

P(D | F=1, H = 2)

1	2	3
1/3	0	2/3

Our first choice is correct with probability 1/3. So, if we switch, we get the car with probability 1/3, if we don't, we get the car with probability 1/3.

Analytics with Bayesian networks



Support Vector Machines

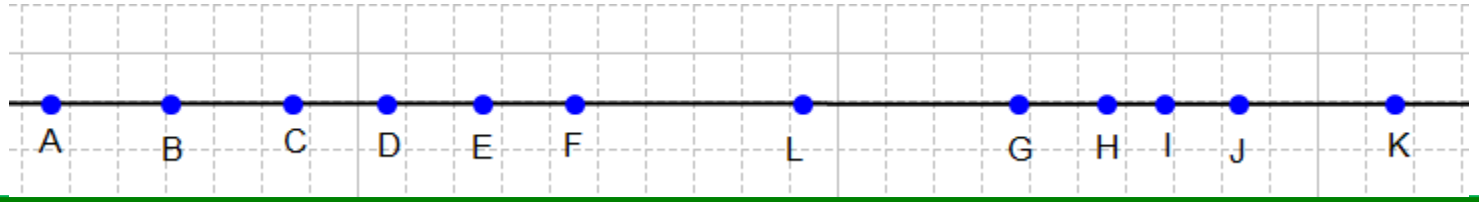
Support Vector Machine (SVM) is an algorithm which can be used for both classification or regression challenges. However, it is mostly used in two-group classification problems. In the SVM algorithm, each data item is plotted as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyperplane that differentiates the two classes.

What is hyperplane?

A hyperplane is a generalization of a plane.

- In one dimension, it is called a point.
- In two dimensions, it is a line.
- In three dimensions, it is a plane.
- In more dimensions one can call it an hyperplane.

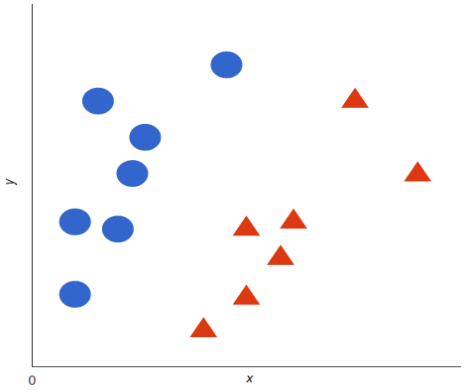
The following figure represents datapoint in one dimension and the point L is a separating hyperplane.



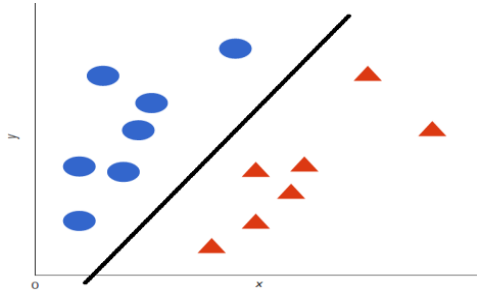
Support Vector Machines cont...

How does SVM works?

- Let's imagine we have two tags: red and blue, and our data has two features: x and y. We want a classifier that, given a pair of (x, y) coordinates, outputs if it's either red or blue. We plot our already labeled training data on a plane.

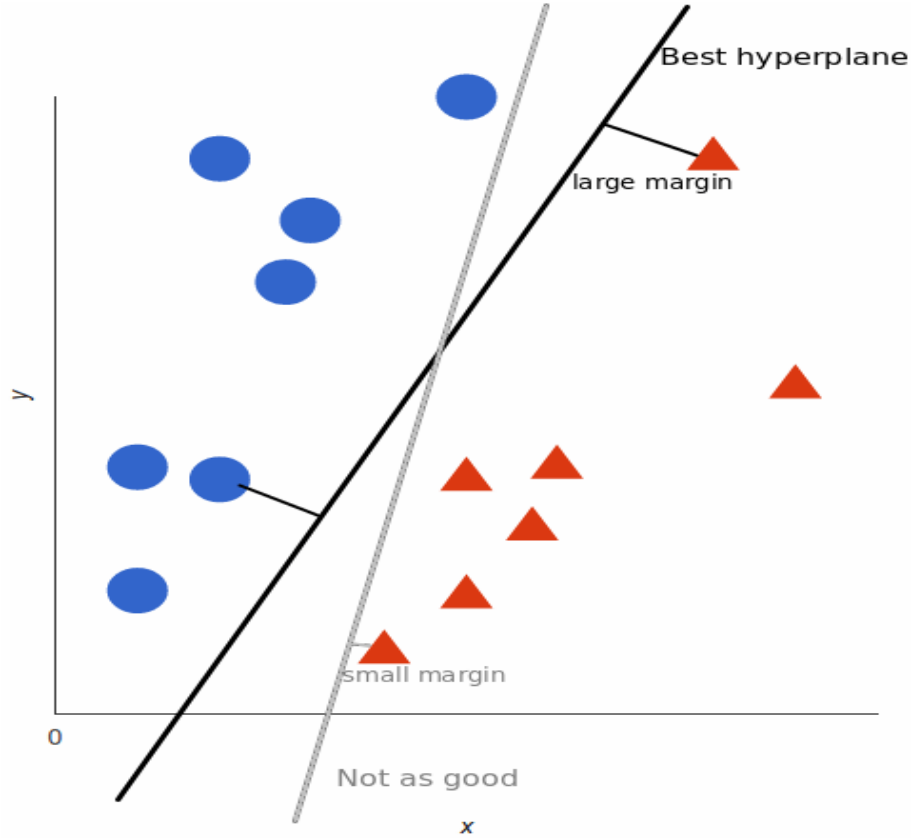


- A support vector machine takes these data points and outputs the hyperplane (which in two dimensions it's simply a line) that best separates the tags. This line is the **decision boundary**: anything that falls to one side of it we will classify as blue, and anything that falls to the other as red.
- In 2D, the best hyperplane is simply a line.



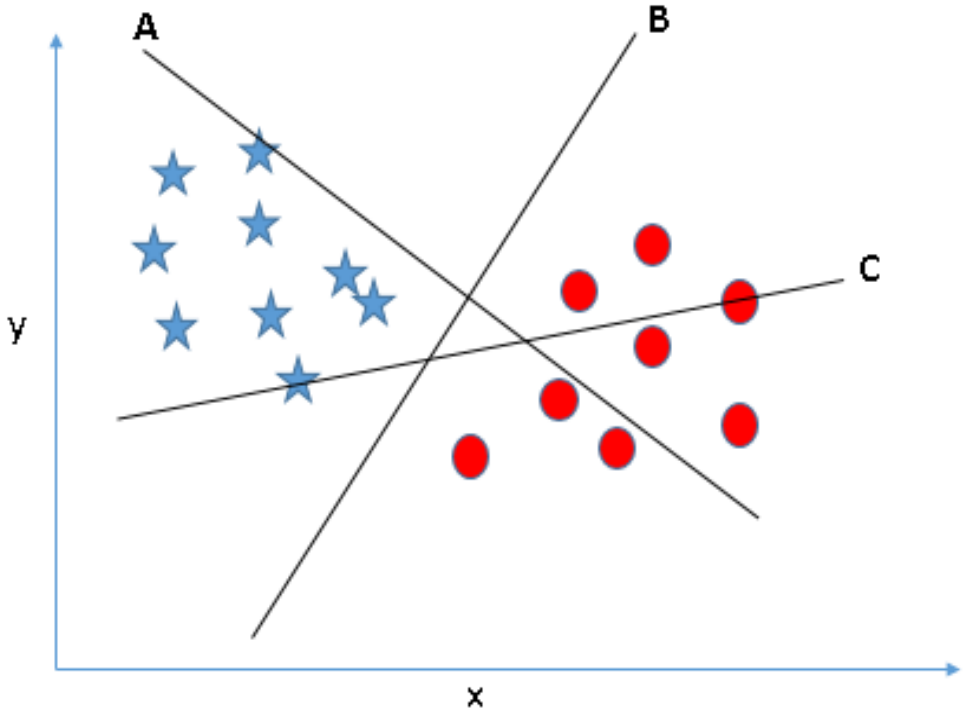
Support Vector Machines cont...

But, what exactly is the best hyperplane? For SVM, it's the one that maximizes the **margins** from both tags. In other words: the hyperplane (in 2D, it's a line) whose distance to the nearest element of each tag is the largest.



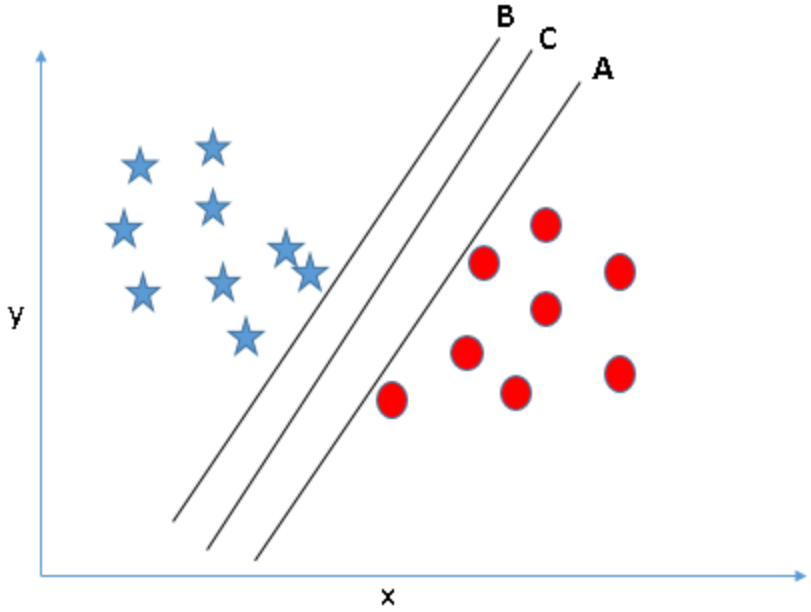
Support Vector Machines cont...

(Scenario-1) Identification of the right hyperplane: Here, we have three hyperplanes (A, B and C). Now, the job is to identify the right hyperplane to classify star and circle. Remember a thumb rule - identify the right hyperplane: "Select the hyperplane which segregates the two classes better". In this scenario, hyperplane "B" has excellently performed this job.



Support Vector Machines cont...

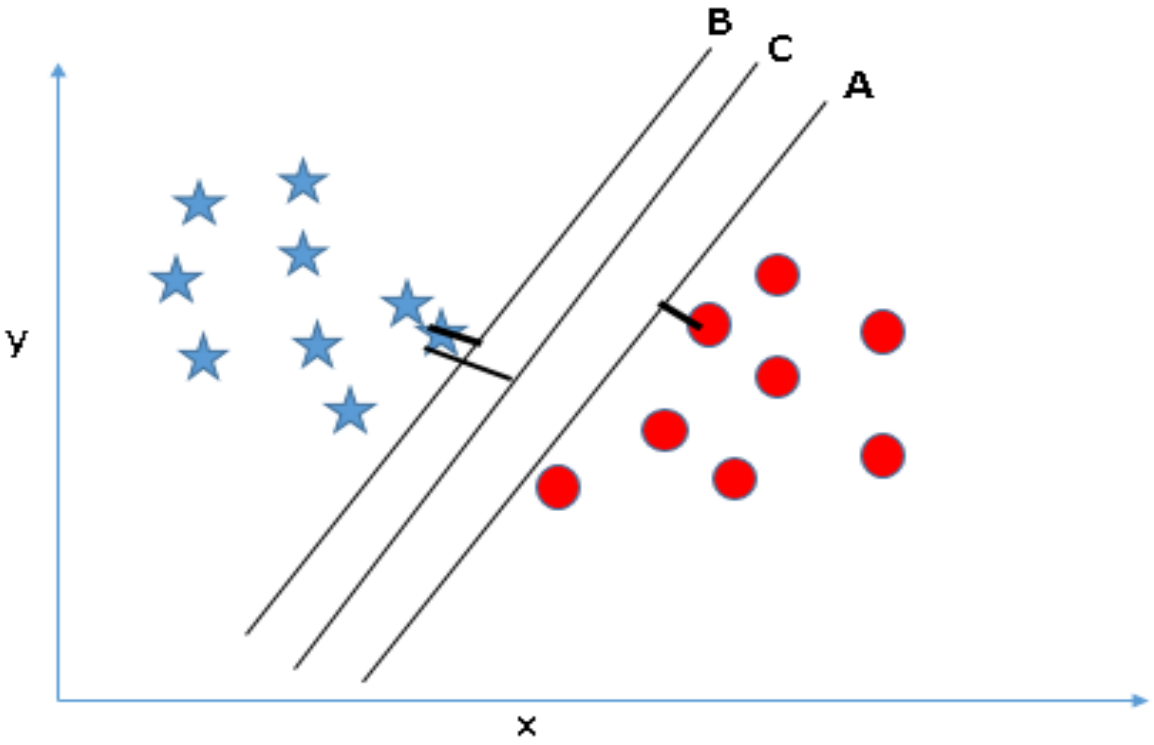
(Scenario-2) Identify the right hyperplane: Here, we have three hyperplanes (A, B and C) and all are segregating the classes well. Now, How can we identify the right hyperplane?



Here, maximizing the distances between nearest data point (either class) and hyperplane will help to decide the right hyperplane. This distance is called as **margin**.

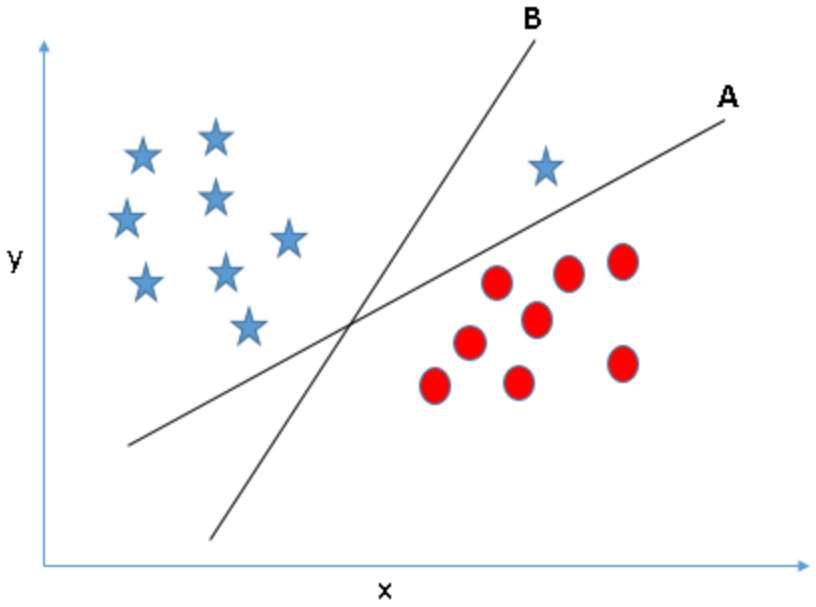
Support Vector Machines cont...

Below, you can see that the margin for hyperplane C is high as compared to both A and B. Hence, we name the right hyperplane as C. Another lightning reason for selecting the hyperplane with higher margin is robustness. If we select a hyperplane having low margin then there is high chance of misclassification.



Support Vector Machines cont...

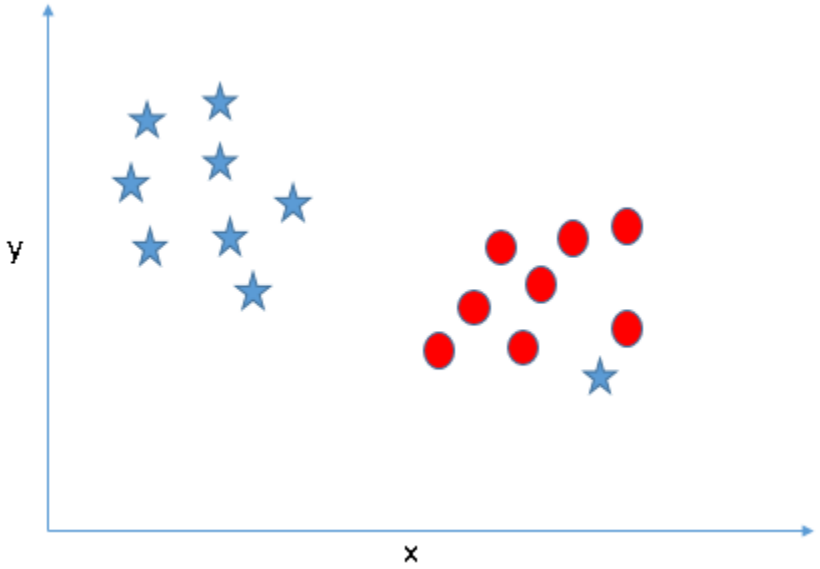
(Scenario-3) Identify the right hyperplane: Use the rules as discussed in previous section to identify the right hyperplane.



Some of you may have selected the hyperplane B as it has higher margin compared to A. But, here is the catch, SVM selects the hyperplane which classifies the classes accurately prior to maximizing margin. Here, hyperplane B has a classification error and A has classified all correctly. Therefore, the right hyperplane is A.

Support Vector Machines cont...

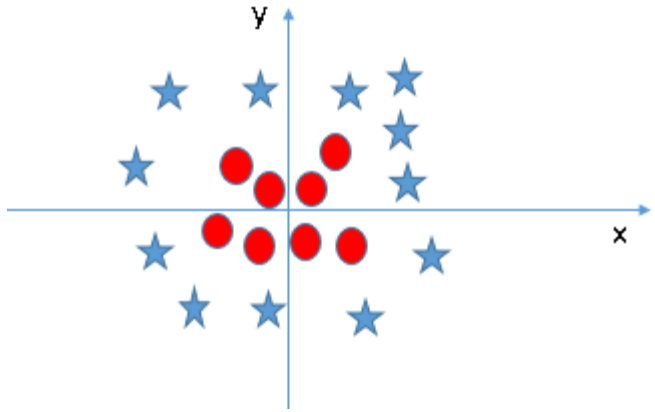
(Scenario-4) Below, we are unable to segregate the two classes using a straight line, as one of the stars lies in the territory of other(circle) class as an outlier.



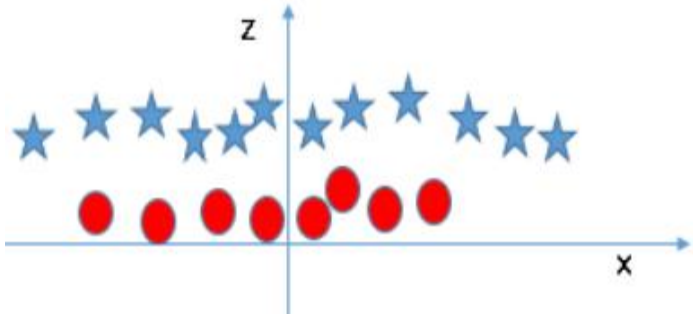
One star at other end is like an outlier for star class. The SVM algorithm has a feature to ignore outliers and find the hyperplane that has the maximum margin. Hence, we can say, SVM classification is robust to outliers.

Support Vector Machines cont...

(Scenario-5) Find the hyperplane to segregate to classes: In the scenario, we can't have linear hyperplane between the two classes, so how does SVM classify these two classes? Till now, we have only looked at the linear hyperplane.



SVM can solve this problem. Easily! It solves this problem by introducing additional feature. Here, we can add a new feature $z=x^2+y^2$. Now, plotting the data points on axis x and z:

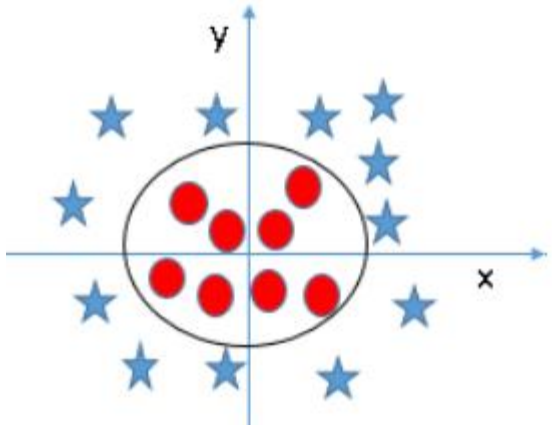


In above plot, points to consider are:

- All values for z would be positive as z is the squared sum of both x and y.
- In the original plot, red circles appear close to the origin of x and y axes, leading to lower value of z and star relatively away from the origin result to higher value of z.

Support Vector Machines cont…

- In the SVM classifier, it is easy to have a linear hyperplane between these two classes. But, another burning question which arises is, should we need to add this feature manually to have a hyperplane. No, the SVM algorithm has a technique called the **kernel** trick. The SVM kernel takes low dimensional input space and transforms it to a higher dimensional space i.e. it converts not separable problem to separable problem. It is mostly useful in non-linear separation problem. Simply put, it does some extremely complex data transformations, then finds out the process to separate the data based on the labels or outputs you've defined.
- When we look at the hyperplane in original input space it looks like a circle.



Types of SVM

- **Linear SVM:** Linear SVM is used for data that are linearly separable i.e. for a dataset that can be categorized into two categories by utilizing a single straight line. Such data points are termed as linearly separable data, and the classifier is used described as a linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is used for data that are non-linearly separable data i.e. a straight line cannot be used to classify the dataset. For this, we use something known as a kernel trick that sets data points in a higher dimension where they can be separated using planes or other mathematical functions. Such data points are termed as non-linear data, and the classifier used is termed as a non-linear SVM classifier.

Time Series Analysis

- Whether to predict the trend in financial markets or electricity consumption, time is an important factor that must be considered in the model. For example, it would be interesting to forecast at what hour during the day there going to be a peak consumption in electricity, such as to adjust the price or the production of electricity.
- A time series is simply a series of data points ordered in time. In a time series, time is often the independent variable and the goal is usually to make a forecast for the future. As the name suggests, it involves working on time (years, weeks, days, hours, minutes) based data, to derive hidden insights to make informed decision making.
- **Example of Time Series Data**:

Field	Example Topics
Economics	Gross Domestic Product (GDP), Consumer Price Index (CPI), and unemployment rates
Medicine	Blood pressure tracking, weight tracking, cholesterol measurements, heart rate monitoring
Physical sciences	Global temperatures, monthly sunspot observations, pollution levels.
Social Sciences	Birth rates, population, migration data, political indicators
Epidemiology	Disease rates, mortality rates, mosquito populations

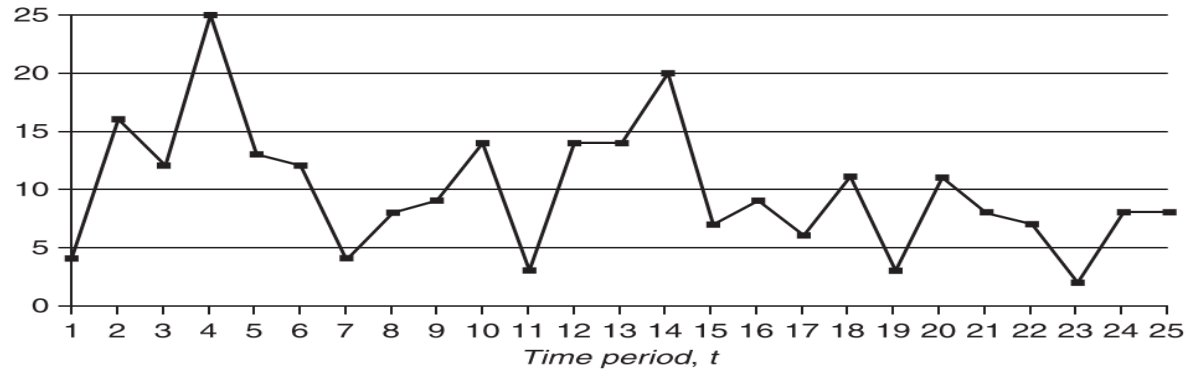
Time Series Model

- A time series is a sequential set of data points, measured typically at successive times. It is mathematically defined as a set of vectors x(t) where t = 0,1, 2… where t represents the time elapsed. The variable x(t) is treated as random variable.
- A time series model generally reflect the fact that observations close together in time which are closely related than the observations further apart.
- The data shown below represent the weekly demand of some product. The model uses x to indicate an observation and t to represent the index of the time period. The data from 1 to t is: $x_1, x_2, \ldots x_t$. Time series of 25 periods is shown below.

Weekly demand

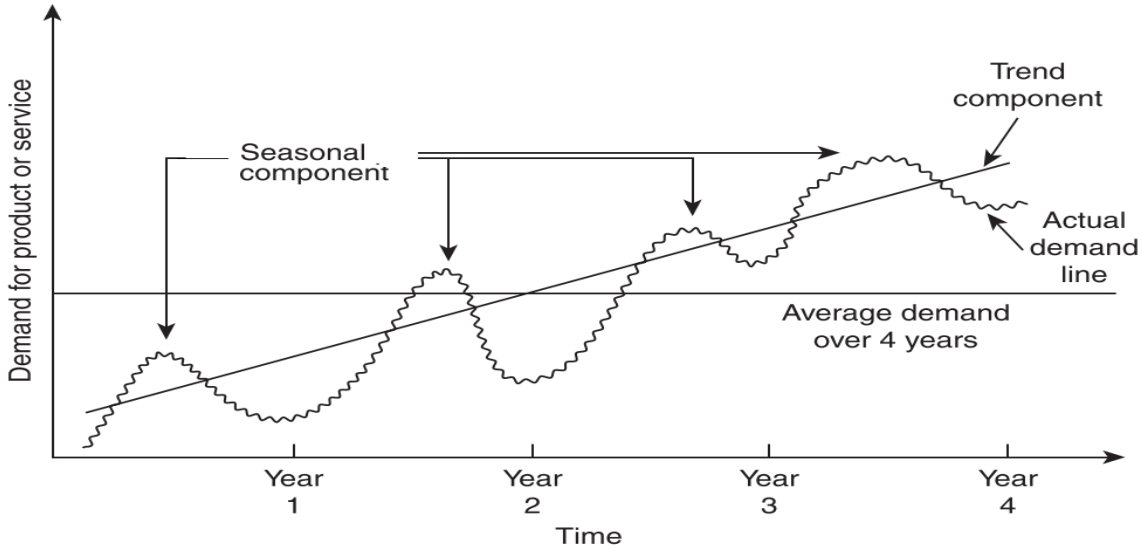
Time	Observations									
1 – 10	4	16	12	25	13	12	4	8	9	14
11 – 20	3	14	14	20	7	9	6	11	3	11
21 – 30	8	7	2	8	8	10	7	16	9	4

Time series of weekly demand -25 periods



Time Series Model cont...

- Any time series is composition of many individual component times series. Some of these components are predictable whereas other components may be almost random which can be difficult to predict.
- This calls for the decomposition methods that will generate individual component series from the original series. Decomposing a series into such components enable to analyze the behaviour of each component and thus improve the accuracy of the final forecast.
- **Example:** A typical sales time series.



Time Series Model Component

Time series models are characterized of four components:

- Trend component
- Seasonal component
- Cyclical component
- Irregular component

Trend component

- The trend shows the general tendency of the data to increase or decrease during a long period of time. A trend is a smooth, general, long-term, average tendency.
- It is not always necessary that the increase or decrease is in the same direction throughout the given period of time.
- It is observable that the tendencies may increase, decrease or are stable in different sections of time. But the overall trend must be upward, downward or stable.
- The population, agricultural production, items manufactured, number of births and deaths, number of industry or any factory, number of schools or colleges are some of its example showing some kind of tendencies of movement.

Time Series Model Component cont...

Seasonal component

- These are the rhythmic forces which operate in a regular and periodic manner over a span of less than a year. They have the same or almost the same pattern during a period of 12 months. This variation will be present in a time series if the data are recorded hourly, daily, weekly, quarterly, or monthly.
- These variations come into play either because of the natural forces or person-made conventions. The various seasons or climatic conditions play an important role in seasonal variations. Such as production of crops depends on seasons, the sale of umbrella and raincoats in the rainy season, and the sale of electric fans and A.C. shoots up in summer seasons.
- The effect of person-made conventions such as some festivals, customs, habits, fashions, and some occasions like marriage is easily noticeable. They recur themselves year after year. An upswing in a season should not be taken as an indicator of better business conditions.

Time Series Model Component cont...

Cyclical component

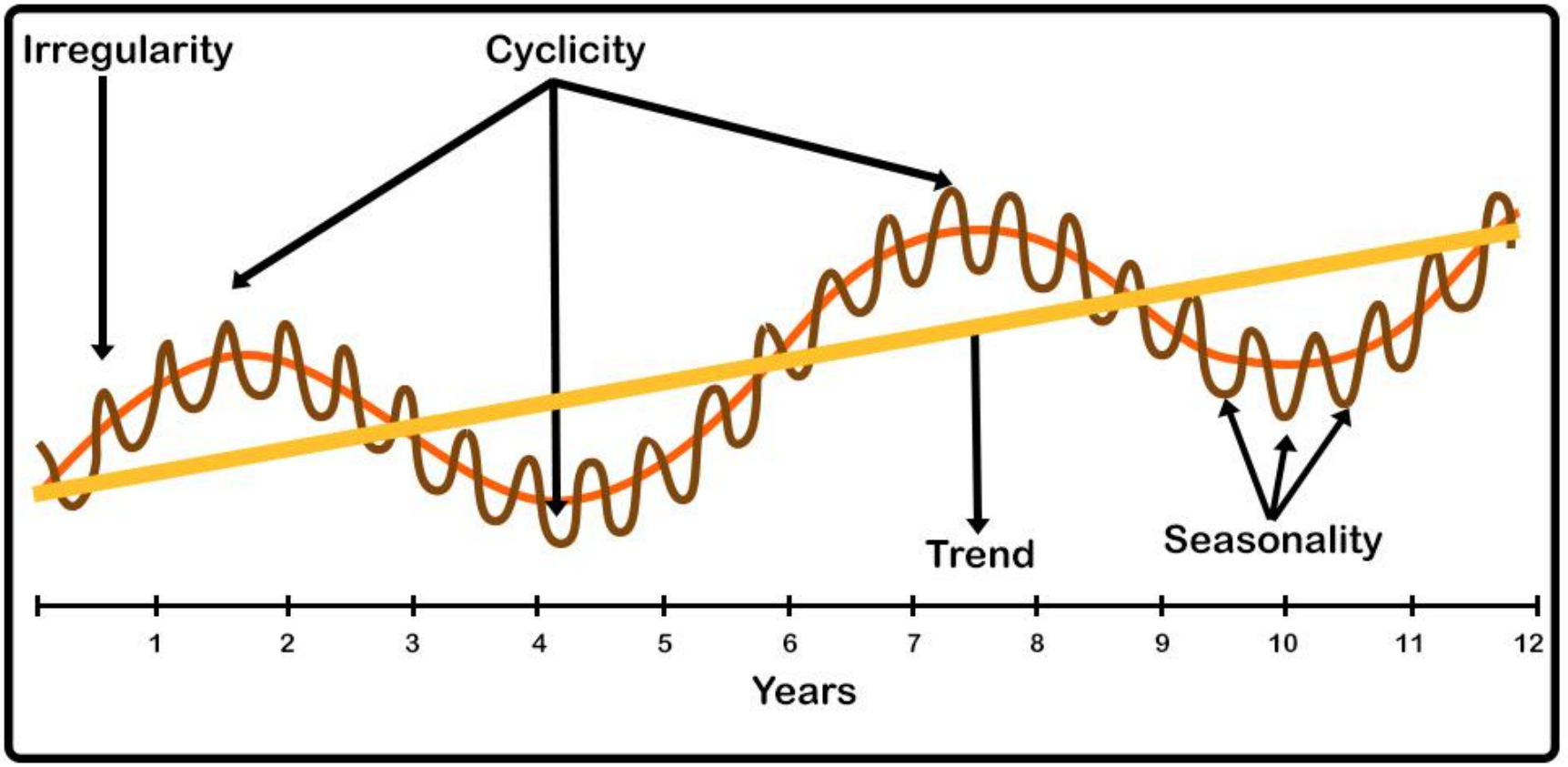
- The variations in a time series which operate themselves over a span of more than one year are the cyclic variations. This oscillatory movement has a period of oscillation of more than a year. One complete period is a cycle. This cyclic movement is sometimes called the 'Business Cycle'.
- It is a four-phase cycle comprising of the phases of prosperity, recession, depression, and recovery. The cyclic variation may be regular are not periodic. The upswings and the downswings in business depend upon the joint nature of the economic forces and the interaction between them.

Irregular component

- They are not regular variations and are purely random or irregular. These fluctuations are unforeseen, uncontrollable, unpredictable, and are erratic. These forces are earthquakes, wars, flood, famines, and any other disasters.

Time Series Model Component cont...

Pictorial depiction of different component



Decomposition Model

- Mathematical representation of the decomposition approach is $Y_t = f(T_t, S_t, C_t, I_t)$ where Y_t is the time series value at time t. T_t, S_t, C_t, and I_t are the trend, seasonal, cyclic and irregular component value at time t respectively.
- There are 3 types of decomposition model:
  - Additive model
  - Multiplicative model
  - Mixed model

Additive model

- According to this model, a time series is expressed as $Y_t = T_t + S_t + C_t + I_t$
- The model is appropriate when the amplitude of both the seasonal and irregular variations do not change as the level of trend rises or falls.
- This model assumes that all four components of the time series act independently of each other.

Multiplicative model

- According to this model, a time series is expressed as $Y_t = T_t * S_t * C_t * I_t$
- The model is appropriate when the amplitude of both the seasonal and irregular variations increase as the level of trend rises.
- The model assumes that the various components operate proportionately to each other.

Decomposition Model cont…

Mixed model

- Different assumptions lead to different combinations of additive and multiplicative models as $Y_t = T_t + S_t + C_t * I_t$
- The time series analysis can also be done using the model as:
  - $Y_t = T_t + S_t * C_t * I_t$
  - $Y_t = T_t * S_t + C_t * I_t$

Home Work

- How to determine if a time series has a trend component?
- How to determine if a time series has a seasonal component?
- How to determine if a time series has both a trend and seasonal component?

Time Series Forecasting Model

- Time series forecasting models can be classified into 2 categories.
- One group is called as **averaging methods** in which all observations (time series values) are equally weighted.
- The second group called **exponential smoothing methods** that applies unequal weights to past data, typically decaying in an exponential manner as one goes from recent to distinct past.

Averaging Model

- The simple average method uses the mean of all the past values to forecast the next value. This method is seen to be no use in a practical scenario.
- This method is used when the time series has attained some level of stability and no longer dependent on any external parameters.
- This would happen in sales forecasting, only when the product for which the forecast is needed is at a mature stage in its life cycle.
- The Averaging Model is represented as follows where F is the forecasted value at instance of time t+1, t is the current time and Y_i is the value of series at time instant i.

$$F_{t+1} = \frac{1}{t}\sum_{i=1}^{t} Y_i$$

Averaging Model

Supplier	Amount
1	9
2	8
3	9
4	12
5	9
6	12
7	11
8	7
9	13
10	9
11	11
12	10

A manager of a warehouse wants to know how much a typical supplier delivers in 10 dollar units. He/she has taken a sample of 12 suppliers at random, obtaining the result as shown in the table.

The computed mean of the amount is 10 and hence the manager decides to use this as the estimate for the expenditure of a typical supplier.

It is more reasonable to assume that the recent points in past are better predictors than the whole history. This is particularly true for sales forecasting. Every product has a life cycle, initial stage, middle volatile period and a more or less stable mature stage and an end stage. Hence, a better method of forecasting would be to use moving averages (MAs).

Moving Averages (MAs)

- The MA approach calculates an average of a finite number of past observations and then employs that average as the forecast for the next period.
- The number of sample observations to be included in the calculation of the average is specified at the start of the process. The term MA refer to the fact that as a new observation becomes available, a new average is calculated by dropping the oldest observation in order to include the newest one.
- An MA of order k, represented with MA(k) is calculated as:

$$F_{t+1} = \frac{1}{k} \sum_{i=t-k+1}^{t} Y_i$$

- MA(3), MA(5) and MA(12) are commonly used for monthly data and MA(4) is normally used for quarterly data.
- MA(4), and MA(12) would average out the seasonality factors in quarterly and monthly data respectively.
- The advantage of MA method is that the data requirement is very small.
- The major disadvantage is that it assumes the data to be stationary.
- MA also called as **simple moving average.**

Moving Averages (MAs) cont...

Month	Demand
1	89
2	57
3	144
4	221
5	177
6	280
7	223
8	286
9	212
10	275
11	188
12	312

- MA(3) = (275 + 188 + 312) / 3 = 258.33
- MA(6) = (223+286+212+275+188+312)/6 = 249.33
- MA(12) = (89+57+144+221+177+280+223+286+212+ 275+188+312)/ 12 = 205.33

Home Work

Calculate:

- MA(5)
- MA(4)
- MA(10)

Exponential Smoothing Model

- The extension to the MA method is to have a weighted MA, whereas in Single Moving Averages the past observations are weighted equally, Exponential Smoothing assigns exponentially decreasing weights as the observation get older. In other words, recent observations are given relatively more weight in forecasting than the older observations.
- In the case of moving averages, the weights assigned to the observations are the same and are equal to 1/N. In exponential smoothing, however, there are one or more smoothing parameters to be determined (or estimated) and these choices determine the weights assigned to the observations.
- This class of techniques consists of a range of methods, starting from simple exponential smoothing (SES) used for the data with no trend or seasonality, to the sophisticated widely used Holt's or Holt-Winters' method which is able to provide forecasts for data that exhibit both seasonality and trend.
- ***In these methods, the observations are weighted in an exponentially decreasing manner as they become older.***

Simple Exponential Smoothing

- For any time period t, the smoothed value S_t is found by computing $S_t=\alpha * y_{t-1}+(1-\alpha) * S_{t-1}$ where $0<\alpha\leq 1$ and $t\geq 2$ and y_t represents the actual value at time t, S_t smoothed observation at time t, α is the smoothing constant.
- **Why is it called Exponential?**

  Let us expand the basic equation by first substituting for S_{t-1} in the basic equation to obtain:

  $S_t=\alpha * y_{t-1}+(1-\alpha) * [\alpha * y_{t-2}+(1-\alpha) * S_{t-2}]$

  $= \alpha * y_{t-1}+ \alpha * (1-\alpha) * y_{t-2}+ (1-\alpha)^2 * S_{t-2}$

  By substituting for S_{t-2}, then for S_{t-3}, and so forth, until we reach S2 (which is just y_1), it can be shown that the expanding equation can be written as:

$$S_t = \alpha \sum_{i=1}^{t-2}(1-\alpha)^{i-1} y_{t-i} + (1-\alpha)^{t-2} S_2\,, \quad t \geq 2$$

  This illustrates the exponential behavior. The weights, $\alpha * (1-\alpha)^t$ decrease geometrically.

Simple Exponential Smoothing cont…

- **What is the best value for α?**

  The speed at which the older responses are dampened (smoothed) is a function of the value of α. When α is close to 1, dampening is quick and when α is close to 0, dampening is slow. This is illustrated in the table below.

α	(1-α)	$(1-\alpha)^2$	$(1-\alpha)^3$	$(1-\alpha)^4$
0.9	0.1	0.01	0.001	0.0001
0.5	0.5	0.25	0.125	0.0625
0.1	0.9	0.81	0.729	0.6561

Error calculation

- The error is calculated as $E_t = y_t - S_t$ (i.e. difference of actual and smooth at time t)
- Then error square is calculated i.e. $ES_t = E_t * E_t$
- Then, sum of the squared errors (SSE) is calculated i.e. $SSE = \Sigma ES_i$ for i = 1 to n where n is the number of observations.
- Then, the mean of the squared errors is calculated i.e. MSE = SSE/(n-1)
- ***The best value for α is choose so the value which results in the smallest MSE.***

Simple Exponential Smoothing cont…

Let us illustrate this principle with an example. Consider the following data set consisting of 12 observations taken over time with α as 0.1:

Time	y_t	S_t	E_t	ES_t
1	71			
2	70	0.1 * 70 + (1-0.1) * 71 = 71	70 - 71= -1.0	$(-1.0)^2 = 1.00$
3	69	0.1 * 69 + (1-0.1) * 71 = 70.9	69 – 70.9 = -1.90	$(-1.90)^2 = 3.61$
4	68	70.71	-2.71	7.34
5	64	70.44	-6.44	41.47
6	65	69.80	-4.80	23.04
7	72	69.32	2.68	7.18
8	78	69.58	8.42	70.90
9	75	70.43	4.57	20.88
10	75	70.88	4.12	16.97
11	75	71.29	3.71	13.76
12	70	71.67	-1.67	2.79

Simple Exponential Smoothing cont…

- The sum of the squared errors (SSE) = 208.94. The mean of the squared errors (MSE) is the SSE /11 = 19.0.
- In the similar fashion, the MSE was again calculated for α=0.5 and turned out to be 16.29, so in this case we would prefer an α of 0.5.
- **Can we do better?**
  - We could apply the proven trial-and-error method. This is an iterative procedure beginning with a range of α between 0.1 and 0.9.
  - We determine the best initial choice for α and then search between $\alpha-\Delta$ and $\alpha+\Delta$. We could repeat this perhaps one more time to find the best α to 3 decimal places.

In general, most well designed statistical software programs should be able to find the value of α that minimizes the MSE.

Holt's Method

- Holt (1957) extended simple exponential smoothing to allow the forecasting of data with a trend. This method involves a forecast equation and two smoothing equations (i.e. one for the level and one for the trend).
- This method is used when a series has no seasonality but exhibits some form of trend.
- The k step ahead forecast function for a given time series X is
  $\mathbf{X_{t+k} = \ell_t + k * b_t}$ where ℓ_t denotes an estimate of the level of the series at time t, b_t denotes an estimate of the trend (slope) of the time series at time t.
- The equation for level is
  $\boldsymbol{\ell_t = \alpha * y_t + (1 - \alpha) * (\ell_{t-1} + b_{t-1})}$
- The equation for trend is
  $\mathbf{b_t = \beta * (\ell_t - \ell_{t-1}) + (1 - \alpha) * (1 - \beta) * b_{t-1}}$
  where,
  α is the smoothing parameter for the level, $0 \le \alpha \le 1$,
  β is the smoothing parameter for the trend, $0 \le \beta \le 1$.
- Reasonable starting values for level and slope are $\boldsymbol{\ell_1 = X_1}$ and $\mathbf{b_1 = X_2 - X_1}$

Evaluation of Forecasting Accuracy

- What makes a good forecast? Of course, a good forecast is an accurate forecast.
- A forecast "error" is the difference between an observed value and its forecast. The "error" does not mean a mistake, it means the unpredictable part of an observation.
- Error measure plays an important role in calibrating and refining forecasting model/method and helps the analyst to improve forecasting method.
- The choice of an error measure may vary according to the situation , number of time series available and on whether the task is to select the most accurate method or to calibrate a given model.
- The popular and highly recommended error measures are
  - Mean Square Error (MSE)
  - Root Mean Square Error (RMSE)
  - Mean Absolute Percentage Error (MAPE)

Mean Square Error (MSE)

MSE is defined as mean or average of the square of the difference between actual and estimated values. Mathematically it is represented as:

$$\text{MSE} = \frac{\sum_{j=1}^{N} (\text{observation } (j) - \text{prediction } (j))^2}{N}$$

Month	1	2	3	4	5	6	7	8	9	10	11	12
Actual Demand	42	45	49	55	57	60	62	58	54	50	44	40
Forecasted Demand	44	46	48	50	55	60	64	60	53	48	42	38
Error	-2	-1	1	5	2	0	-2	-2	1	2	2	2
Squared Error	4	1	1	25	4	0	4	4	1	4	4	4

Sum of Square Error = 56 and MSE = 56 / 12 = 4.6667

Root Mean Square Error (RMSE)

It is just the square root of the mean square error. Mathematically it is represented as:

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N} (\text{observation } (j) - \text{prediction } (j))^2}{N}}$$

Month	1	2	3	4	5	6	7	8	9	10	11	12
Actual Demand	42	45	49	55	57	60	62	58	54	50	44	40
Forecasted Demand	44	46	48	50	55	60	64	60	53	48	42	38
Error	-2	-1	1	5	2	0	-2	-2	1	2	2	2
Squared Error	4	1	1	25	4	0	4	4	1	4	4	4

Sum of Square Error = 56, MSE = 56 / 12 = 4.6667, RMSE = SQRT(4.667) = 2.2

Mean Absolute Percentage Error (MAPE)

The formula to calculate MAPE is as follows:

$$\text{MAPE} = (100/n) \times \sum_{i=1}^{n} \frac{(|X'(t) - X(t)|}{X(t)}$$

Here, X'(t) represents the forecasted data value of point t and X(t) represents the actual data value of point t. Calculate MAPE for the below dataset.

Month	1	2	3	4	5	6	7	8	9	10	11	12
Actual Demand	42	45	49	55	57	60	62	58	54	50	44	40
Forecasted Demand	44	46	48	50	55	60	64	60	53	48	42	38

- MAPE is commonly used because it's easy to interpret and easy to explain. For example, a MAPE value of 11.5% means that the average difference between the forecasted value and the actual value is 11.5%.
- The lower the value for MAPE, the better a model is able to forecast values e.g. a model with a MAPE of 2% is more accurate than a model with a MAPE of 10%.


THANK
YOU!