CSE3506-Essentials of Data Analytics

Dr. Vergin Raja Sarobin

School of Computer Science and Engineering
VIT Chennai
verginraja.m@vit.ac.in

Course Objectives

- ✔ To understand the concepts of analytics using various machine learning models.
- ✔ To appreciate supervised and unsupervised learning for predictive analysis
- ✔ To understand data analytics as the next wave for businesses looking for competitive advantage
- ✔ Validate the results of their analysis according to statistical guidelines
- ✔ Validate and review data accurately and identify anomalies
- ✔ To learn aspects of computational learning theory
- ✔ Apply statistical models to perform Regression Analysis, Clustering and Classification

Course Outcomes

- ✔ Identify and apply the appropriate supervised learning techniques to solve real world problems.
- ✔ Choose and implement typical unsupervised algorithms for different types of applications.
- ✔ Implement statistical analysis techniques for solving practical problems.
- ✔ Understand different techniques to optimize the learning algorithms.
- ✔ Aware of health and safety policies followed in organization, data and information management and knowledge & skill development.

Syllabus

Module-1: Regression Analysis

- ✔ Linear regression: simple linear regression - Regression Modelling - Correlation, ANOVA, Forecasting, Autocorrelation **(6 Hours)**

Syllabus

Module-2: Classification

- ✔ Logistic Regression, Decision Trees, Naïve Bayes-conditional probability - Random Forest - SVM Classifier **(6 Hours)**

Syllabus

Module-3: Clustering

- ✔ K-means, K-medoids, Hierarchical clustering **(4 Hours)**

VIT
Vellore Institute of Technology

Module-4: Optimization

- ✔ Gradient descent - Variants of gradient descent - Momentum - Adagrad - RMSprop - Adam – AMSGrad **(3 Hours)**

Syllabus

Module-5: Managing Health and Safety

- ✔ Comply with organization's current health, safety and security policies and procedures - Report any identified breaches in health, safety, and security policies and procedures to the designated person - Identify and correct any hazards that they can deal with safely, competently and within the limits of their authority - Report any hazards that they are not competent to deal with to the relevant person in line with organizational procedures and warn other people who may be affected. **(4 Hours)**

Syllabus

Module-6: Data and Information Management

- ✔ Establish and agree with appropriate people the data/information they need to provide, the formats in which they need to provide it, and when they need to provide it - Obtain the data/information from reliable sources - Check that the data/information is accurate, complete and up-to-date **(4 Hours)**

Syllabus

Module-7: Learning and Self Development

- ✔ Obtain advice and guidance from appropriate people to develop their knowledge, skills and competence - Identify accurately the knowledge and skills they need for their job role - Identify accurately their current level of knowledge, skills and competence and any learning and development needs - Agree with appropriate people a plan of learning and development activities to address their learning needs **(3 Hours)**

Syllabus

Text Book

- ✔ Cathy O'Neil and Rachel Schutt. "Doing Data Science, Straight talk from the Frontline", O'Reilly. 2014.
- ✔ Dan Toomey, "R for Data Science", Packt Publishing, 2014.
- ✔ Trevor Hastie, Robert Tibshirani and Jerome Friedman. "Elements of Statistical Learning", Springer , Second Edition. 2009.
- ✔ Kevin P. Murphy. "Machine Learning: A Probabilistic Perspective", MIT Press; 1st Edition, 2012.

Syllabus

Reference Books

- ✔ Glenn J. Myatt, "Making Sense of Data : A Practical Guide to Exploratory Data Analysis and Data Mining", John Wiley & Sons, Second Edition, 2014.
- ✔ G. K. Gupta, —Introduction to Data Mining with Case Studies", Easter Economy Edition, Prentice Hall of India, 2006.
- ✔ Michael Berthold, David J. Hand, "Intelligent Data Analysis", Springer, 2007.
- ✔ Colleen Mccue, "Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis", Elsevier, 2007.
- ✔ R N Prasad, Seema Acharya, "Fundamentals of Business Analytics", Wiley; Second edition, 2016.
- ✔ https://www.sscnasscom.com/qualification-pack/SSC/Q2101/

Assessment Process (Theory)

CAT-1	**15**
CAT-2	**15**
Assignments/Quizzes	**30**
FAT	**40**
Total	**100**

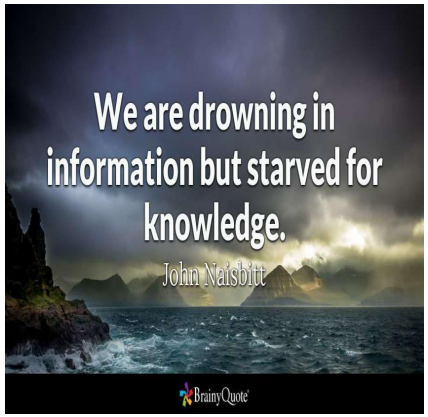
Module-1: Regression Analysis

Linear regression: simple linear regression - Regression Modelling - Correlation, ANOVA, Forecasting, Autocorrelation **(6 Hours)**

Data Analytics – What?

- **Science of analyzing raw data in order to make conclusions about that information** *[Investopedia]*
- **Analytics** is the systematic computational analysis of data or statistics *[Wikipedia]*
- **Data analysis** is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, deriving conclusions, and supporting decision-making *[Wikipedia]*

Data Analytics – Why?



- Helps business to optimize their performances
  - Reduce cost
  - Improved business
  - Make better decision

Data Analytics – Types

- **Descriptive analytics** - describes what has happened over a given period of time.
  - Have the number of views gone up?
  - Are sales stronger this month than last?
- **Diagnostic analytics** - focuses more on why something happened. This involves more diverse data inputs and a bit of hypothesizing.
  - Did the weather affect sales of a cool drink?
  - Did that latest marketing campaign impact sales?

Data Analytics – Types

- **Predictive analytics** - focuses on what is likely going to happen in the near term.
  - What happened to sales the last time we had a hot summer?
  - How many weather models predict a hot summer this year?
- **Prescriptive analytics** suggests a course of action.
  - If the likelihood of a hot summer is measured as an average of say five weather models is above 58%, we should add an evening shift to the workers to increase output.

What is Machine Learning?

- Large volume of data demands automated methods of data analysis which is what machine learning provides.
- Machine learning is defined as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty.

Machine Learning Paradigms

- Three Learning Paradigms
  - Predictive or Supervised Learning
  - Descriptive or Unsupervised Learning
  - Reinforcement Learning

Statistical Learning

Statistical learning refers to

- getting inference from a vast data set using

  Supervised learning models

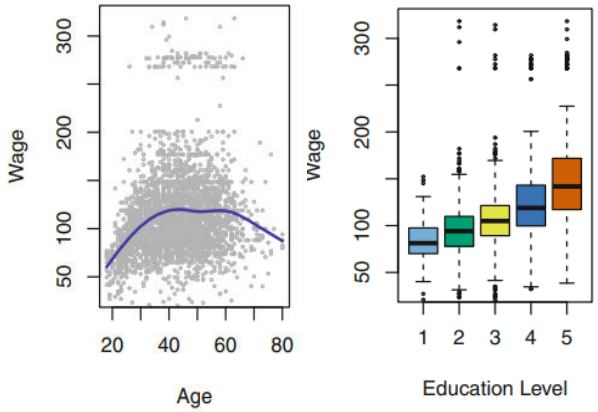
  or

  Unsupervised learning models

Statistical Learning

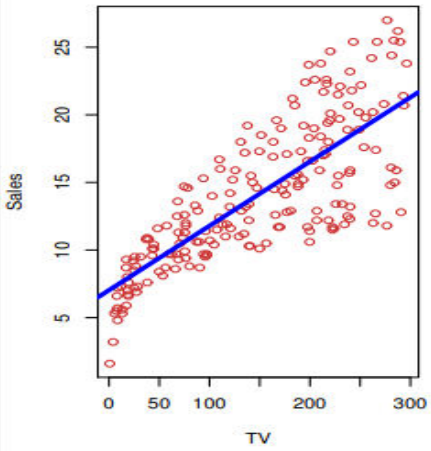
- **Supervised statistical learning:**
  - ✓ It is defined by its use of labelled datasets to train algorithms that to classify data or predict outcomes accurately.
  - ✓ **Examples**: Problems occur in business, medicine, astrophysics, and public policy
- **Unsupervised statistical learning:**
  - ✓ Unsupervised learning uses unlabelled data. From that data, it discovers patterns that help solve for clustering or we can learn relationships and structure from such data
  - ✓ **Example**: Input dataset containing images of different types of cats and dogs

Statistical Learning – Wage Data



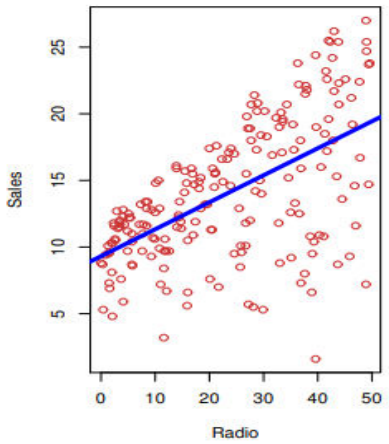
- The Wage data involves predicting a continuous or quantitative output value.
- This is often referred to as a **regression problem**

Statistical Learning – Advertising Data



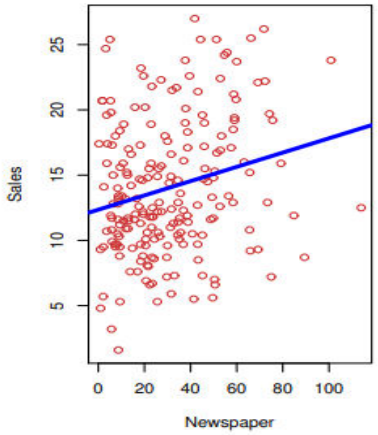
✓ The Advertising data set consists of the sales of a product in 200 different cities, along with advertising budgets for three different media: **TV**, radio, and newspaper

Statistical Learning – Advertising Data



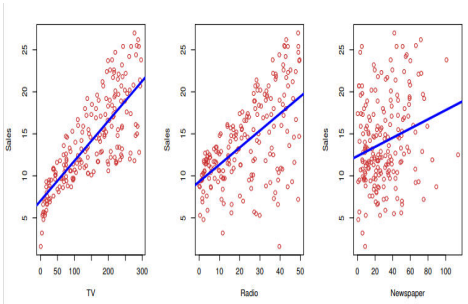
- ✓ The Advertising data set consists of the sales of a product in 200 different cities, along with advertising budgets for three different media: TV, **Radio**, and newspaper

Statistical Learning – Advertising Data



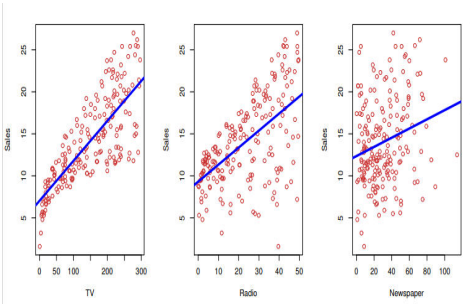
- ✓ The Advertising data set consists of the sales of a product in 200 different cities, along with advertising budgets for three different media: TV, Radio, and **newspaper**
- ✓ Goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets

Statistical Learning – Advertising Data



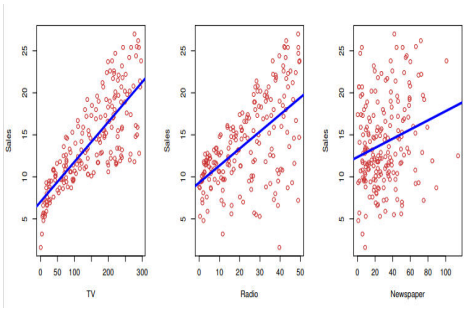
- Input Variables: **Advertising budgets**
- Input Variables are denoted by X
- X_1 – TV budget
- X_2 – Radio budget
- X_3 – Newspaper budget
- Input variables are called by different names like
  - **Predictors**
  - **Independent variables**
  - **Features**
  - **Variables**

Statistical Learning – Advertising Data



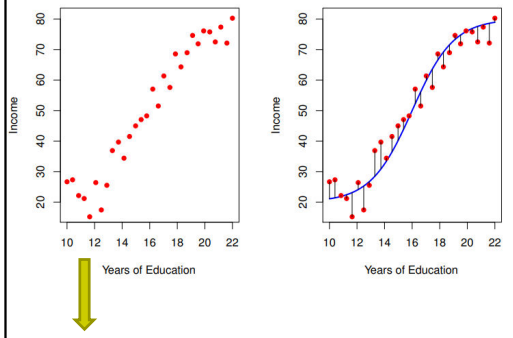
- Output Variable: **Sales**
- Output Variables are denoted by Y
- Output variables are called by different names like
  - **Responses,**
  - **Dependent variables**

Statistical Learning – Advertising Data



- ✓ There is some relationship between Y and X = $(X_1, X_2,...,X_p)$
- ✓ General form of relationship is
  - ✓ **$Y = f(X) + \varepsilon$**
- ✓ where
  - ✓ f is some fixed but unknown function of $X_1,...,X_p$
  - ✓ ε is a random error term, which is independent of X and has mean zero

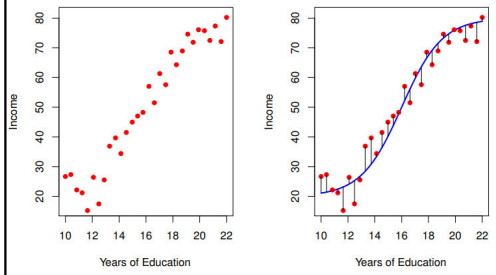
Statistical Learning – Income Data



Observed values of income and years of education for 30 individuals

- The black lines represent the error associated with each observation.
- Here some errors are positive (if an observation lies above the blue curve) and some are negative (if an observation lies below the curve)
- **Overall, these errors have approximately mean zero**

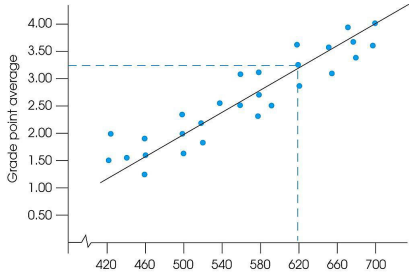
Statistical Learning – Income Data



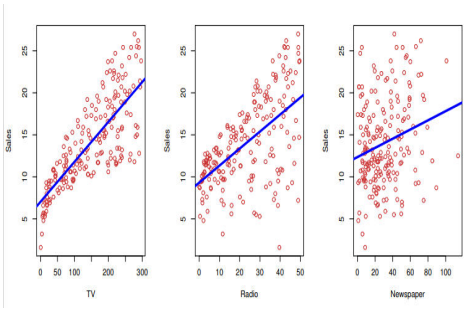
- ✓ Statistical Learning refers to a set of approaches for estimating f in the equation
  - **$Y = f(X) + \varepsilon$**
- ✓ Reasons to estimate 'f',:
  - ✓ **Prediction**
  - ✓ **Inference**

Linear Regression - Introduction

- **Linear Regression** is a statistical procedure that determines the equation for the straight line that best fits a specific set of data.
- This is a very simple approach for supervised learning
- In particular, it is a useful tool for predicting a quantitative response.



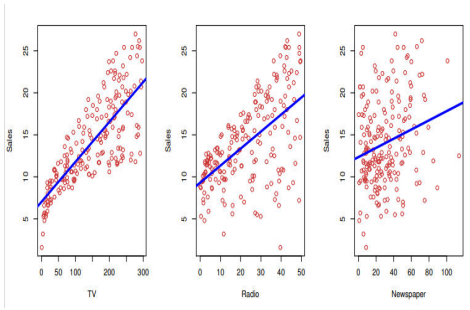
Advertising Data



On the basis of given advertising data,

- Marketing plan for next year can be made
- To develop the marketing plan, some information is required.
  - Is there a relationship between advertising budget and sales?
  - Is the relationship linear?
  - Predicting sales with a high level of accuracy requires a strong relationship.
  - If it is strong relationship then
  - In marketing, it is known as a **synergy effect**, while in statistics it is called **an interaction effect**

Advertising Data



The important questions are

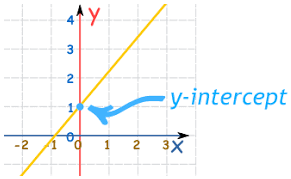
- ✓ **Which media contribute more to sales?**
- ✓ Do all three contribute to sales, or do just one or two.
- ✓ The individual effects of each medium on the money spent
- ✓ For every dollar spent on advertising in TV or Radio or Newspaper, by what amount will sales increase?
- ✓ How accurately can we predict this amount of increase?

Linear regression can be used to answer each of these questions

Linear Regression - Types

- **Types:**
  - ✓ Based on the number of independent variables, there are two types of linear regression
  - ✓ Simple Linear Regression
  - ✓ Multiple Linear Regression
- Mathematically, the linear relationship is approximately modeled as
  - $y = \beta_0 + \beta_1 x$

β0 - Intercept
β1 - Slope
β0 and β1 – Model coefficients



Simple Linear Regression

Estimating the coefficients β_0 and β_1

- Once we produce the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ using the training data, we can predict y given x:

$$\hat{y} \approx \hat{\beta}_0 + \hat{\beta}_1 x.$$

- Let $\hat{y}_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for i^{th} value of y based on the i^{th} value of x. Then

$$e_i = y_i - \hat{y}_i$$

represents the i^{th} residual. This is the difference between the i^{th} observed response value and the i^{th} predicted response value.

- The residual sum of squares (RSS) is defined as

$$\text{RSS} = \sum_{i}^{n} e_i^2 = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

where n is the number of predictions or simply, the number of samples in the training data.

Simple Linear Regression

Estimating the coefficients β_0 and β_1

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. Using some calculus, one can show that the minimizers are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

These $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares coefficient estimates for simple linear regression, and they give **the best linear fit** on the given training data.

Simple Linear Regression

Question-1:

Consider the following five training examples

X = [2 3 4 5 6]

Y = [12.8978 17.7586 23.3192 28.3129 32.1351]

We want to learn a function f(x) of the form f(x) = ax + b which is parameterized by (a, b). Using squared error as the loss function, which of the following parameters would you use to model this function.

(a) (4 3)

(b) (5 3)

(c) (5 1)

(d) (1 5)

Simple Linear Regression

Solution:

1. Calculate $Y_{predicted}$ for the given X using the given (a, b) values
2. For each (a, b) value, calculate the RSS
3. The best set of parameters is the one that gives minimum RSS

To calculate RSS, use the following formula

$$\text{RSS} = \sum_{i}^{n} e_i^2 = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

where $e_i = y_i - \hat{y}_i$

Simple Linear Regression

Solution:

For a = 4 and b = 3

Formula

$$\text{RSS} = \sum_{i}^{n} e_i^2 = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

X	Y	$Y_{predicted}$	$(Y-Y_{Predicted})^2$
2	12.8978	11	3.6016
3	17.7586	15	7.6099
4	23.3192	19	18.6555
5	28.3129	23	28.2269
6	32.1351	27	26.3693
		RSS	**84.4632**

Simple Linear Regression

Solution:

For a = 5 and b = 3

Formula

$$RSS = \sum_{i}^{n} e_i^2 = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

X	Y	$Y_{predicted}$	$(Y-Y_{Predicted})^2$
2	12.8978	13	0.0104
3	17.7586	18	0.0583
4	23.3192	23	0.1019
5	28.3129	28	0.0979
6	32.1351	33	0.7481
		RSS	**1.0166**

Simple Linear Regression

Solution:

For a = 5 and b = 1

Formula

$$RSS = \sum_{i}^{n} e_i^2 = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

X	Y	$Y_{predicted}$	$(Y-Y_{Predicted})^2$
2	12.8978	11	3.6016
3	17.7586	16	3.0927
4	23.3192	21	5.3787
5	28.3129	26	5.3495
6	32.1351	31	1.2885
		RSS	**18.7110**

Simple Linear Regression

Solution:

For a = 1 and b = 5

Formula

$$RSS = \sum_{i}^{n} e_i^2 = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

X	Y	$Y_{predicted}$	$(Y-Y_{Predicted})^2$
2	12.8978	7	34.7840
3	17.7586	8	95.2303
4	23.3192	9	205.0395
5	28.3129	10	335.3623
6	32.1351	11	446.6925
		RSS	**1117.1086**

Answer: The parameter (5,3) which gives least RSS (1.016). Hence (5,3) is used to model this function

Simple Linear Regression

VIT Vellore Institute of Technology

Question-2:

Consider the following five training examples

X = [2 3 4 5 6]

Y = [12.8978 17.7586 23.3192 28.3129 32.1351]

(a) Find the best linear fit

(b) Determine the minimum RSS

(c) Draw the residual plot for the best linear fit and comment on the suitability of the linear model to this training data.

Simple Linear Regression

Solution:

(a) To find the best fit, calculate the model coefficients using the formula

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

Simple Linear Regression

Solution:

	X	Y	$(X-X_{mean})$	$(Y-Y_{mean})$	$(X-X_{mean})(Y-Y_{mean})$	$(X-X_{mean})^2$
	2	12.8978	-2	-9.9869	19.9738	4
	3	17.7586	-1	-5.1261	5.1261	1
	4	23.3192	0	0.4345	0.0000	0
	5	28.3129	1	5.4282	5.4282	1
	6	32.1351	2	9.2504	18.5008	4
Sum	20	114.4236	0	0.0000	49.0289	10
Mean	4	22.88472				

The best linear fit is

Y = 4.9029X + 3.2732

Substituting in the formula $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$,

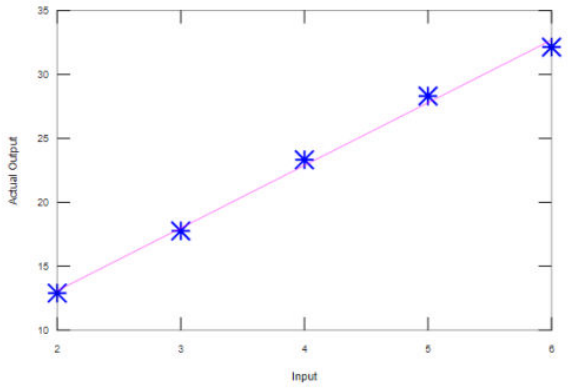
β_0 =3.2732

β_1 = 4.9029

Simple Linear Regression

Solution:

Best Linear Fit



Simple Linear Regression

Solution:

(b) To determine RSS

	X	Y	$(X-X_{mean})$	$(Y-Y_{mean})$	$(X-X_{mean})(Y-Y_{mean})$	$(X-X_{mean})^2$	$Y_{predicted}$	$(Y-Y_{Predicted})^2$
	2	12.8978	-2	-9.9869	19.9738	4	13.0789	0.0328
	3	17.7586	-1	-5.1261	5.1261	1	17.9818	0.0498
	4	23.3192	0	0.4345	0.0000	0	22.8847	0.1888
	5	28.3129	1	5.4282	5.4282	1	27.7876	0.2759
	6	32.1351	2	9.2504	18.5008	4	32.6905	0.3085
Sum	20	114.4236	0	0.0000	49.0289	10	RSS	0.8558
Mean	4	22.88472						

Y predicted is calculated using the best linear fit

Y = 4.9029 + 3.2732 X

RSS_{min} = 0.8558

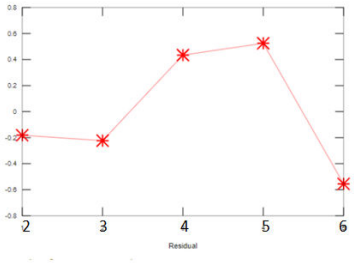
Simple Linear Regression

Solution:

(c) Residual plot for the best linear fit

X	Y	$Y_{predicted}$	Residual $(Y-Y_{Predicted})$
2	12.8978	13.0789	-0.1811
3	17.7586	17.9818	-0.2232
4	23.3192	22.8847	0.4345
5	28.3129	27.7876	0.5253
6	32.1351	32.6905	-0.5554

Residual Plot



The random pattern in it is an indication that a linear model is suitable for this data

Linear regression model validation

1. R^2 values are just one such measure.

Formula

R-squared, otherwise known as R^2 typically has a value in the range of 0 through to 1. **The closer the r-squared value is to 1, the better the fit**.

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_{i}^{n} e_i^2 = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

where $e_i = y_i - \hat{y}_i$

This is the difference between original and predicted sample

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

TSS = total sum of squares
n = number of observations
y_i = value in a sample
$\bar{y}$ = mean value of a sample

R2 squared is also known as Coefficient of Determination or sometimes also known as Goodness of fit

2. Residual Plot:

For regression, there are numerous methods to evaluate the goodness of your fit i.e. how well the model fits the data. One such method is residual plot.

A typical residual plot has the residual values on the Y-axis and the independent variable on the x-axis.

If the points are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

3. Mean Absolute Percentage Error (MAPE)

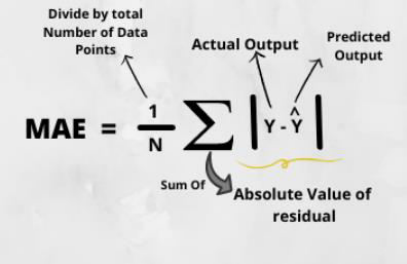
$$M = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

M = mean absolute percentage error
n = number of times the summation iteration happens
A_t = actual value
F_t = forecast value

What is a good value of MAPE?

The closer the MAPE value is to zero, the better the predictions.
A MAPE **less than 5%** is considered as an indication that the forecast is acceptably accurate.

4. Mean Absolute Error(MAE)



We aim to get a minimum MAE because this is a loss.

5. Mean Squared Error(MSE)

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)}^{2}$$

The square of the difference between actual and predicted

The lower the value the better and 0 means the model is perfect.

6. Root Mean Squared Error(RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)}$$

This produces a value **between 0 and 1**, where values closer to 0 represent better fitting models. Based on a rule of thumb, it can be said that RMSE values **between 0.2 and 0.5** shows that the model can relatively predict the data accurately.

7. Standard Error (SE) and Residual Standard Error (RSE)

- Assuming the errors ϵ_i for each observation are uncorrelated with common variance σ^2, the *standard errors* associated with $\hat{\beta}_0$ and $\hat{\beta}_1$ can be expressed as

$$\text{SE}(\hat{\beta}_0) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

and

$$\text{SE}(\hat{\beta}_1) = \sigma\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

- In general, $\sigma = \sqrt{\text{Var}(\epsilon)}$ is not known, but can be estimated from the data. This estimate is known as the *residual standard error* (RSE), and is expressed as

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}}.$$

8. Confidence Interval)

- For linear regression, the 95% confidence interval for β_0 approximately takes the form

$$\hat{\beta}_0 \pm 2\ \mathrm{SE}(\hat{\beta}_0).$$

- That is, there is approximately a 95 % chance that the interval

$$[\hat{\beta}_0 - 2\ \mathrm{SE}(\hat{\beta}_0)\ ,\ \hat{\beta}_0 + 2\ \mathrm{SE}(\hat{\beta}_0)]$$

will contain the true value of β_0

A confidence interval, in statistics, refers to the probability that a population parameter will fall between a set of values for a certain proportion of times. Analysts often use confidence intervals than contain either 95% or 99% of expected observations.

95% confidence interval

- Similarly, a confidence interval for β_1 approximately takes the form

$$[\hat{\beta}_1 - 2\ \mathrm{SE}(\hat{\beta}_1)\ ,\ \hat{\beta}_1 + 2\ \mathrm{SE}(\hat{\beta}_1)]$$

will contain the true value of β_1

- The word 'approximately' is included mainly because
  - ✓ The errors are assumed to be Gaussian and
  - ✓ The factor '2' in front of $\mathrm{SE}(\hat{\beta}_1)$ term will vary slightly depending on the number of observations 'n' in the linear regression

Example: Confidence Interval for Regression Coefficient in R

Suppose we'd like to fit a simple linear regression model using **hours studied** as a predictor variable and **exam score** as a response variable for 15 students in a particular class:

Hours Studied	Exam Score
1	64
2	66
4	76
5	73
5	74
6	81
6	83
7	82
8	80
10	88
11	84
11	82
12	91
12	93
14	89

We can use the lm() function to fit this simple linear regression model in R:

```
#create data frame
df <- data.frame(hours=c(1, 2, 4, 5, 5, 6, 6, 7, 8, 10, 11, 11, 12, 12, 14),
                 score=c(64, 66, 76, 73, 74, 81, 83, 82, 80, 88, 84, 82, 91, 93, 89))

#fit linear regression model
fit <- lm(score ~ hours, data=df)

#view model summary
summary(fit)
```

```
Call:
lm(formula = score ~ hours, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-5.140 -3.219 -1.193  2.816  5.772

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   65.334      2.106  31.023 1.41e-13 ***
hours          1.982      0.248   7.995 2.25e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.641 on 13 degrees of freedom
Multiple R-squared:  0.831,     Adjusted R-squared:  0.818
F-statistic: 63.91 on 1 and 13 DF,  p-value: 2.253e-06
```

We can use the **confint()** function to calculate a 95% confidence interval for the regression coefficient:

```
#calculate confidence interval for regression coefficient for 'hours'
confint(fit, 'hours', level=0.95)


          2.5 %   97.5 %
hours 1.446682 2.518068
```

Since this confidence interval doesn't contain the value 0, we can conclude that there is a statistically significant association between hours studied and exam score.

We can also confirm this is correct by calculating the 95% confidence interval for the regression coefficient by hand:

Alpha=0.05

- 95% C.I. for β_1: $b_1 \pm t_{1-\alpha/2,\ n-2} * se(b_1)$
- 95% C.I. for β_1: $1.982 \pm t_{.975,\ 15-2} * .248$
- 95% C.I. for β_1: $1.982 \pm 2.1604 * .248$
- 95% C.I. for β_1: [1.446, 2.518]

The 95% confidence interval for the regression coefficient is **[1.446, 2.518]**.

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Correlation

Covariance and correlation

Covariance and correlation are two mathematical concepts used in statistics.

Both terms are used to describe how two variables relate to each other.

Covariance

Covariance signifies the **direction of the linear relationship** between the two variables.

By direction we mean if the *variables* are directly proportional or inversely proportional to each other. (Increasing the value of one variable might have a positive or a negative impact on the value of the other variable).

The values of covariance can be any number between the two opposite infinities.

Covariance Formula

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

$cov_{x,y}$ = covariance between variable x and y

x_i = data value of x

y_i = data value of y

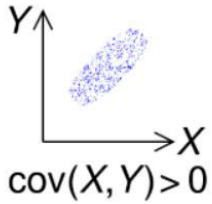
$\bar{x}$ = mean of x

$\bar{y}$ = mean of y

N = number of data values

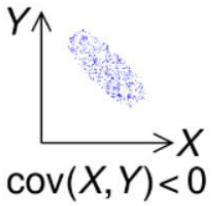
Types of covariance

Positive covariance
Positive covariance means both the variables (X, Y) move in the same direction (i.e. show similar behavior). So, if greater values of one variable (X) seem to correspond with greater values of another variable (Y), then the variables are considered to have positive covariance. This tells you something about the linear relationship between the two variables. So, for example, if an increase in a person's height corresponds with an increase in a person's weight, there is positive covariance between the two.



Types of covariance…

Negative covariance
Negative covariance means both the variables (X, Y) move in the opposite direction. As opposed to positive covariance, if the greater values of one variable (X) correspond to lesser values of another variable (Y) and vice-versa, then the variables are considered to have negative covariance.



Variables whose covariance is zero are called uncorrelated variables

Example:

X	Y
10	40
12	48
14	56
8	32

Step 1: Calculate Mean of X and Y

Mean of X (μx) : 10+12+14+8 / 4 = 11

Mean of Y(μy) = 40+48+56+32 = 44

Step 2:

$x_i - \bar{x}$	$y_i - \bar{y}$
10 − 11 = -1	40 − 44 = − 4
12 − 11 = 1	48 − 44 = 4
14 − 11 = 3	56 − 44 = 12
8 − 11 = -3	32 − 44 = 12

Step 3: Substitute the above values in the covariance formula

$$\text{Cov}(x,y) = \frac{(-1)(-4) + (1)(4) + (3)(12) + (-3)(12)}{3}$$

Cov(x,y)=2.67, It is a positive covariance

Correlation

What is the relationship between two variables?

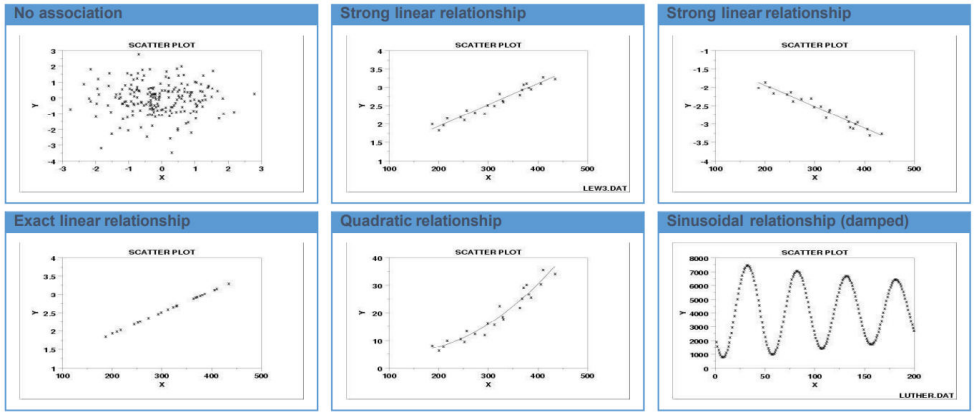
Relationship between hours studying (X) and grades on a midterm (Y)?

Relationship between self-esteem (X) and depression (Y)?

The relationship between two variables over a period, especially one that shows a close match between the variables' movements

Direction and strength of relationship between two variables

Graphical representation of data in a bivariate setup



Examples

- Increase in height results in weight increase for children
- Attending lessons leads to improved grades
- Age of the car impact its stopping distances
- More the years of education higher the income

Business Examples

- Rising unemployment leads to a decrease in sales of taste the difference products
- Increase in demand of a product leads to increase in supply
- More efficient the workers higher the productivity

Correlation Computation

Correlation between the two variables is obtained from normalizing the covariance by dividing it with the product of the standard deviations of the two variables.

$$Correlation = \frac{Cov\ (x, y)}{\sigma x * \sigma y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

Correlation…

Correlation value ranges from -1 to +1.

The closer it is to +1 or -1, the more closely the two variables are related.

If there is no relationship at all between two variables, then the correlation coefficient will certainly be 0.

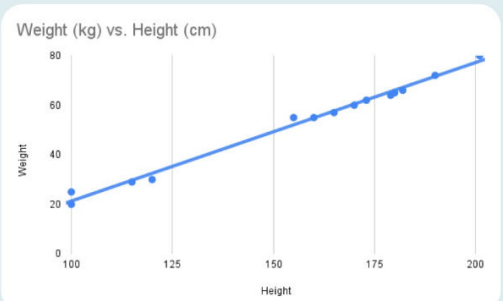
When the correlation coefficient is positive, an increase in one variable also increases the other.

When the correlation coefficient is negative, the changes in the two variables are in opposite directions.

Types of correlation

Positive correlation

Two variables are considered to have a positive correlation if they are directly proportional. That is, if the value of one variable increases, then the value of the other variable will also increase. A perfect positive correlation holds a value of "1". On a graph, positive correlation appears as follows:

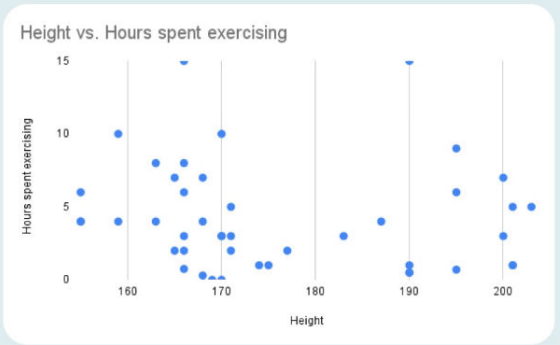


Negative correlation

A perfect negative correlation holds a value of "-1" which means that, as the value of one variable increases, the value of the second variable decreases (and vice versa). In graph form, this is how negative correlation might look:

Zero or no correlation

The value "0" denotes that there is no correlation. It indicates that there is no relationship between the two variables, so an increase or decrease in one variable is unrelated to an increase or decrease in the other variable. A graph showing zero correlation will follow a random distribution of data points, as opposed to a clear line:



What is a correlation matrix

A correlation matrix is essentially a table depicting the correlation coefficients for various variables. The rows and columns contain the value of the variables, and each cell shows the correlation coefficient.

	Hours spent exercising	Cardio fitness level	Height	Age
Hours spent exercising	1	0.82	0.03	-0.44
Cardio fitness level	0.82	1	0.2	-0.05
Height	0.03	0.2	1	0.1
Age	-0.44	-0.05	0.1	1

ANOVA- Analysis of Variance

Population

Sampling
- IoE inspection to get feedback from students/faculty/parent/Alumni/Industry
- Quality control (Statistical Quality Control)
  - 100% inspection
  - Sample inspection
- Conducting Experiments

Note:
There should not be significant variation between the sample mean and the population mean.
This is to be proved statistically.

Why ANOVA?
Helps us to understand how different sample groups respond.

ANOVA

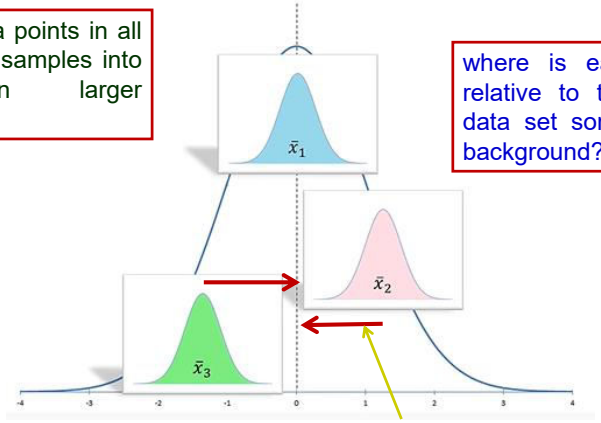
- **ANOVA – ANalysis of Variance**
- **Variance:**
- The variance measures the average degree to which each data point is different from the mean.
- The variance is greater when there is a wider range of numbers in the group.
- The calculation of variance uses squares because it weighs outliers more heavily than data point closer to the mean.
- This prevents differences above the mean from canceling out those below, which would result in a variance of zero.
- Thus variance is the **average of the squared differences from the mean**.
- **ANOVA** is a **hypothesis testing procedure** that is used to evaluate differences between 2 or more samples.

Standard Deviation:

- Standard Deviation tells how far the data points are from the mean.
- It is the square root of variance
- These two statistical concepts are closely related
- For Data analysts, these two mathematical concepts are of paramount importance as they are used to measure volatility of data distribution.
- In stock trading, if the standard deviation is less, it indicates the investment is less risky.

ANOVA

Put all the data points in all of the THREE samples into a common larger distribution

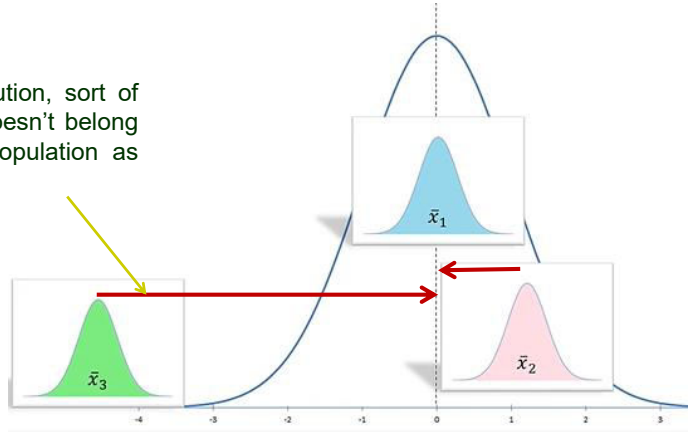


where is each mean relative to the overall data set sorted in the background?

Shows how far the mean it is away from the mean of the larger sort of combined population

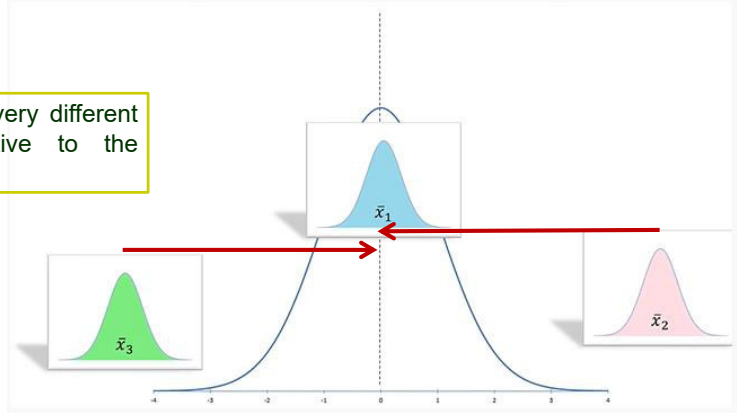
ANOVA

Oddball distribution, sort of the one that doesn't belong in the same population as the other two



ANOVA

Means are in very different locations relative to the overall mean



Step1:
Setting the hypothesis (Null hypothesis or alternate hypothesis)

- Null Hypothesis (H0: $\mu1=\mu2=\mu3$)
- Alternate Hypothesis (Ha: Alteast one difference among the means)
  And
- Fixing the confidence interval (90%, 95%)
  α=0.1 or 0.05

Step2: Find the df

- df between the groups/columns
- df within the groups/columns
- df_total

Step3:Calculating the Means

- Means for each group and
- Grand mean

Step4: All variability across the columns/groups

- SST
- SSC (Sum of Squares between/Columns)
- SSE(Sum of Squares within/Errors)

Step5: To calculate the variance between and within

- Mean Squares_between = $\frac{SS\_between}{df\_between}$
- Mean Squares_within = $\frac{SS\_within}{df\_within}$

Step 6: To perform F test (To calculate F_ratio)

- F_statistic = Mean Square_between / Mean Square_within
- F_critical from F distribution table (Corr to df_numerator and df_denominator)

of the *F* Distribution

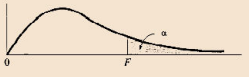


Table 1 $\alpha = 0.05$

Degrees of Freedom for Denominator	Degrees of Freedom for Numerator															
	1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50
1	161.4	199.5	215.8	224.8	230.0	233.8	236.5	238.6	240.1	242.1	245.2	248.4	248.9	250.5	250.8	252.6
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.44	19.46	19.47	19.48	19.48
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.63	8.62	8.59	8.58
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.77	5.75	5.72	5.70
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.52	4.50	4.46	4.44
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.83	3.81	3.77	3.75
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.40	3.38	3.34	3.32
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.11	3.08	3.04	3.02
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.89	2.86	2.83	2.80
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.73	2.70	2.66	2.64
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.60	2.57	2.53	2.51
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.50	2.47	2.43	2.40
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.41	2.38	2.34	2.31
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.34	2.31	2.27	2.24
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.28	2.25	2.20	2.18
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.23	2.19	2.15	2.12
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.18	2.15	2.10	2.08
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.14	2.11	2.06	2.04
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.11	2.07	2.03	2.00
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.07	2.04	1.99	1.97
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	2.02	1.98	1.94	1.91
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11	2.03	1.97	1.94	1.89	1.86
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07	1.99	1.94	1.90	1.85	1.82
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04	1.96	1.91	1.87	1.82	1.79
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.88	1.84	1.79	1.76
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.78	1.74	1.69	1.66

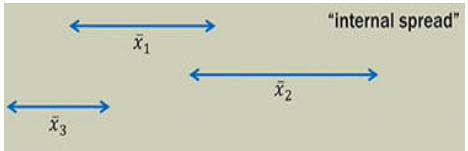
F_statistic < F_critical

ANOVA

ANOVA: Analysis of Variance is a *variability ratio*

Variability AMONG / BETWEEN the means.

Variability AROUND / WITHIN the distributions.



$$= \frac{Variance\ Between}{Variance\ Within}$$

ANOVA

ANOVA: Analysis of Variance is a *variability ratio*

$$\frac{Variance\ Between}{Variance\ Within}\ \Big\}\ Total\ Variance\ Components$$

$$Variance\ Between + Variance\ Within = Total\ Variance$$

"Partitioning" – separating total variance into its component parts

This is One way ANOVA/ Single Factor ANOVA

If the variability BETWEEN the means (distance from overall mean) in the numerator is relatively large compared to the variance WITHIN the samples (internal spread) in the denominator, the ratio will be much larger than 1. The samples then most likely do NOT come from a common population; REJECT NULL HYPOTHESIS that means are equal.

ANOVA

ANOVA: Analysis of Variance is a *variability ratio*

$\frac{LARGE}{small} = Reject\ H_0$ — At least one mean is an outlier and each distribution is narrow; distinct from each other

$\frac{Variance\ Between}{Variance\ Within}$

$\frac{similar}{similar} = Fail\ to\ Reject\ H_0$ — Means are fairly close to overall mean and/ or distributions overlap a bit; hard to distinguish

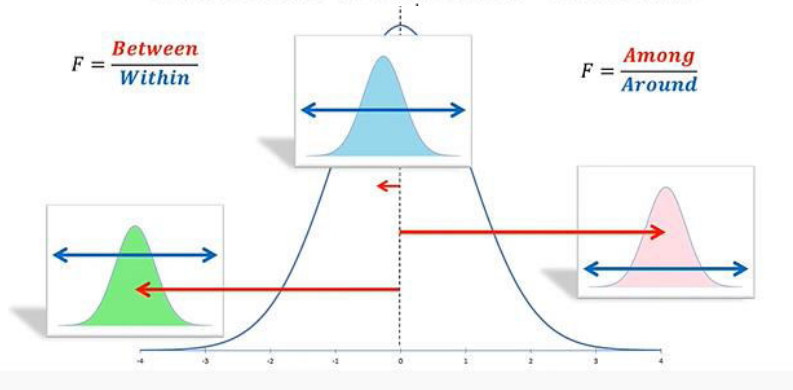
$\frac{small}{LARGE} = Fail\ to\ Reject\ H_0$ — The means are very close to overall mean and/ or distribution "melt" together

ANOVA

ANOVA: Analysis of Variance is a *variability ratio*

$$Variance\ Between + Variance\ Within = Total\ Variance$$

$F = \frac{Between}{Within}$ $F = \frac{Among}{Around}$



ANOVA

Question-4:

18 students (six each from first year to third year) were selected for an informal study about their understanding skill level. The evaluation was done for a score of 100. Using One-way ANOVA technique, find out whether or not a difference exists somewhere between the three different year levels

Scores		
First Year	Second Year	Third Year
82	62	64
93	85	73
61	94	87
74	78	91
69	71	56
53	66	78

ANOVA

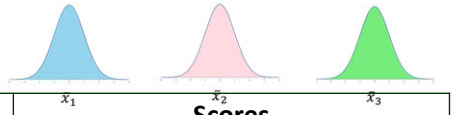
Groups/ Columns

Random Sample within each group

Scores		
First Year	Second Year	Third Year
82	62	64
93	85	73
61	94	87
74	78	91
69	71	56
53	66	78

ANOVA

Calculate the mean of each column



	Scores		
	First Year	Second Year	Third Year
	82	62	64
	93	85	73
	61	94	87
	74	78	91
	69	71	56
	53	66	78
Mean $\bar{x}$	**72**	**76**	**74.83**

Calculate Grand Mean/ Overall Mean $\bar{\bar{x}}$

The mean of all 18 scores is

$\bar{\bar{x}} = 74.28$

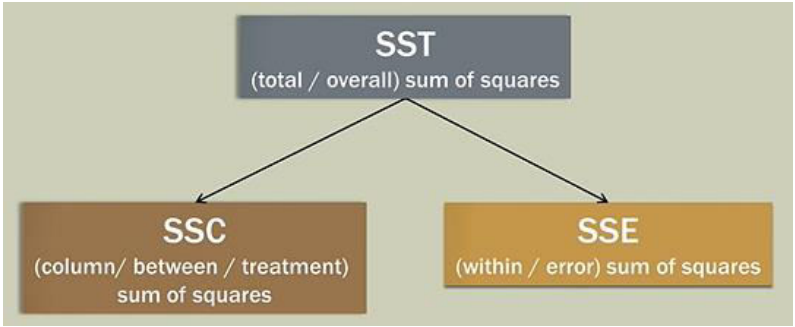
ANOVA

Sum of Squares (SS)

Sum of squares of the difference of the dependent variable and its mean

ANOVA

Partitioning Sum of Squares



ANOVA

	Scores		
	First Year	Second Year	Third Year
	82	62	64
	93	85	73
	61	94	87
	74	78	91
	69	71	56
	53	66	78
Mean $\bar{x}$	**72**	**76**	**74.83**

SST (total / overall) sum of squares

1. Find difference between each data point and the overall mean.
2. Square the difference.
3. Add them up

$\bar{\bar{x}} = 74.28$

ANOVA

	Scores					
	First Year	Second Year	Third Year	$(X_A-X_{mean})^2$	$(X_B-X_{mean})^2$	$(X_C-X_{mean})^2$
	82	62	64	59.633	150.744	105.633
	93	85	73	350.522	114.966	1.633
	61	94	87	176.299	388.966	161.855
	74	78	91	0.077	13.855	279.633
	69	71	56	27.855	10.744	334.077
	53	66	78	452.744	68.522	13.855
Sum	432	456	449	**1067.130**	**747.796**	**896.685**
Mean	72	76	74.83			

SST = 1067.130 + 747.796 + 896.685 = 2711.611

SST (total / overall) sum of squares

1. Find difference between each data point and the overall mean.
2. Square the difference.
3. Add them up

$\bar{\bar{x}} = 74.28$

ANOVA

	Scores		
	First Year	Second Year	Third Year
	82	62	64
	93	85	73
	61	94	87
	74	78	91
	69	71	56
	53	66	78
Mean $\bar{x}$	**72**	**76**	**74.83**

Sum of Squares_between

1. Find difference between each group mean and the overall mean
2. Square the deviations
3. Multiply with no. of values of each column
4. Add them up

$\bar{\bar{x}} = 74.28$

ANOVA

	Scores		
	First Year	Second Year	Third Year
	82	62	64
	93	85	73
	61	94	87
	74	78	91
	69	71	56
	53	66	78
Mean $\bar{x}$	**72**	**76**	**74.83**

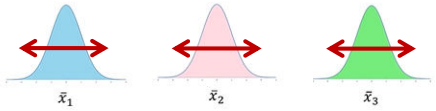
SSC = $6(72 - 74.28)^2 + 6(76 - 74.28)^2 + 6(74.83 - 74.28)^2 = 50.778$

Sum of Squares_between

1. Find difference between each group mean and the overall mean
2. Square the deviations
3. Multiply with no. of values of each column
4. Add them up

$\bar{\bar{x}} = 74.28$

ANOVA



	Scores		
	First Year	Second Year	Third Year
	82	62	64
	93	85	73
	61	94	87
	74	78	91
	69	71	56
	53	66	78
Mean $\bar{x}$	**72**	**76**	**74.83**

Sum of Squares_between

1. Find difference between each data point and its column mean.
2. Square each deviation.
3. Add them up the squared deviations.

ANOVA

	Scores					
	First Year	Second Year	Third Year	$(X_A - x_{a\_mean})^2$	$(X_B - x_{b\_mean})^2$	$(X_C - x_{c\_mean})^2$
	82	62	64	100	196	117.361
	93	85	73	441	81	3.361
	61	94	87	121	324	148.028
	74	78	91	4	4	261.361
	69	71	56	9	25	354.694
	53	66	78	361	100	10.028
Sum	432	456	449	**1036**	**730**	**894.833**
Mean	72	76	74.83			

SSE = 1036 + 730 + 894.833 = 2660.833

Sum of Squares_within

1. Find difference between each data point and its column mean.
2. Square each deviation.
3. Add them up the squared deviations.

ANOVA

Formulas for One-Way ANOVA

SSC	Sum of squares (columns/treatments)
SSE	Sum of squares (within/error)
SST	Sum of squares (total)

df = Degrees of Freedom

1. DoF b/w the columns

$$df_{columns} = C - 1$$

Mean Squares_between $= \frac{SS\_between}{df\_between}$

2. DoF within the columns

$$df_{error} = N - C$$

Mean Squares_within $= \frac{SS\_within}{df\_within}$

$$df_{total} = N - 1$$

ANOVA F - statistic

$$F = \frac{\text{Mean Squares\_between}}{\text{Mean Squares\_within}}$$

N = total observations

C = # columns/treatments

MSC = Mean Square Columns/ Treatments

MSE = Mean Square Error/ Within

ANOVA

Substituting the values

Mean Squares_between= $\frac{50.778}{3-1} = 25.389$

Mean Squares_within = $\frac{2660.833}{18-3} = 177.389$

$$F = \frac{MSC}{MSE} = \frac{25.389}{177.389} = 0.1431$$

Formula to calculate Critical Value in Excel:

F.INV.RT(ALPHA,NUMERATOR DOF, DENOMINATOR DOF)

- F-statistic value is less than $F_{critical}$
- Null hypothesis is accepted.
- **It means there is no significant difference in mean values**

Critical value of F: $F_{\alpha,\ dfc,\ dfe} = F_{0.05,\ 2,\ 15} = 3.68$

Forecasting

Time Series Forecasting

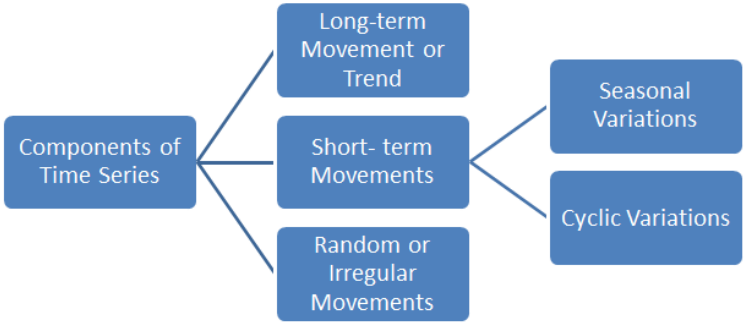
A time series is a sequence of observations recorded over a certain period of time. A simple example of time series is how we come across different temperature changes day by day or in a month.

Timeseries forecasting in simple words means to forecast or to predict the future value(eg-stock price) over a period of time.

Terminologies

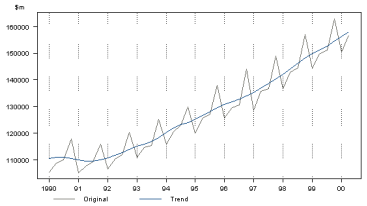
- Time series data
  - experimental data that have been observed at different points in time
- Examples:
  - daily stock market quotations
  - monthly unemployment figures
  - No. of COVID-19 cases observed over a period of time
  - BP measured over time

Components of Time Series



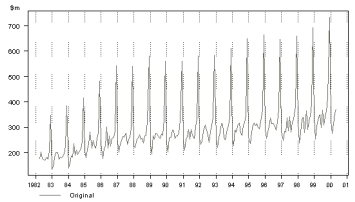
Components of Time Series

Trend



Direction in which something is increasing or decreasing

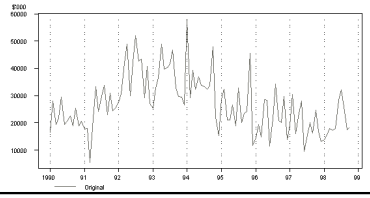
Seasonal



Repetition of peak or dip at regular intervals

Random

Irregular fluctuations – uncontrolled situations contributing to changes in values



Factors associated with time-series forecasting

- **Amount of data**
- **Data quality**
- **Seasonality**
- **Trends**
- **Unexpected events**

The **amount of data** is probably the most important factor (assuming that the data is accurate). A good rule of thumb would be *the more data we have, the better our model will generate forecasts*.

Data quality entails some basic requirements, such as having no duplicates, a standardized data format, and for the data to be collected consistently or at regular intervals.

Seasonality means that there are distinct periods of time when the data contains consistent irregularities. For example, if an online web shop analyzed its sales history, it would be evident that the holiday season results in an increased amount of sales.

Trends are probably the most important information you are looking for. They indicate whether a variable in the time series will increase or decrease in a given period.

Unexpected events (sometimes also referred to as noise or irregularities) can always occur, and we need to consider that when creating a prediction model. They present noise in historical data, and they are also not predictable.

Time Series Forecast Methods

1. The Average as a Forecast

The simplest form of an average as a forecast can be represented by the following formula:

$$\text{Forecast}_{t+1} = \text{Average Sales}_{1 \text{ to } t} = \sum_{t=1}^{N} S_t / N$$

where: S = Sales
N = Number of Periods of Sales Data (t)

In other words, our forecast for next month (or any month in the future, for that matter) is the average of all sales that have occurred in the past

2. Autoregression

In a multiple regression model, we forecast the variable of interest using a linear combination of predictors.

y= b0 + b1*X1

This technique can be used on time series where we can predict the value for the next time step (t+1) given the observations at the last two time steps (t and t-1).

As a regression model, this would look as follows,

X(t+1) = b0 + b1*X(t) + b2*X(t-1)

Because the regression model uses data from the same input variable at previous time steps, it is referred to as an autoregression (regression of self).

3. Simple Moving Average

Rather than use all the previous data in the calculation of an average as the forecast, why not just use some of the more recent data? This is precisely what a moving average does, with the following formula.

$$F_{t+1} = (S_t + S_{t-1} + S_{t-2} + \cdots + S_{t-N-1})/N$$

where: F_{t+1} = Forecast for Period t + 1
S_{t-1} = Sales for Period t – 1
N = Number of Periods in the Moving Average

So a three-period moving average would be:

$$F_{t+1} = (S_t + S_{t-1} + S_{t-2})/3$$

a four-period moving average would be:

$$F_{t+1} = (S_t + S_{t-1} + S_{t-2} + S_{t-3})/4$$

a five-period moving average would be:

$$F_{t+1} = (S_t + S_{t-1} + S_{t-2} + S_{t-3} + S_{t-4})/5$$

Simple Moving Average

Example 1: 3 year Simple Moving Average forecast

year	1	2	3	4	5	6	7	8	9	10	11	12
Sales	5.2	4.9	5.5	4.9	5.2	5.7	5.4	5.8	5.9	6	5.2	4.8

Calculate 3 year Simple Moving Average forecast

Solution:
Calculation of 3 year moving averages of the data

Year	Sales	3 year MA	
1	5.2		
2	4.9		
3	5.5		
4	4.9	(5.2+4.9+5.5)/3=5.2	
5	5.2	(4.9+5.5+4.9)/3=5.1	
6	5.7	(5.5+4.9+5.2)/3=5.2	
7	5.4	(4.9+5.2+5.7)/3=5.267	
8	5.8	(5.2+5.7+5.4)/3=5.433	
9	5.9	(5.7+5.4+5.8)/3=5.633	
10	6	(5.4+5.8+5.9)/3=5.7	
11	5.2	(5.8+5.9+6)/3=5.9	
12	4.8	(5.9+6+5.2)/3=5.7	

(1) year	(2) Sales	(3) 3 year moving average	(4) Error	(5) |Error|	(6) $Error^2$	(7) | % Error|
1	5.2					
2	4.9					
3	5.5					
4	4.9	5.2	4.9 - 5.2 = - 0.3	0.3	0.09	6.12 %
5	5.2	5.1	5.2 - 5.1 = 0.1	0.1	0.01	1.92 %
6	5.7	5.2	5.7 - 5.2 = 0.5	0.5	0.25	8.77 %
7	5.4	5.2667	5.4 - 5.2667 = 0.1333	0.1333	0.0178	2.47 %
8	5.8	5.4333	5.8 - 5.4333 = 0.3667	0.3667	0.1344	6.32 %
9	5.9	5.6333	5.9 - 5.6333 = 0.2667	0.2667	0.0711	4.52 %
10	6	5.7	6 - 5.7 = 0.3	0.3	0.09	5 %
11	5.2	5.9	5.2 - 5.9 = - 0.7	0.7	0.49	13.46 %
12	4.8	5.7	4.8 - 5.7 = - 0.9	0.9	0.81	18.75 %
13		5.3333	Total	3.5667	1.9633	67.34 %

(7) Calculation: |((Actual-forecast)/actual)|%= |(4.9-5.2)/4.9|%=6.12%

Forecasting errors

$$\text{MAE} = \frac{1}{n}\sum\left|e_i\right| = \frac{3.5667}{9} = 0.3963$$

2. Mean squared error (MSE)

$$\text{MSE} = \frac{1}{n}\sum\left|e_i^2\right| = \frac{1.9633}{9} = 0.2181$$

3. Root mean squared error (RMSE)

$$\text{RMSE} = \sqrt{MSE} = \sqrt{0.2181} = 0.4671$$

4. Mean absolute percentage error (MAPE)

$$\text{MAPE} = \frac{1}{n}\sum\left|\frac{e_i}{y_i}\right| = \frac{67.34}{9} = 7.48$$

Example 2:
Calculate a four-year moving average from the following data set:

Year	Sales ($M)
2003	4
2004	6
2005	5
2006	8
2007	9
2008	5
2009	4
2010	3
2011	7
2012	8

4. Weighted Moving Average forecast example

Example 1: 3 year Weighted Moving Average forecast

year	1	2	3	4	5	6	7	8	9	10	11	12
Sales	5.2	4.9	5.5	4.9	5.2	5.7	5.4	5.8	5.9	6	5.2	4.8

Calculate 3 year Weighted Moving Average forecast with weight=1,2,1

Solution:

The weights of the 3 years are respectively 1,2,1 and their sum is 4.

Calculation of 3 year moving averages of the data

Year	Sales	3 year MA
1	5.2	
2	4.9	
3	5.5	
4	4.9	(1*5.2+2*4.9+1*5.5)/4=5.125
5	5.2	(1*4.9+2*5.5+1*4.9)/4=5.2
6	5.7	(1*5.5+2*4.9+1*5.2)/4=5.125
7	5.4	(1*4.9+2*5.2+1*5.7)/4=5.25
8	5.8	(1*5.2+2*5.7+1*5.4)/4=5.5
9	5.9	(1*5.7+2*5.4+1*5.8)/4=5.575
10	6	(1*5.4+2*5.8+1*5.9)/4=5.725
11	5.2	(1*5.8+2*5.9+1*6)/4=5.9
12	4.8	(1*5.9+2*6+1*5.2)/4=5.775

(1) year	(2) Sales	(3) 3 year weighted moving average	(4) Error	(5) |Error|	(6) $Error^2$	(7) | % Error|
1	5.2					
2	4.9					
3	5.5					
4	4.9	5.125	4.9 - 5.125 = - 0.225	0.225	0.0506	4.59 %
5	5.2	5.2	5.2 - 5.2 = 0	0	0	0 %
6	5.7	5.125	5.7 - 5.125 = 0.575	0.575	0.3306	10.09 %
7	5.4	5.25	5.4 - 5.25 = 0.15	0.15	0.0225	2.78 %
8	5.8	5.5	5.8 - 5.5 = 0.3	0.3	0.09	5.17 %
9	5.9	5.575	5.9 - 5.575 = 0.325	0.325	0.1056	5.51 %
10	6	5.725	6 - 5.725 = 0.275	0.275	0.0756	4.58 %
11	5.2	5.9	5.2 - 5.9 = - 0.7	0.7	0.49	13.46 %
12	4.8	5.775	4.8 - 5.775 = - 0.975	0.975	0.9506	20.31 %
13		5.3	Total	3.525	2.1156	66.5 %

Forecasting errors

1. Mean absolute error (MAE), also called mean absolute deviation (MAD)

$$\text{MAE} = \frac{1}{n}\sum\left|e_i\right| = \frac{3.525}{9} = 0.3917$$

2. Mean squared error (MSE)

$$\text{MSE} = \frac{1}{n}\sum\left|e_i^2\right| = \frac{2.1156}{9} = 0.2351$$

3. Root mean squared error (RMSE)

$$\text{RMSE} = \sqrt{MSE} = \sqrt{0.2351} = 0.4848$$

4. Mean absolute percentage error (MAPE)

$$\text{MAPE} = \frac{1}{n}\sum\left|\frac{e_i}{y_i}\right| = \frac{66.5}{9} = 7.39$$

5. Exponential Smoothing forecast

Example: 3 year Single Exponential Smoothing forecast

year	1	2	3	4	5	6	7	8	9	10
Sales	30	25	35	25	20	30	35	40	30	45

Calculate 3 year Single Exponential Smoothing forecast

Solution:

(1) year	(2) Sales	(3) Exponential Smoothing ($\alpha = 0.3$)
1	30	30
2	25	$0.3 \cdot 30 + 0.7 \cdot 30 = 30$
3	35	$0.3 \cdot 25 + 0.7 \cdot 30 = 28.5$
4	25	$0.3 \cdot 35 + 0.7 \cdot 28.5 = 30.45$
5	20	$0.3 \cdot 25 + 0.7 \cdot 30.45 = 28.815$
6	30	$0.3 \cdot 20 + 0.7 \cdot 28.815 = 26.1705$
7	35	$0.3 \cdot 30 + 0.7 \cdot 26.1705 = 27.3193$
8	40	$0.3 \cdot 35 + 0.7 \cdot 27.3193 = 29.6235$
9	30	$0.3 \cdot 40 + 0.7 \cdot 29.6235 = 32.7365$
10	45	$0.3 \cdot 30 + 0.7 \cdot 32.7365 = 31.9155$
11		$0.3 \cdot 45 + 0.7 \cdot 31.9155 = 35.8409$

(1) year	(2) Sales	(3) Exponential Smoothing	(4) Error	(5) |Error|	(6) Error^2	(7) | % Error|
1	30	30				
2	25	30				
3	35	28.5				
4	25	30.45	25 - 30.45 = - 5.45	5.45	29.7025	21.8 %
5	20	28.815	20 - 28.815 = - 8.815	8.815	77.7042	44.07 %
6	30	26.1705	30 - 26.1705 = 3.8295	3.8295	14.6651	12.77 %
7	35	27.3193	35 - 27.3193 = 7.6807	7.6807	58.9924	21.94 %
8	40	29.6235	40 - 29.6235 = 10.3765	10.3765	107.6708	25.94 %
9	30	32.7365	30 - 32.7365 = - 2.7365	2.7365	7.4883	9.12 %
10	45	31.9155	45 - 31.9155 = 13.0845	13.0845	171.2032	29.08 %
11		35.8409	Total	51.9725	467.4265	164.72 %

Forecasting errors

1. Mean absolute error (MAE), also called mean absolute deviation (MAD)

$$\text{MAE} = \frac{1}{n}\sum\left|e_i\right| = \frac{51.9725}{7} = 7.4246$$

2. Mean squared error (MSE)

$$\text{MSE} = \frac{1}{n}\sum\left|e_i^2\right| = \frac{467.4265}{7} = 66.7752$$

3. Root mean squared error (RMSE)

$$\text{RMSE} = \sqrt{MSE} = \sqrt{66.7752} = 8.1716$$

4. Mean absolute percentage error (MAPE)

$$\text{MAPE} = \frac{1}{n}\sum\left|\frac{e_i}{y_i}\right| = \frac{164.72}{7} = 23.53$$

6. ARIMA Models

AutoRegressive Integrated Moving Average, or ARIMA, is a forecasting method that combines both an autoregressive model and a moving average model.

Autoregression uses observations from previous time steps to predict future values using a regression equation.
An autoregressive model utilizes a linear combination of past variable values to make forecasts:
Thus, an autoregressive model of order p can be written as:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

Likewise a pure **Moving Average (MA only) model** is one where Yt depends only on the lagged forecast errors.

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

If we combine differencing with autoregression and a moving average model, we obtain a non-seasonal ARIMA model. The full model can be represented with the following equation:

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Autocorrelation

Autocorrelation analysis is an important step in the Exploratory Data Analysis of time series forecasting.
The autocorrelation analysis helps detect patterns and check for randomness.
Autocorrelation is the correlation between a time series with a lagged version of itself.

Any autocorrelation that may be present in time series data is determined using a correlogram, also known as an ACF plot. This is used to help you determine whether your series of numbers is exhibiting autocorrelation at all, at which point you can then begin to better understand the pattern that the values in the series may be predicting.

An **autoregressive model** is when a value from a time series is regressed on previous values from that same time series.

for example, y_t on y_{t-1}:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t.$$

Autocorrelation

The coefficient of correlation between two values in a time series is called the **autocorrelation function** (**ACF**) For example the ACF for a time series yt is given by:

$$\text{Corr}(y_t, y_{t-k}), k = 1, 2, \ldots.$$

The ACF is a way to measure the linear relationship between an observation at time *t* and the observations at previous times.

The key statistics in time series analysis is the autocorrelation coefficient (or the correlation of the time series with itself, lagged by 1, 2, or more periods), which is given by the following formula,

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-K} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

ACF Example:

Y	Y_{t-1}	$Y_t - \bar{Y}$	$Y_{t-1} - \bar{Y}$	$(Y_t - \bar{Y})(Y_{t-1} - \bar{Y})$	$(Y_t - \bar{Y})(Y_t - \bar{Y})$
2		-4			16
3	2	-3	-4	12	9
5	3	-1	-3	3	1
7	5	1	-1	-1	1
9	7	3	1	3	9
10	9	4	3	12	16
Total				29	52

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-K} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} = \frac{29}{52} = 0.557692$$

CSE3506-Essentials of Data Analytics

Dr. Vergin Raja Sarobin

School of Computer Science and Engineering
VIT Chennai
verginraja.m@vit.ac.in

Module 2 : Classification

- Logistic Regression
- Decision Tree
- Naïve Bayes
- Conditional Probability
- Random Forest
- SVM Classifier

Logistic Regression

Introduction

- Prediction / Forecast
- The predicted data can be
  - Continuous
  - Discrete
- If it is continuous data we use regression
- If the predicted output is discrete the classification is used.

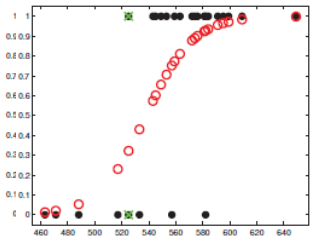
A classification Problem

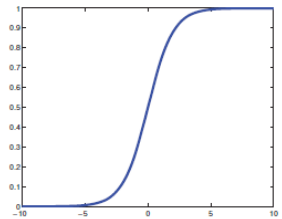
- VITEE score vs. admission in VIT
  - Admitted (1)
  - Not admitted (0)

Binary classification

- $y \in \{0,1\}$

Can it be solved by regression technique?





Sigmoidal function or squashing function of logistic/logit function

In linear regression,
$\hat{y} = H_\theta(x)$ can be <0 or >1

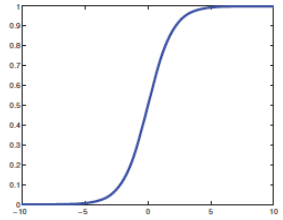
But in logistic classification we want,
$\hat{y} = 0$ or 1

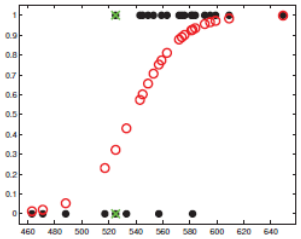
In linear regression,
$\hat{y} = H_\theta(x) = \theta^T x$

In logistic regression,
$\hat{y} = H_\theta(x) = \mathrm{f}(\theta^T x)$

$where, f(a) = \frac{1}{1+e^{-a}}$

Decision Boundary





Prediction:

$y = 1 \ \ if \ H_\theta(x) \geq 0.5$

$y = 0 \ \ if \ H_\theta(x) < 0.5$

Logistic Regression

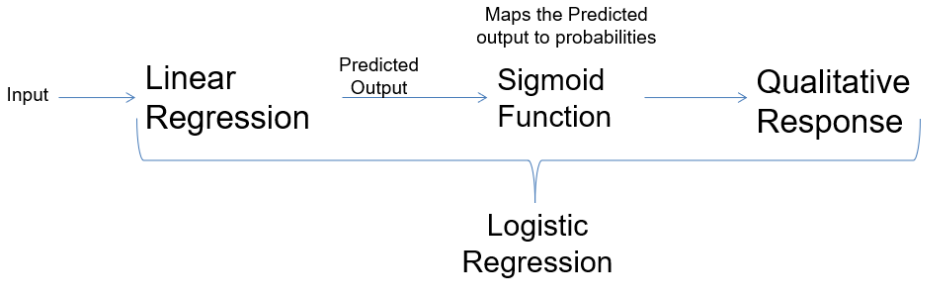
- The linear regression model assumes that the response (dependent) variable Y is **quantitative**. But in many situations, the response variable is instead **qualitative or categorical**
- **Examples:**
  - ✓ Eye color is qualitative, (eg black, brown, blue or green)
  - ✓ Student receives or does not receives scholarship
  - ✓ After an emergency surgery patient may be alive or dead
  - ✓ Your mobile phone signal is available or no coverage

Logistic Regression

- In general, there are two outcomes of the response variable "success" or "failure" and represent them by 1 (for a success) and 0 (for a failure).
- The data has at least one explanatory variable x and the probability p depends on the value of x.
- For example, suppose that we are studying whether a student applicant receives scholarship (y = 1) or not (y = 0). Here, p is the probability that an applicant receives aid, and possible explanatory variables include
  - (a) the financial support of the parents,
  - (b) the income and savings of the applicant, and
  - (c) whether the applicant has received financial aid before.

Logistic Regression

- For example, suppose that we are studying whether a student applicant receives scholarship (y = 1) or not (y = 0). Here, p is the probability that an applicant receives aid, and possible explanatory variables include
  - (a) the financial support of the parents,
  - (b) the income and savings of the applicant, and
  - (c) whether the applicant has received financial aid before.



Odds:

- Logistic regressions work with odds.
- The odds are simply the ratio of the proportions for the two possible outcomes.
- If $\hat{p}$ is the proportion for one outcome, then $(1 - \widehat{p})$ is the proportion for the second outcome:

$$\text{odds} = \frac{\hat{p}}{1 - \hat{p}}$$

$$odds = \frac{P(occurring)}{P(not\ occurring)}$$

Odds - Example:

- For the customer service data, the proportion of customers who would recommend the service in the sample of customers is $\hat{p} = 0.84$. Find the odds of recommending the service department

$$1 - \hat{p} = 1 - 0.84 = 0.16$$

$$\text{odds} = \frac{\hat{p}}{1 - \hat{p}} = \frac{0.84}{0.16} = 5.25$$

- ✓ If we round 5.25 to 5 = 5 / 1, we would say that the odds are approximately 5 to 1 that a customer would recommend the service to a friend.
- ✓ In a similar way, we could describe the odds that a customer would not recommend the service as 1 to 5.

Estimated Regression Equation

The natural logarithm of the odds ratio is equivalent to a *linear* function of the independent variables. The antilog of the logit function allows us to find the estimated regression equation.

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 \quad \text{antilog} \quad \frac{p}{1-p} = e^{\beta_0+\beta_1 x_1}$$

$$p = e^{\beta_0+\beta_1 x_1}(1-p)$$

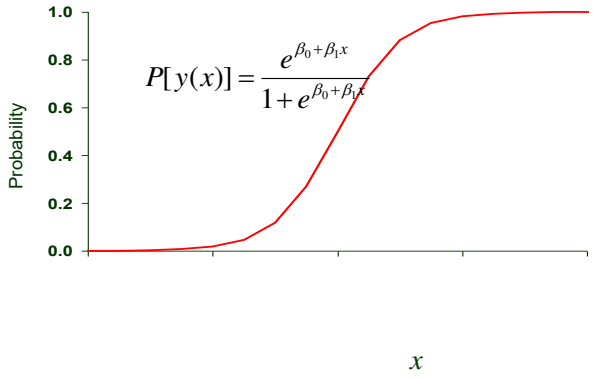
$$p = e^{\beta_0+\beta_1 x_1} - e^{\beta_0+\beta_1 x_1} * p$$

$$p + e^{\beta_0+\beta_1 x_1} * p = e^{\beta_0+\beta_1 x_1}$$

$$p(1 + e^{\beta_0+\beta_1 x_1}) = e^{\beta_0+\beta_1 x_1}$$

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1}}{1 + e^{\beta_0+\beta_1 x_1}}$$ Estimated Regression Equation

Logistic function (1)



Estimated Regression Equation

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}} \qquad \frac{1}{1+e^{-y}}$$

- To fit the above model (i.e. to determine β_0 and β_1), a method called *maximum likelihood* is used.
- The estimates β_0 and β_1 are chosen to maximize the *likelihood* function:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \times \prod_{i':y_{i'}=1} (1 - p(x_{i'})).$$

Example:

Consider the following training examples:

Marks scored: X = [81 42 61 59 78 49]

Grade (Pass/Fail): Y = [Pass Fail Pass Fail Pass Fail]

Assume we want to model the probability of Y of the form

which is parameterized by (β_0, β_1).

(i) Which of the following parameters would you use to model p(x).

(a) (-119, 2) (b) (-120, 2) (c) (-121, 2)

(ii) With the chosen parameters, what should be the minimum mark to ensure the student gets a 'Pass' grade with 95% probability?

Solution:

$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}$

X	β_0	β_1	p(x)	Grade Y	
81	-119	2	1.0000	Pass	1.0000
42	-119	2	0.0000	Fail	1.0000 (1-p)
61	-119	2	0.9526	Pass	0.9526
59	-119	2	0.2689	Fail	0.7311 (1-p)
78	-119	2	1.0000	Pass	1.0000
49	-119	2	0.0000	Fail	1.0000 (1-p)
				Likelihood value	**0.6964**

Likelihood Function:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \times \prod_{i':y_{i'}=1} (1 - p(x_{i'})).$$

Solution:

$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}$

X	β_0	β_1	p(x)	Grade Y	
81	-120	2	1.0000	Pass	1.0000
42	-120	2	0.0000	Fail	1.0000
61	-120	2	0.8808	Pass	0.8808
59	-120	2	0.1192	Fail	0.8808
78	-120	2	1.0000	Pass	1.0000
49	-120	2	0.0000	Fail	1.0000
				Likelihood value	**0.7758**

Likelihood Function:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \times \prod_{i':y_{i'}=1} (1 - p(x_{i'})).$$

Solution:

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}$$

X	β_0	β_1	p(x)	Grade Y	
81	-121	2	1.0000	Pass	1.0000
42	-121	2	0.0000	Fail	1.0000
61	-121	2	0.7311	Pass	0.7311
59	-121	2	0.0474	Fail	0.9526
78	-121	2	1.0000	Pass	1.0000
49	-121	2	0.0000	Fail	1.0000
				Likelihood value	**0.6964**

Likelihood Function:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \times \prod_{i':y_{i'}=1} (1 - p(x_{i'})).$$

- Among three, the maximum likelihood value is for β_0 = -120 , β_1 = 2.
- Therefore, we have to use these values to model p(x)
- With the chosen parameters, what should be the minimum mark to ensure the student gets a 'Pass' grade with 95% probability?

- Substituting p(x) = 0.95, β_0 = -120 and β_1 = 2, we will get

$$x_{min} = 61.47$$

Problem:

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}$$

Consider the following training examples:

Marks scored: X = [75 40 64 53 82 45]

Grade (Pass/Fail): Y = [Pass Fail Pass Fail Pass Fail]

Assume we want to model the probability of Y of the form

which is parameterized by (β_0, β_1).

(i) Which of the following parameters would you use to model p(x).

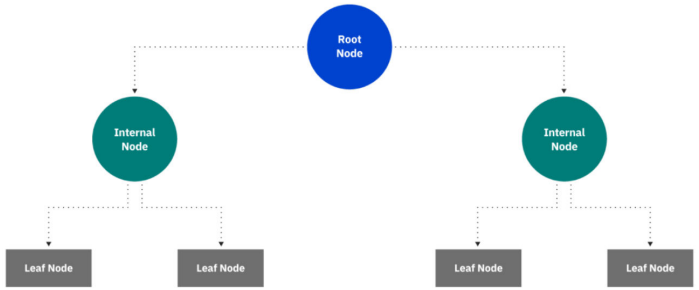
(a) (-119, 2) (b) (-120, 2) (c) (-121, 2)

(ii) With the chosen parameters, what should be the minimum mark to ensure the student gets a 'Pass' grade with 95% probability?

Decision Tree

Decision Trees

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.



Decision Trees

A tree can be *"learned"* by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called *recursive partitioning*. The recursion is completed with leaf nodes.

As you can see from the diagram in the previous page, a decision tree starts with a root node, which does not have any incoming branches. The outgoing branches from the root node then feed into the internal nodes, also known as decision nodes. The leaf nodes represent all the possible outcomes within the dataset.

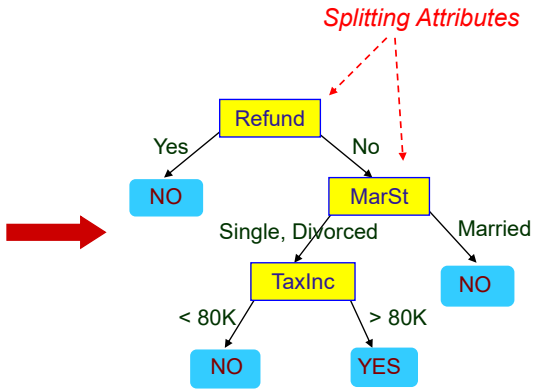
Decision trees used in data analytics are of two main types –

- **Classification tree** – when the response is a nominal variable, for example if an email is spam or not.
- **Regression tree** – when the predicted outcome can be considered a real number (e.g. the salary of a worker).

Example of a Decision Tree

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

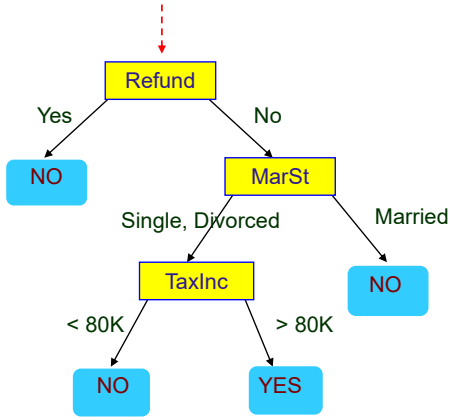
Training Data



Model: Decision Tree

Apply Model to Test Data

Start from the root of tree.



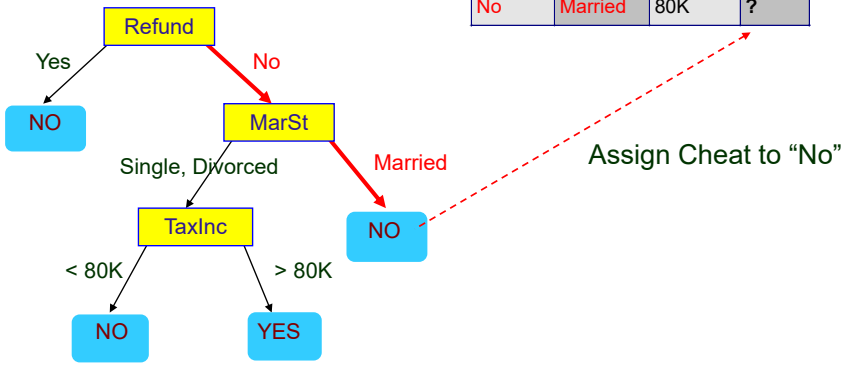
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Assign Cheat to "No"

CLASSIFICATION METHODS

Challenges

How to represent the entire information in the dataset using minimum number of rules?

How to develop the smallest tree?

Solution

Select the variable with maximum information for first split

Training

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Attribute Selection Measure: Information Gain (ID3/C4.5)

- Select the attribute with the highest information gain
- Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$
- Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

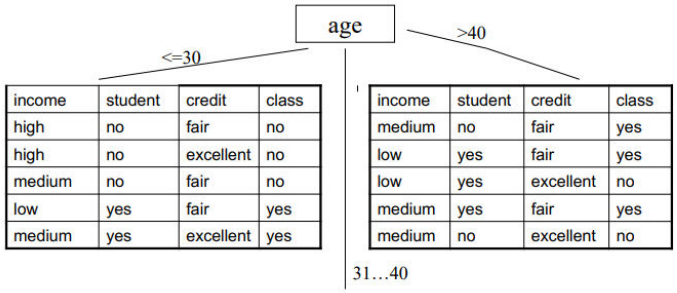
- Information needed (after using A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

- Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

Building The Tree: we choose "age" as a root

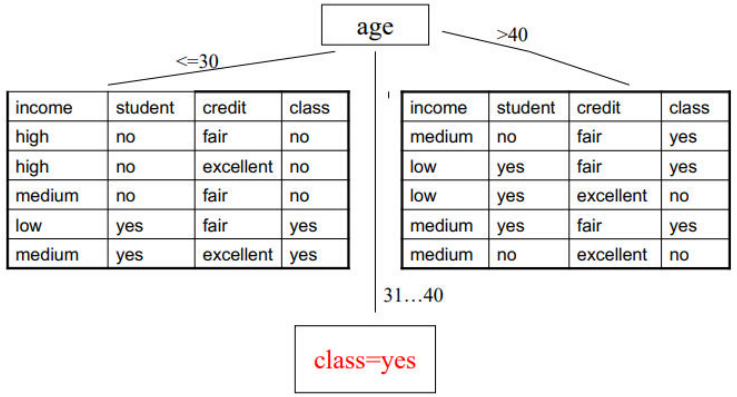


income	student	credit	class
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	student	credit	class
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

income	student	credit	class
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

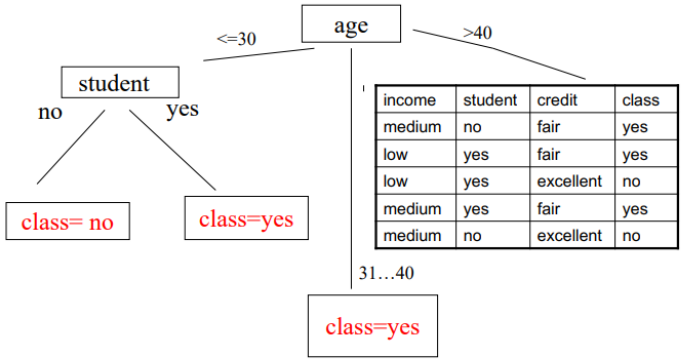
Building The Tree: "age" as the root



income	student	credit	class
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

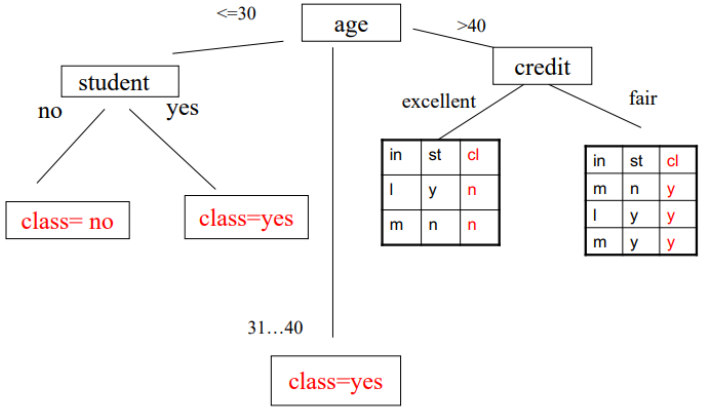
income	student	credit	class
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

Building The Tree: we chose "student" on <=30 branch



income	student	credit	class
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

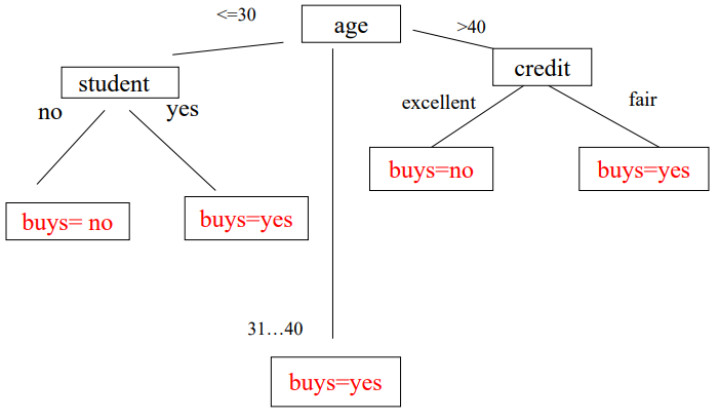
Building The Tree: we chose "credit" on >40 branch



in	st	cl
l	y	n
m	n	n

in	st	cl
m	n	y
l	y	y
m	y	y

Finished Tree for class="buys"



Extracting Classification Rules from Trees

- **Goal:** Represent the knowledge in the form of
- IF-THEN determinant rules
- **One rule** is created for **each path** from the root to a leaf;
- **Each attribute-value** pair along a **path** forms a conjunction;
- The **leaf node** holds the class prediction

- Rules are easier to understand

Discriminant RULES extracted from our TREE

- The rules are:

IF *age* = "<=30" AND *student* = "*no*" THEN
buys_computer = "*no*"
IF *age* = "<=30" AND *student* = "*yes*" THEN
buys_computer = "*yes*"
IF *age* = "31…40" THEN
buys_computer = "*yes*"
IF *age* = ">40" AND *credit_rating* = "*excellent*" THEN
buys_computer = "no"
IF *age* = ">40" AND *credit_rating* = "*fair*" THEN
buys_computer = "yes"

Training

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Attribute Selection Measure: Information Gain (ID3/C4.5)

- Select the attribute with the highest information gain
- Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$
- Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Information needed (after using A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

- Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

Information Gain

- Information gain is used as an attribute selection measure
- Pick the attribute that has the highest Information Gain

$$Gain(D, A) = Entropy(D) - \sum_{j=1}^{v} \frac{|D_j|}{|D|} Entropy(D_j)$$

D: A given data partition
A: Attribute
v: Suppose we were partition the tuples in ***D*** on some attribute ***A*** having ***v*** distinct values
D is split into v partition or subsets, {***D1, D2, … Dj***}, where ***Dj*** contains those tupes in ***D*** that have outcome a_j of ***A***.

Attribute Selection: Information Gain

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

$$Info_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

$\frac{5}{14} I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes's and 3 no's.

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Attribute Selection: Information Gain

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

$$Info_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

$\frac{5}{14} I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes's and 3 no's. Hence

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

Similarly find the following:

$$Gain(income) = ?$$
$$Gain(student) = ?$$
$$Gain(credit\_rating) = ?$$

Decision Tree:

Decision Tree (1.1)

Data Set

Age	Income	Student	Credit Rating	Buy Computer
<=30	high	no	fair	no
<=30	high	no	Ex	no
31...40	high	no	F	Yes
>40	med	no	F	Yes
>40	low	yes	F	Yes
>40	low	Yes	Ex	no
31...40	low	Yes	Ex	Yes
<=30	med	no	F	no
<=30	low	yes	F	yes
>40	medium	yes	F	yes
<=30	med.	yes	Ex	yes
31...40	med.	no	Ex	yes
31...40	high	yes	F	yes
>40	med.	no	Ex	no

Attribute Selection (1.2)

$Gain_{(Attribute)} = Info_{(Dataset)} - Info_{Attribute\ (Dataset)}$

$Info\ (Dataset) = -\sum_{i=1}^{m} P_i \log_2 (P_i)$

m → no. of classification

$P_i = \frac{S_i}{S}$, S_i (w.r.t → yes) & S_i (w.r.t → no)

S → Total Data.

$Info_{(Attr)}\ Dataset = \sum_{j=1}^{v} \left|\frac{D_j}{D}\right| Info\ (D_j)$

v → w.r.t. attributes.

Log. calculation

$\log_2(x) = \frac{\ln(x)}{\ln(2)}$

$Info\ (D) \rightarrow Info\ (9,5) = -\frac{9}{14}\log_2 \frac{9}{14} - \frac{5}{14}\log_2 \frac{5}{14}$

$= -\frac{9}{14} \times -0.637 - \frac{5}{14}$

$= -0.643 \times 0.637 - 0.357 \times -1.485$

$= 0.4096 + 0.531$

$= 0.94$

Attribute	P_i Yes	N_i No	$I(P_i, N_i)$
<=30	2	3	I(2,3)
31...40	4	0	I(4,0)
>40	3	2	I(3,2)

(1) Age

$Info_{Age}(D)$: $Info\ (2,3) = -\frac{2}{14}\log\frac{2}{14} - \frac{3}{14}\log\frac{3}{14}$

$= -0.143 \times -2.806 - 0.214 \times -2.224$

$= 0.847$

$Info\ (4,0) = -\frac{4}{4}\log_2 \frac{4}{4} - \frac{0}{4}\log_2 \frac{0}{4}$

$= 0$

$Info\ (3,2) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log_2 \frac{2}{5}$

$Info_{Age}(D) = \sum_{i=1}^{3} \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2)$

$Gain_{(Age)} = Info(D) - Info_A(D)$

$= 0.94 - 0.693$

$Gain(Age) = 0.247$

ii) Income

Attribute	P_i Yes	N_i No	$I(P_i, N_i)$
High	2	2	I(2,2)
Med	4	2	I(4,2)
Low	3	1	I(3,1)

$I(2,2) = -\frac{2}{4}\log_2 \frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = -(-0.5 \times -1) \times 2 = 1$

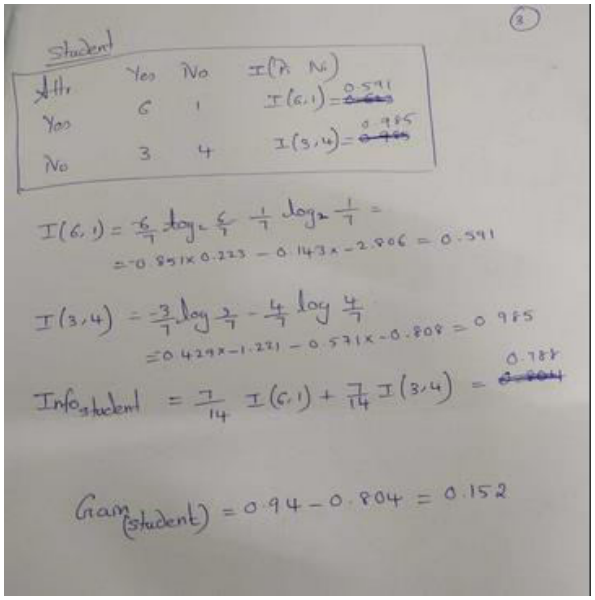
$I(4,2) = -\frac{4}{6}\log\frac{4}{6} - \frac{2}{6}\log\frac{2}{6} = (0.667 \times -0.584) - (-.333 \times -1.585) = 0.917$

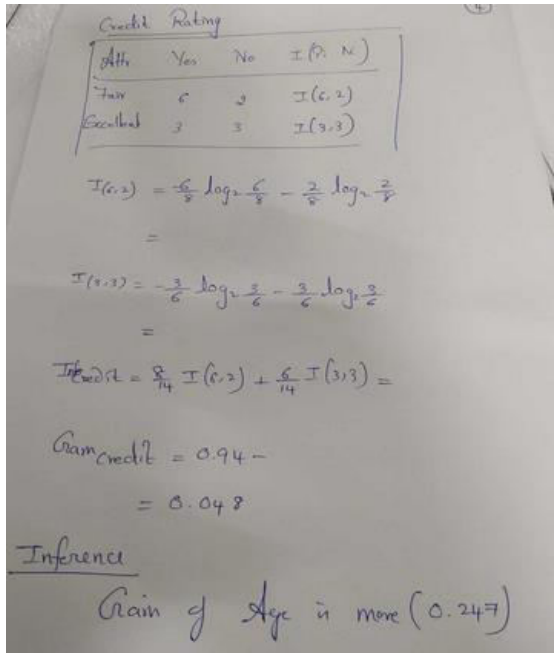
$I(3,1) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = (0.75 \times -0.415) - (-.25 \times -2) = 0.811$

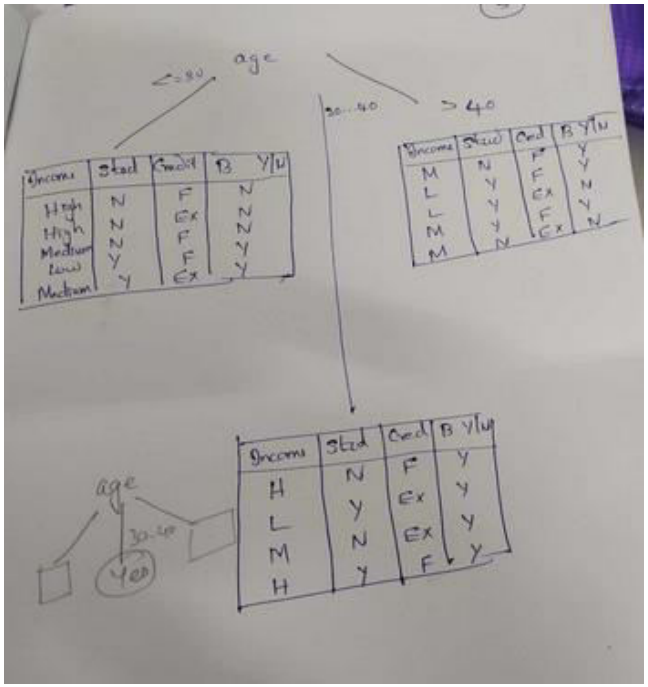
$Info_{Income} = \frac{4}{14} I(2,2) + \frac{6}{14} I(4,2) + \frac{4}{14} I(3,1) = 0.91$

$Gain\ (Income) = 0.94 - 0.91$

$= 0.03$







Problem 2

	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

Conditional Probability

Probability

A ***probability model*** is a mathematical representation of a random phenomenon.

An ***event A*** is a subset of the sample space S.

A ***probability*** is a numerical value assigned to a given event *A*. The probability of an event is written *P(A)*

The first two basic rules of probability are the following:

Rule 1: *Any probability P(A) is a number between 0 and 1 ($0 \leq P(A) \leq 1$).*

Rule 2: *The probability of the sample space S is equal to 1 (P(S) = 1).*

Example:

Let us take "weather" as a sample space S,

To find P(weather), consider the probability value of all possible events.
Weather – **random variable**

P(Weather = Sunny) = 0.1
P(Weather = Rain) = 0.7
P(Weather = Snow) = 0.2

P(Weather) = (0.1, 0.7, 0.2) – probability distribution

Axioms of probability

1. All probability are between 0 and 1: $0 \leq P(A) \leq 1$
1. P(true) = 1 and P(false) = 0
2. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
3. P(~A) = 1 – P(A)

Prior Probability

P(H)
is the *a priori* probability that a specified hypothesis is true. This is often called the prior probability, or just the *prior*.
This is the **unconditional** probability, without taking any evidence into consideration.

Prior probability

Consider the random variables,
cavity={true, false}
weather={sunny, rain, cloudy, snow}

Prior or unconditional probability,
P(cavity=true)=0.1
P(weather=sunny)=0.72

Probability distribution gives values of all possible assignments:
P(weather)={0.72, 0.1, 0.08, 0.1}(normalized i.e, sums to 1)

Joint probability distribution

Joint probability distribution for a set of random variables gives the probability of every atomic event (sample point) on those variables

P(Weather, Cavity) = a 4 X 2 matrix of values:

	Weather =			
	sunny	*rain*	*cloudy*	*snow*
Cavity = *true*	0.144	0.02	0.016	0.02
Cavity = *false*	0.576	0.08	0.064	0.08

Joint probability distribution

With two variables

	Fever	¬ *Fever*
flu	p_1	p_2
¬*flu*	p_5	p_6

With three variables

	cough		¬ *cough*	
	Fever	¬ *Fever*	*Fever*	¬ *Fever*
flu	p_1	p_2	p_3	p_4
¬*flu*	p_5	p_6	p_7	p_8

Joint probability distribution

With four variables

	smokes			
	cough		¬ cough	
	Fever	¬ Fever	Fever	¬ Fever
flu	p_1	p_2	p_3	p_4
¬flu	p_5	p_6	p_7	p_8

	¬ smokes			
	cough		¬ cough	
	fever	¬ fever	fever	¬ fever
flu	p_9	p_{10}	p_{11}	p_{12}
¬flu	p_{13}	p_{14}	p_{15}	p_{16}

Joint probability distribution

With five variables

	Allergy			
	smokes			
	cough		¬ cough	
	Fever	¬Fever	Fever	¬ Fever
flu	p_1	p_2	p_3	p_4
¬flu	p_5	p_6	p_7	p_8

	¬Allergy			
	smokes			
	cough		¬ cough	
	Fever	¬Fever	Fever	¬ Fever
flu	p_{17}	p_{18}	p_{19}	p_{20}
¬flu	p_{21}	p_{22}	p_{23}	p_{24}

	Allergy			
	¬ smokes			
	cough		¬ cough	
	fever	¬fever	fever	¬fever
flu	p_9	p_{10}	p_{11}	p_{12}
¬flu	p_{13}	p_{14}	p_{15}	p_{16}

	¬Allergy			
	¬ smokes			
	cough		¬ cough	
	fever	¬fever	fever	¬fever
flu	p_{25}	p_{26}	p_{27}	p_{28}
¬flu	p_{29}	p_{30}	p_{31}	p_{32}

Joint Probability

We can express the conditional probability,

in terms of the joint probability

P(H|E)
(probability that hypothesis H is true given that evidence E is observed)
P(H ∩ E),
(probability that both H is true and E is observed).

$$P(H|E) = \frac{P(H \cap E)}{P(E)}$$

This states that the probability that hypothesis H is true if evidence E is observed is equal to the joint probability of H and E, divided by the prior probability of E.

Bayes' Rule

The condition probability of the occurrence of A if event B occurs

$P(A|B) = P(A \wedge B) / P(B)$

This can be written also as:
$P(A \wedge B) = P(A|B) * P(B)$
$P(A \wedge B) = P(B|A) * P(A)$

Hence $P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$

Conditional Probability and Product Rule

$$P(H|E) = \frac{P(H \cap E)}{P(E)}$$

and similarly

$$P(E|H) = \frac{P(E \cap H)}{P(H)}$$

Multiplying by P(H) we get

$$P(E \cap H) = P(E|H)\, P(H)$$

Product rule

Product Rule:

$$P(E \cap H) = P(E|H)P(H) = P(H|E)P(E)$$

The probability of A and B being true is equal to the probability of A being true if we observe B, times the probability that B is true.

The probability of A and B being true is equal to the probability of B being true if we observe A, times the probability that A is true.

Start with the joint distribution $\mathbf{P}(Cavity, Catch, Toothache)$:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

$$P(\neg cavity|toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)} = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

Naïve Bayes

Naïve Bayes Classifier Algorithm

•Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.

• The Naive Bayes classifier works on the principle of conditional probability, as given by the Bayes theorem.

According to Bayes theorem:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

where:

P(A|B) = Conditional Probability of A given B
P(B|A) = Conditional Probability of A given B
P(A) = Probability of event A
P(B) = Probability of event A

Working of Naïve Bayes' Classifier:

Suppose we have a dataset of **weather conditions** and corresponding target variable "**Play**". So using this dataset we need to decide that whether we should play or not on a particular day according to the weather conditions.

To solve this problem, we need to follow the below steps:

1.Convert the given dataset into frequency tables.
2.Generate Likelihood table by finding the probabilities of given features.
3.Now, use Bayes theorem to calculate the posterior probability.

Problem 1: If the weather is sunny, then the Player should play or not?

S

	Outlook	Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rainy	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rainy	No
9	Sunny	No
10	Sunny	Yes
11	Rainy	No
12	Overcast	Yes
13	Overcast	Yes

Frequency table for the Weather Conditions:

Weather	Yes	No
Overcast	5	0
Rainy	2	2
Sunny	3	2
Total	10	4

Likelihood table weather condition:

Weather	No	Yes	
Overcast	0	5	5/14= 0.35
Rainy	2	2	4/14=0.29
Sunny	2	3	5/14=0.35
All	4/14=0.29	10/14=0.71	

Applying Bayes'theorem:

P(Yes|Sunny)= P(Sunny|Yes)*P(Yes)/P(Sunny)

P(Sunny|Yes)= 3/10= 0.3

P(Sunny)= 0.35

P(Yes)=0.71

So P(Yes|Sunny) = 0.3*0.71/0.35= **0.60**

P(No|Sunny)= P(Sunny|No)*P(No)/P(Sunny)

P(Sunny|NO)= 2/4=0.5

P(No)= 0.29

P(Sunny)= 0.35

So P(No|Sunny)= 0.5*0.29/0.35 = **0.41**

So as we can see from the above calculation that **P(Yes|Sunny)>P(No|Sunny)**

Hence on a Sunny day, Player can play the game.

Problem 2

	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

Let us test it on a new set of features (let us call it today):
today = (Sunny, Hot, Normal, False)

Outlook

	Yes	No	P(yes)	P(no)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Temperature

	Yes	No	P(yes)	P(no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

Humidity

	Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

Wind

	Yes	No	P(yes)	P(no)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

Play		P(Yes)/P(No)
Yes	9	9/14
No	5	5/14
Total	14	100%

So, probability of playing golf is given by:

$$P(Yes|today) = \frac{P(SunnyOutlook|Yes)P(HotTemperature|Yes)P(NormalHumidity|Yes)P(NoWind|Yes)P(Yes)}{P(today)}$$

and probability to not play golf is given by:

$$P(No|today) = \frac{P(SunnyOutlook|No)P(HotTemperature|No)P(NormalHumidity|No)P(NoWind|No)P(No)}{P(today)}$$

P(today)=P(sunny)*P(hot)*p(normal)*p(nowind)
=(5/14)*(4/14)*(7/14)*(8/14)

P(Yes/today)=((2/9)*(2/9)*(6/9)*(6/9)*(9/14))/((5/14)*(4/14)*(7/14)*(8/14))

=0.484

P(No/today)=((3/5)*(2/5)*(1/5)*(2/5)*(5/14))/((5/14)*(4/14)*(7/14)*(8/14))

=0.235

Since

$P(Yes|today) > P(No|today)$

So, prediction that golf would be played is 'Yes'.

Problem 3: Practice

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

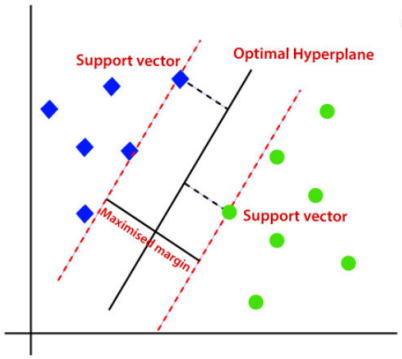
Support Vector Machine

SVM

- Support Vector Machine" (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

SVM

- It is a supervised machine learning problem where we try to find a hyperplane that best separates the two classes.
- Support Vectors are simply the coordinates of individual observation. The SVM classifier is a frontier that best segregates the two classes (hyper-plane/ line).



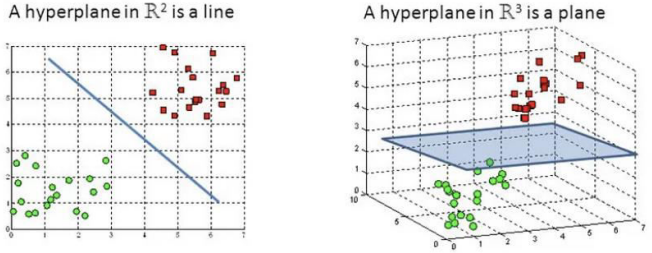
Types of SVM

•**Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

•**Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

Hyperplane and Support Vectors in the SVM algorithm:

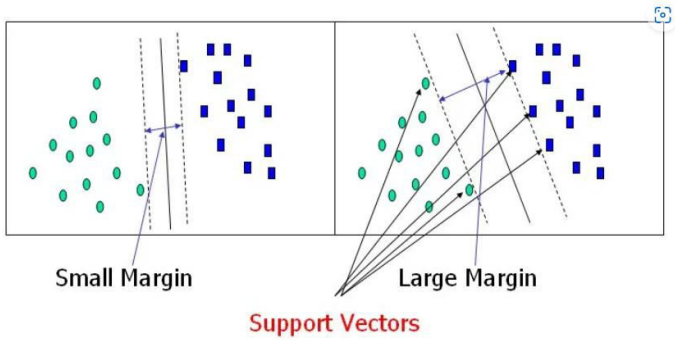
Hyperplane: There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM. The dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine wh[illegible]



Hyperplanes in 2D and 3D feature space

Hyperplane and Support Vectors in the SVM algorithm:

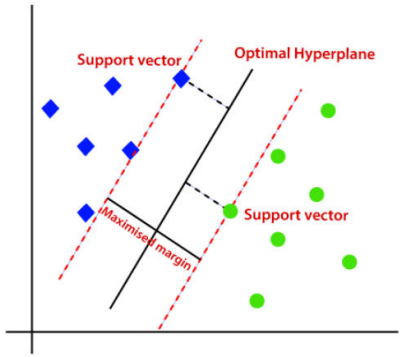
Support Vectors:
The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.



Support Vectors

Selecting the best hyper-plane:

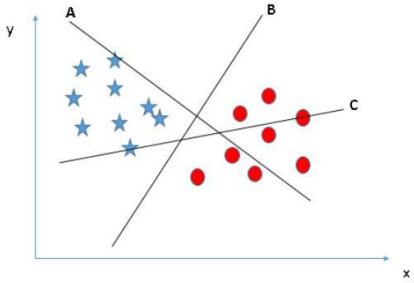
The distance between the support vectors and the hyperplane is called as **margin**. And the goal of SVM is to maximize this margin.
The **hyperplane** with maximum margin is called the **optimal hyperplane**.



How SVM works in different scenario?

Identify the right hyper-plane (Scenario-1)

Here, we have three hyper-planes (A, B, and C). Now, identify the right hyper-plane to classify stars and circles.

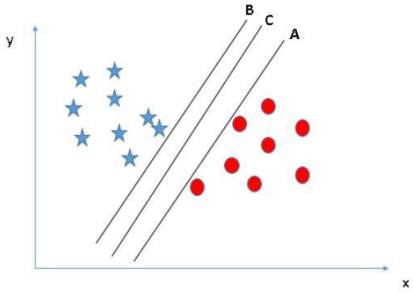


"Select the hyper-plane which segregates the two classes better." In this scenario, hyper-plane "B" has excellently performed this job.

How SVM works in different scenario?

Identify the right hyper-plane (Scenario-2)

we have three hyper-planes (A, B, and C), and all segregate the classes well. Now, How can we identify the right hyper-plane?

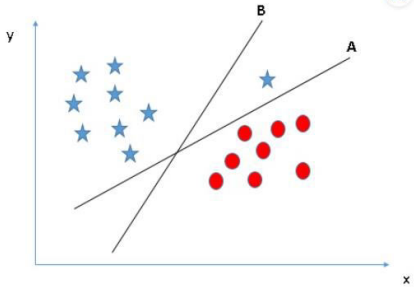


Maximizing the distances between the nearest data point (either class) and the hyper-plane will help us to decide the right hyper-plane. This distance is called a **Margin**. We name the right hyper-plane as C

How SVM works in different scenario?

Identify the right hyper-plane (Scenario-3)

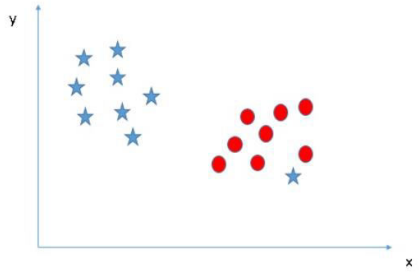
Use the rules as discussed in the previous section to identify the right hyper-plane.



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

How SVM works in different scenario?

Can we classify two classes (Scenario-4)?

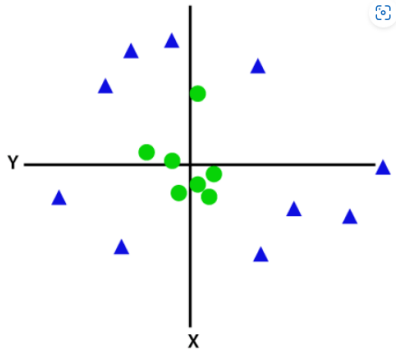


One star at the other end is like an outlier for the star class. The SVM algorithm has a feature to ignore outliers and find the hyper-plane that has the maximum margin. Hence, we can say SVM classification is robust to outliers.

How SVM works in different scenario?

Find the hyper-plane to segregate to classes (Scenario-5)

In the scenario below, we can't have a linear hyper-plane between the two classes, so how does SVM classify these two classes? This could be possible by Non-linear SVM.

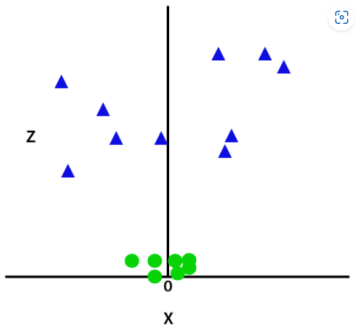


How SVM works in different scenario?

Find the hyper-plane to segregate to classes (Scenario-5)

SVM can solve this problem. It solves this problem by introducing additional features. Here, we will add a new feature, $z=x^2+y^2$

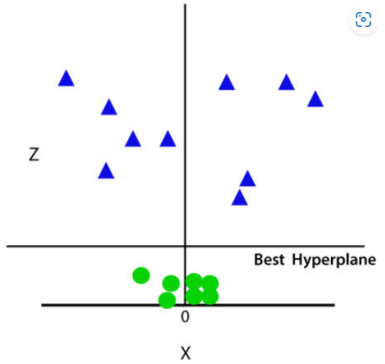
By adding the third dimension, the sample space will become as below image:



How SVM works in different scenario?

Find the hyper-plane to segregate to classes (Scenario-5)

So now, SVM will divide the datasets into classes in the following way.



How SVM works in different scenario?

Find the hyper-plane to segregate to classes (Scenario-5)

How to add this feature manually to have a hyper-plane. This process in SVM has a technique called the **kernel trick**.

The SVM kernel is a function that takes low dimensional input space and transforms it to a higher dimensional space, i.e., it converts not separable problem to a separable problem. It is mostly useful in non-linear data separation problems.

Different Kernel functions

1. Polynomial kernel

Following is the formula for the polynomial kernel:

$$f(X1, X2) = (X1^T . X2 + 1)^d$$

Here d is the degree of the polynomial, which we need to specify manually.

2. Sigmoid kernel

We can use it as the proxy for neural networks. Equation is:

$$f(x1, x2) = \tanh(\alpha x^T y + x)$$

It is just taking your input, mapping them to a value of 0 and 1 so that they can be separated by a simple straight line.

Advantages of SVM

- Support vector machine works comparably well when there is an understandable margin of dissociation between classes.
- It is more productive in high-dimensional spaces.
- It is effective in instances where the number of dimensions is larger than the number of specimens.
- Support vector machine is comparably memory systematic. Support Vector Machine (SVM) is a powerful supervised machine learning algorithm with several advantages. Some of the main advantages of SVM include:
- Handling high-dimensional data: SVMs are effective in handling high-dimensional data, which is common in many applications such as image and text classification.
- Handling small datasets: SVMs can perform well with small datasets, as they only require a small number of support vectors to define the boundary.
- Modeling non-linear decision boundaries: SVMs can model non-linear decision boundaries by using the kernel trick, which maps the data into a higher-dimensional space where the data becomes linearly separable.
- Robustness to noise: SVMs are robust to noise in the data, as the decision boundary is determined by the support vectors, which are the closest data points to the boundary.
- Generalization: SVMs have good generalization performance, which means that they are able to classify new, unseen data well.
- Versatility: SVMs can be used for both classification and regression tasks, and it can be applied to a wide range of applications such as natural language processing, computer vision and bioinformatics.
- Sparse solution: SVMs have sparse solutions, which means that they only use a subset of the training data to make predictions. This makes the algorithm more efficient and less prone to overfitting.

Disadvantages of support vector machine:

•Computationally expensive: SVMs can be computationally expensive for large datasets, as the algorithm requires solving a quadratic optimization problem.
•Choice of kernel: The choice of kernel can greatly affect the performance of an SVM, and it can be difficult to determine the best kernel for a given dataset.
•Sensitivity to the choice of parameters: SVMs can be sensitive to the choice of parameters, such as the regularization parameter, and it can be difficult to determine the optimal parameter values for a given dataset.
•Memory-intensive: SVMs can be memory-intensive, as the algorithm requires storing the kernel matrix, which can be large for large datasets.
•Limited to two-class problems: SVMs are primarily used for two-class problems, although multi-class problems can be solved by using one-versus-one or one-versus-all strategies.
•Lack of probabilistic interpretation: SVMs do not provide a probabilistic interpretation of the decision boundary, which can be a disadvantage in some applications.
•Not suitable for large datasets with many features: SVMs can be very slow and can consume a lot of memory when the dataset has many features.
•Not suitable for datasets with missing values: SVMs requires complete datasets, with no missing values, it can not handle missing values.

Random Forest Classifier

Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.*

Introduction

As the name suggests, ***"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."*** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

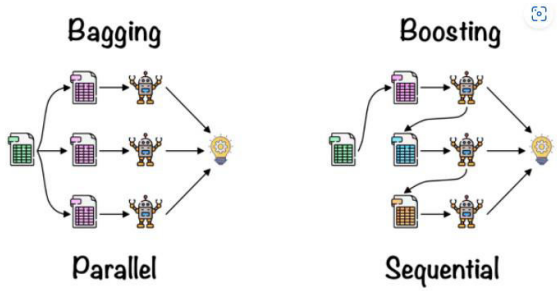
Working of Random Forest Algorithm

Ensemble simply means combining multiple models. Thus a collection of models is used to make predictions rather than an individual model.

Ensemble uses two types of methods:

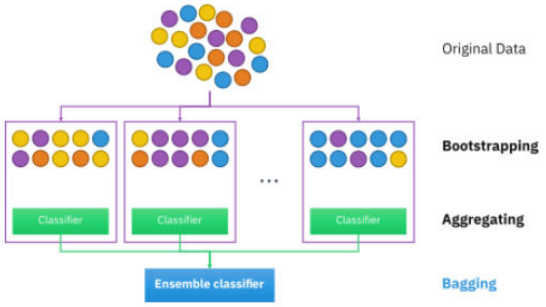
Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.

2. **Boosting**– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.



Bagging

Bagging, also known as ***Bootstrap Aggregation,*** is the ensemble technique used by random forest. Bagging chooses a random sample/random subset from the entire data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as ***row sampling***. This step of row sampling with replacement is called ***bootstrap***.
Now each model is trained independently, which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting, is known as ***aggregation***.



Steps Involved in Random Forest Algorithm

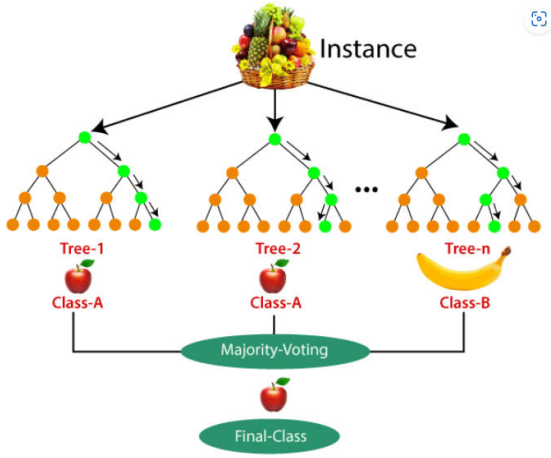
Step 1: In the Random forest model, a subset of data points and a subset of features is selected for constructing each decision tree. Simply put, n random records and m features are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on ***Majority Voting or Averaging*** for Classification and regression, respectively.

Random Forest Algorithm: Example



Important Features of Random Forest

- **Diversity:** Not all attributes/variables/features are considered while making an individual tree; each tree is different.
- **Immune to the curse of dimensionality:** Since each tree does not consider all the features, the feature space is reduced.
- **Parallelization:** Each tree is created independently out of different data and attributes. This means we can fully use the CPU to build random forests.
- **Train-Test split:** In a random forest, we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
- **Stability:** Stability arises because the result is based on majority voting/ averaging.

Difference Between Decision Tree and Random Forest

Decision trees	Random Forest
1. Decision trees normally suffer from the problem of overfitting if it's allowed to grow without any control.	1. Random forests are created from subsets of data, and the final output is based on average or majority ranking; hence the problem of overfitting is taken care of.
2. A single decision tree is faster in computation.	2. It is comparatively slower.
3. When a data set with features is taken as input by a decision tree, it will formulate some rules to make predictions.	3. Random forest randomly selects observations, builds a decision tree, and takes the average result. It doesn't use any set of formulas.

Performance measure

A confusion matrix is a table that is used to define the performance of a classification algorithm.

Let's understand this with the help of an example. We are trying to predict whether some word is a spam word or not. The 4 cases here will be:

1. The Actual value is Spam and the Predicted Value is Spam.
2. The Actual Value is Spam and the Predicted Value is Non-Spam.
3. The Actual Value is Non-Spam and the Predicted Value is Non-Spam.
4. The Actual Value is Non-Spam and the Predicted Value is Spam.

So, for writing this information into a easily readable format for our computer, we are converting this into a 2*2 matrix

		Predicted Class: Spam	Predicted Class: Non-Spam
Actual Class	Spam	TP=45	FN=20
	Non-Spam	FP=5	TN=30

There are several metrics that can be calculated with Confusion Matrix

1. Error Rate
2. Accuracy
3. Precision
4. Recall (Sensitivity)
5. Specificity
6. F score etc.

Example 1:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Here,

TP = 45

TN = 30

FP = 5

FN = 20

Accuracy = (45 + 30) / (45 + 30 + 5 + 20) = 75/100 = 0.75

K-Means Clustering

K Means Algorithm

Let the Input: Data Points (V1,V2,…..Vn)

Step 1: Pick 'K' data points to the centroids randomly (C_k)

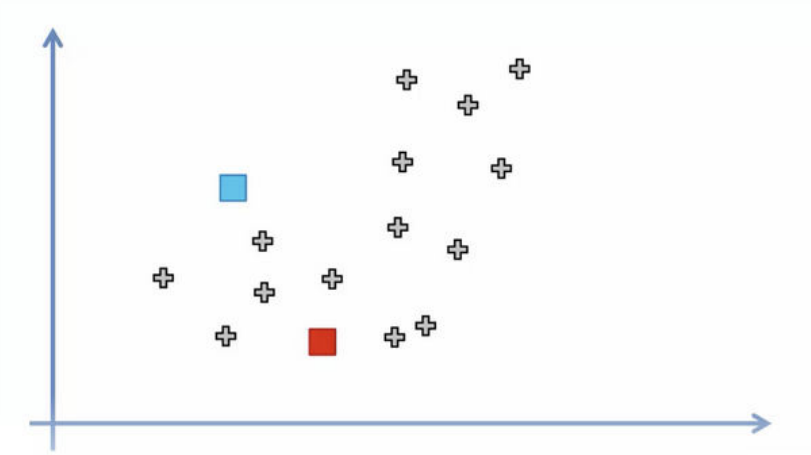
Step 2: Using Euclidean distance assign each data point to closest centroid Cj
arg min dist(Vi,Cj)

Step 3: For each cluster j = 1to k
3.1 find new centroid, Cj(new) = 1/nj Σvi

Step 4: Repeat step 2 and 3 until none of the cluster assignment change

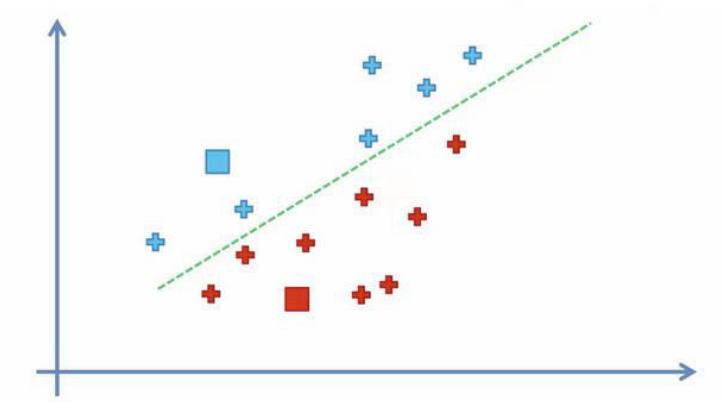
K-Means Clustering Algorithm

1. Select K (i.e. 2) random points as cluster centers called centroids



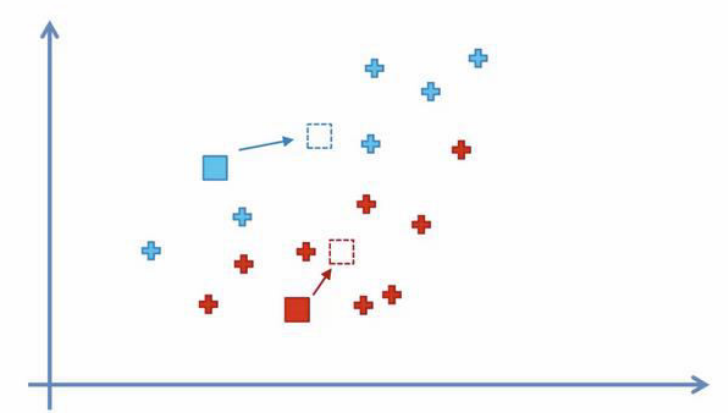
K-Means Clustering Algorithm Cntd..

2. Assign each data point to the closest cluster by calculating its distance with respect to each centroid



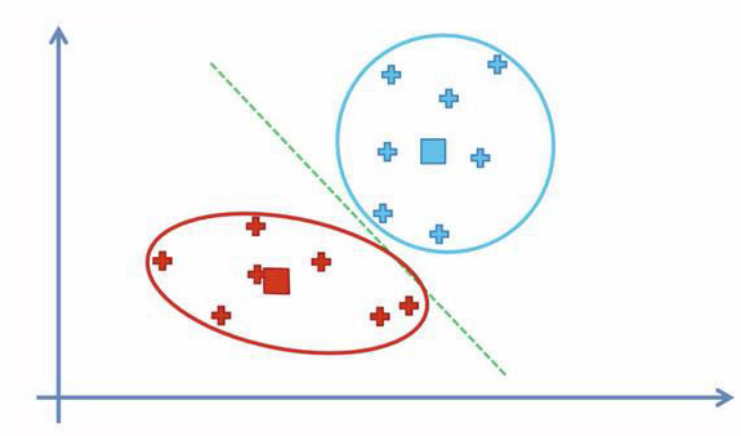
K-Means Clustering Algorithm Cntd..

3. Determine the new cluster center by computing the average of the assigned points



K-Means Clustering Algorithm Cntd..

4. Repeat steps 2 and 3 until none of the cluster assignments change



A Simple example showing the implementation of k-means algorithm (using K=2)

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Step 1:

Initialization: Randomly we choose following two centroids (k=2) for two clusters.

In this case the 2 centroid are: m1=(1.0,1.0) and m2=(5.0,7.0).

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

	Individual	Mean Vector
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

Step 2:

- Thus, we obtain two clusters containing:
  {1,2,3} and {4,5,6,7}.
- Their new centroids are:

$$m_1 = \left(\frac{1}{3}(1.0+1.5+3.0), \frac{1}{3}(1.0+2.0+4.0)\right) = (1.83, 2.33)$$

$$m_2 = \left(\frac{1}{4}(5.0+3.5+4.5+3.5), \frac{1}{4}(7.0+5.0+5.0+4.5)\right)$$

$$= (4.12, 5.38)$$

Individual	Centroid 1	Centroid 2
1	0	7.21
2 (1.5, 2.0)	1.12	6.10
3	3.61	3.61
4	7.21	0
5	4.72	2.5
6	5.31	2.06
7	4.30	2.92

$$d(m_1,2) = \sqrt{|1.0-1.5|^2 + |1.0-2.0|^2} = 1.12$$

$$d(m_2,2) = \sqrt{|5.0-1.5|^2 + |7.0-2.0|^2} = 6.10$$

Step 3:

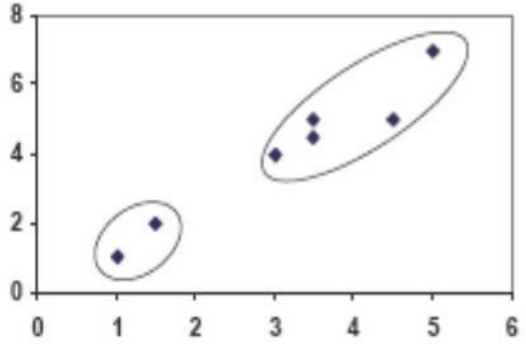
- Now using these centroids we compute the Euclidean distance of each object, as shown in table.
- Therefore, the new clusters are:
  {1,2} and {**3**,4,5,6,7}
- Next centroids are: m1=(1.25,1.5) and m2 = (3.9,5.1)

Individual	Centroid 1	Centroid 2
1	1.57	5.38
2	0.47	4.28
3	2.04	1.78
4	5.64	1.84
5	3.15	0.73
6	3.78	0.54
7	2.74	1.08

- Step 4 :
  The clusters obtained are:
  {1,2} and {3,4,5,6,7}
- Therefore, there is no change in the cluster.
- Thus, the algorithm comes to a halt here and final result consist of 2 clusters {1,2} and {3,4,5,6,7}.

Individual	Centroid 1	Centroid 2
1	0.56	5.02
2	0. 56	3.92
3	3.05	1.42
4	6.66	2.20
5	4.16	0.41
6	4.78	0.61
7	3.75	0.72

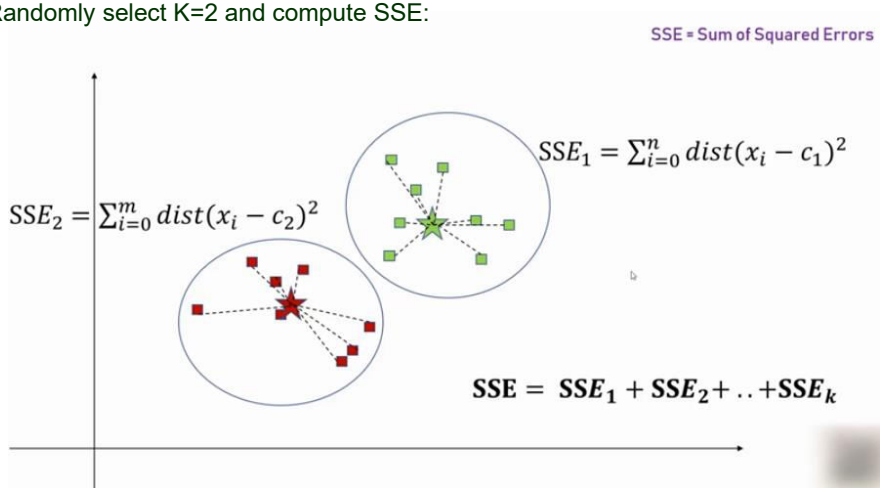
PLOT



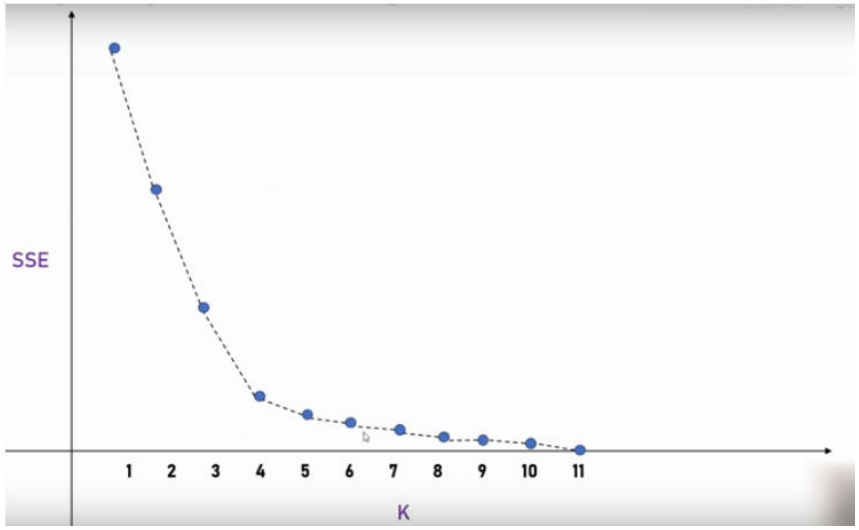
Choosing the right number of clusters in Kmeans Clustering

Elbow Method

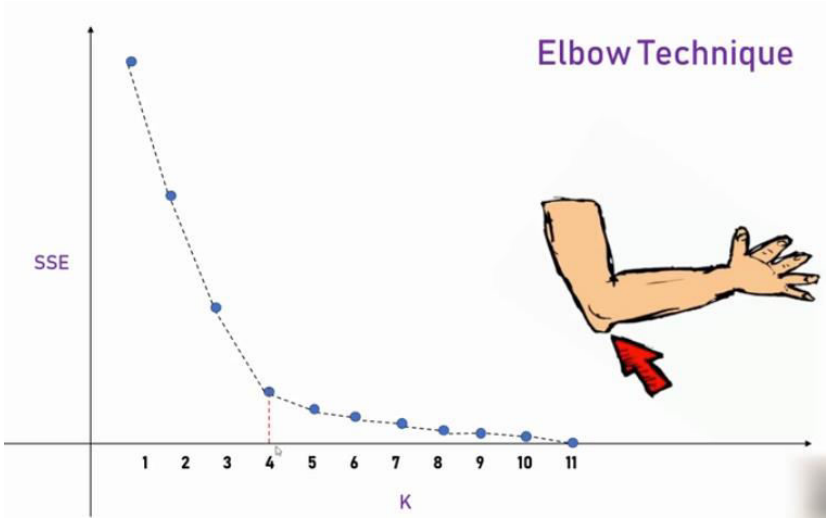
Randomly select K=2 and compute SSE:



Repeat SSE for different value of K and plot it as given below.

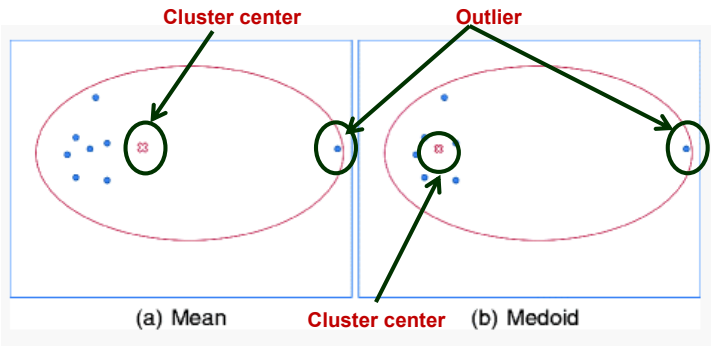


From the above graph, it shows that as we increase the number of cluster, the error got reduced. Find the elbow point.



K- Medoids Clustering

- The K-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values.
- K-medoids clustering is a variant of K-means that is more robust to noises and outliers.
- Instead of using the mean point as the center of a cluster, K-medoids uses an actual point in the cluster to represent it.
- Medoid is the most centrally located object of the cluster, with minimum sum of distances to other points.



- **Mean is greatly influenced by the outlier and thus cannot represent the correct cluster center, while medoid is robust to the outlier and correctly represents the cluster center**

K-Medoids (also called Partitioning Around Medoid) algorithm was proposed in 1987 by Kaufman and Rousseeuw. A medoid can be defined as a point in the cluster, whose dissimilarities with all the other points in the cluster are minimum. The dissimilarity of the medoid(Ci) and object(Pi) is calculated by using E = |Pi – Ci|

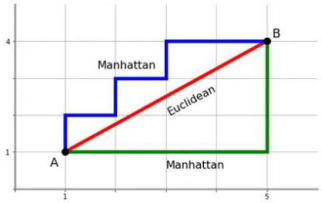
The cost in K-Medoids algorithm is given as

$$c = \sum_{Ci} \sum_{Pi \in Ci} |Pi - Ci|$$

Manhattan Distance

Definition: The distance between two points measured along axes at right angles. In a plane with p_1 at (x_1, y_1) and p_2 at (x_2, y_2), it is $|x_1 - x_2| + |y_1 - y_2|$.

$$\text{Mdist} = |x2 - x1| + |y2 - y1|$$



Partitioning Around Medoids (PAM)

Algorithm:

Given the value of k and unlabelled data:

1. Choose k number of random points from the data and assign these k points to k number of clusters. These are the initial medoids.
2. For all the remaining data points, calculate the distance from each medoid and assign it to the cluster with the nearest medoid.
3. Calculate the total cost (Sum of all the distances from all the data points to the medoids)
4. Select a random point as the new medoid and swap it with the previous medoid. Repeat 2 and 3 steps.
5. If the total cost of the new medoid is less than that of the previous medoid, make the new medoid permanent and repeat step 4.
6. If the total cost of the new medoid is greater than the cost of the previous medoid, undo the swap and repeat step 4.
7. The Repetitions have to continue until no change is encountered with new medoids to classify data points.

The **time complexity** is $O(k * (n-k)^2)$.

Question:

Apply K-Medoid clustering to cluster the following samples/data points: (0,0), (0,1), (1,0), (3,3), (5,6), (8,9), (9,8) and (9,9).

Solution: Assumption: K=2, initial medoids are (0,0) & (5,6), and Manhattan distance is used as a dissimilarity measure/metric.

Data Sample	Manhattan Distance from Cluster-1's medoid (0,0)	Manhattan Distance from Cluster-2's medoid (5,6)
(0,0)		
(0,1)		
(1,0)		
(3,3)		
(5,6)		
(8,9)		
(9,8)		
(9,9)		

Question:

Apply K-Medoid clustering to cluster the following samples/data points: (0,0), (0,1), (1,0), (3,3), (5,6), (8,9), (9,8) and (9,9).

Assumption: K=2, initial medoids are (0,0) & (5,6), and Manhattan distance is used as a dissimilarity measure/metric.

Data Sample	Manhattan Distance from Cluster-1's medoid (0,0)	Manhattan Distance from Cluster-2's medoid (5,6)
(0,0)	0	11
(0,1)	1	10
(1,0)	1	10
(3,3)	6	5
(5,6)	11	0
(8,9)	17	6
(9,8)	17	6
(9,9)	18	7

Cost = 0+1+1+5+0+6+6+7 = 26.

What will be the cost if in Cluster-1, the medoid (0,0) is swapped with the non-medoid data point (1,0)?

Data Sample	Manhattan Distance from Cluster-1's medoid (1,0)	Manhattan Distance from Cluster-2's medoid (5,6)
(0,0)	1	11
(0,1)	2	10
(1,0)	0	10
(3,3)	5	5
(5,6)	10	0
(8,9)	16	6
(9,8)	16	6
(9,9)	17	7

Cost = 1+2+0+5+0+6+6+7 = 27. The cost increases and therefore, the swap should be avoided.

Will there be a decrease in the cost if in Cluster-2, the medoid (5,6) is swapped with the non-medoid data point (8,9)?

Data Sample	Manhattan Distance from Cluster-1's medoid (0,0)	Manhattan Distance from Cluster-2's medoid (8,9)
(0,0)	0	17
(0,1)	1	16
(1,0)	1	16
(3,3)	6	11
(5,6)	11	6
(8,9)	17	0
(9,8)	17	2
(9,9)	18	1

Cost = 0+1+1+6+6+0+2+1 = 17. The cost decreases and therefore, they can be swapped. The medoid for Cluster-2 should be (8,9).

Cluster-1: (0,0), (0,1), (1,0), and (3,3)

Cluster-2: (5,6), (8,9), (9,8), and (9,9)

Apply K Medoids method for grouping the given samples into 2 clusters.

Initial medoids: M1(1, 3) and M2(4, 9)

	x	y
0	5	4
1	7	7
2	1	3
3	8	6
4	4	9

https://www.javatpoint.com/k-medoids-clustering-theoretical-explanation

Advantages:
It is simple to understand and easy to implement.
K-Medoid Algorithm is fast and converges in a fixed number of steps.
PAM is less sensitive to outliers than other partitioning algorithms.

Disadvantages:
The main disadvantage of K-Medoid algorithms is that it is not suitable for clustering non-spherical (arbitrarily shaped) groups of objects. This is because it relies on minimizing the distances between the non-medoid objects and the medoid (the cluster center) – briefly, it uses compactness as clustering criteria instead of connectivity.
It may obtain different results for different runs on the same dataset because the first k medoids are chosen randomly.

Hierarchical Clustering

Hierarchical Clustering

Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster analysis or HCA.

Agglomerative Clustering is a type of hierarchical clustering algorithm.

Hierarchical clustering has a couple of key benefits:

There is no need to pre-specify the number of clusters.

Agglomerative:

Initially, each object is considered to be its own cluster. According to a particular procedure, the clusters are then merged step by step until a single cluster remains. At the end of the cluster merging process, a cluster containing all the elements will be formed.

Step 1: Compute the proximity matrix using a particular distance metric

Step 2: Each data point is assigned to a cluster

Step 3: Merge the clusters based on a metric for the similarity between clusters

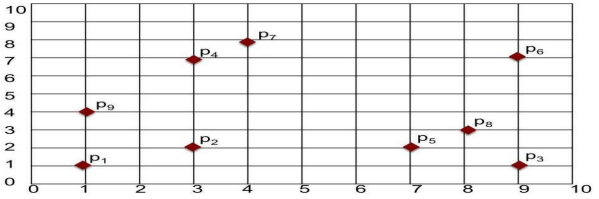
Step 4: Update the distance matrix

Step 5: Repeat Step 3 and Step 4 until only a single cluster remains

Computing a proximity matrix

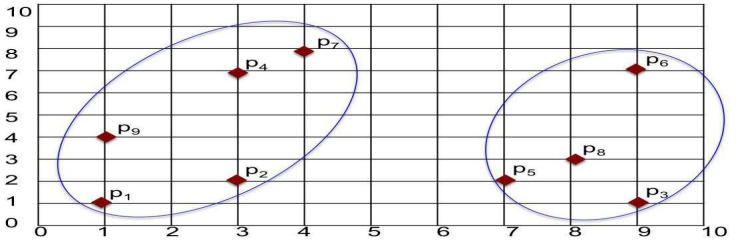
The first step of the algorithm is to create a distance matrix. The values of the matrix are calculated by applying a distance function between each pair of objects. The **Euclidean distance function** is commonly used for this operation.

	p_1	p_2	p_3	...	p_n
p_1	$d(p_1,p_1)$	$d(p_1,p_2)$	$d(p_1,p_3)$	...	$d(p_1,p_n)$
p_2	$d(p_2,p_1)$	$d(p_2,p_2)$	$d(p_2,p_3)$	...	$d(p_2,p_n)$
p_3	$d(p_3,p_1)$	$d(p_3,p_2)$	$d(p_3,p_3)$	...	$d(p_3,p_n)$
...	...	...	...	...	...
p_n	$d(p_n,p_1)$	$d(p_n,p_2)$	$d(p_n,p_3)$	...	$d(p_n,p_n)$



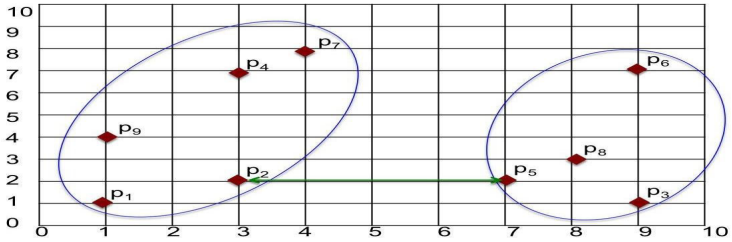
Point	x	y
p_1	1	1
p_2	3	2
p_3	9	1
p_4	3	7
p_5	7	2
p_6	9	7
p_7	4	8
p_8	8	3
p_9	1	4

Suppose we have two clusters in the sample data set, as shown in Figure 2. There are different approaches to calculate the distance between the clusters. Popular methods are listed below.



Min (Single) Linkage

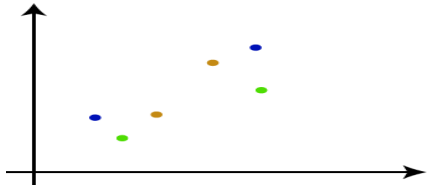
One way to measure the distance between clusters is to find the minimum distance between points in those clusters. That is, we can find the point in the first cluster nearest to a point in the other cluster and calculate the distance between those points



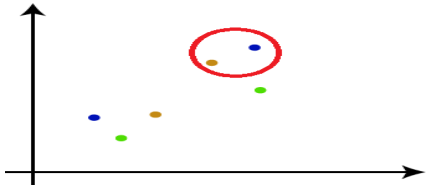
How the Agglomerative Hierarchical clustering Work?

The working of the AHC algorithm can be explained using the below steps:

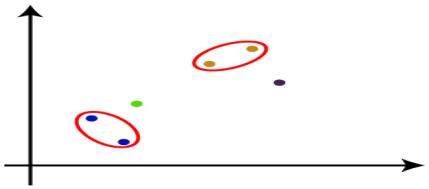
- **Step-1:** Create each data point as a single cluster. Let's say there are N data points, so the number of clusters will also be N.



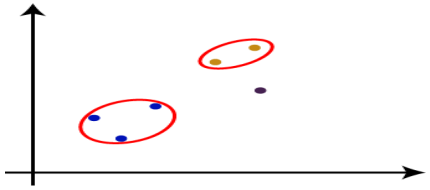
- **Step-2:** Take two closest data points or clusters and merge them to form one cluster. So, there will now be N-1 clusters.

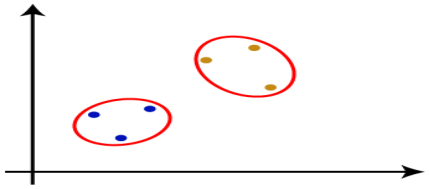


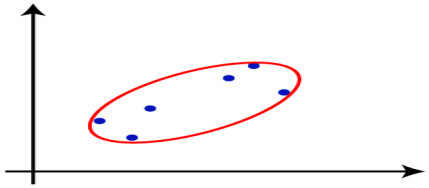
- **Step-3**: Again, take the two closest clusters and merge them together to form one cluster. There will be N-2 clusters.



- **Step-4:** Repeat Step 3 until only one cluster left. So, we will get the following clusters. Consider the below images:





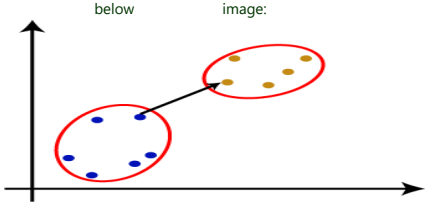


- **Step-5:** Once all the clusters are combined into one big cluster, develop the dendrogram to divide the clusters as per the problem.

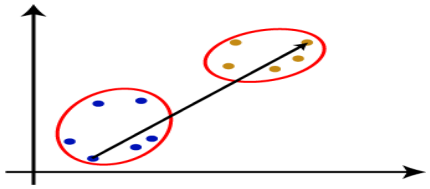
Measure for the distance between two clusters

As we have seen, the **closest distance** between the two clusters is crucial for the hierarchical clustering. There are various ways to calculate the distance between two clusters, and these ways decide the rule for clustering. These measures are called **Linkage methods**. Some of the popular linkage methods are given below:

1. **Single Linkage:** It is the Shortest Distance between the closest points of the clusters. Consider the below image:

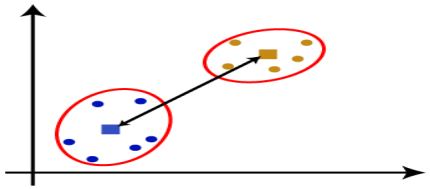


2. **Complete Linkage:** It is the farthest distance between the two points of two different clusters. It is one of the popular linkage methods as it forms tighter clusters than single-linkage.



3. **Average Linkage:** It is the linkage method in which the distance between each pair of datasets is added up and then divided by the total number of datasets to calculate the average distance between two clusters. It is also one of the most popular linkage methods.

4. **Centroid Linkage:** It is the linkage method in which the distance between the centroid of the clusters is calculated. Consider the below image:

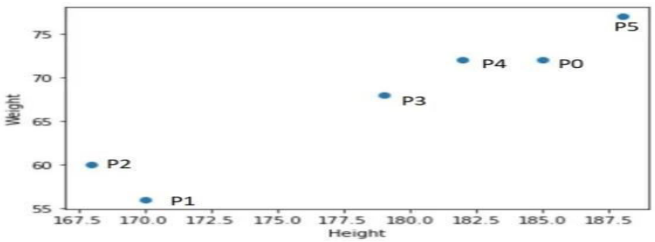


From the above-given approaches, we can apply any of them according to the type of problem or business requirement.

Mathematical Approach to Agglomerative Clustering

Let's take dataset containing Height and Weight of a customer.

Height	Weight
185	72
170	56
168	60
179	68
182	72
188	77



Step 1

P00 = 0, P11 = 0, P22 = 0, P33 = 0, P44 = 0
(this is because distance between self is 0)

Distance between two points P12
= sqrt((P1.X — P2.X)² + (P1.Y — P2.Y)²)
= sqrt((170–168)² + (56–60)²)
= sqrt(4 + 16) = sqrt(20) = 4.47

Similarly, we have to calculate the distance between all the clusters and make a distance matrix.

	P0	P1	P2	P3	P4	P5
P0	0					
P1	21.93	0				
P2	20.81	4.47	0			
P3	7.21	15	13.6	0		
P4	3	20	18.44	5	0	
P5	5.83	27.66	26.25	12.73	7.81	0

Now, we have to see which two cluster has minimum distance.
Its the distance between P0 and P4 which is 3. MERGE PO AND P4

Remove the row P4 and column P4

After merging P0 and P4

How we have arrive the value of P1-[P0,P4] , P2-[P0,P4],P3- [P0,P4],P5-[P0,P4].

We have got these values with the help of single linkage method.

	[P0,P4]	P1	P2	P3	P5
[P0,P4]	0				
P1	20	0			
P2	18.44	4.47	0		
P3	5	15	13.6	0	
P5	5.83	27.66	26.25	12.73	0

It says that, Distance of P1-[P0,P4] = d(P1,[P0-P4])
= min(d(P1,P0),d(P1,P4)) = min(21.93, 20) = 20

Distance of P2-[P0,P4] = d(P2,[P0,P4])
= min(d(P2,P0),d(P2,P4)) = min(20.81, 18.44) = 18.44

Similarly we have calculated all the distances. Step 3: Repeat step 2

Again the minimum distance is P1-P2. So, the next distance matrix will be:

	[P0,P4]	[P1,P2]	P3	P5
[P0,P4]	0			
[P1,P2]	18.44	0		
P3	5	13.6	0	
P5	5.83	26.25	12.73	0

Step 3: Repeat step 2

Now minimum distance is P3-[P0,P4] which is 5. So, the next distance matrix will be:

	[P3,[P0,P4]]	[P1,P2]	P5
[P3,[P0,P4]]	0		
[P1,P2]	13.6	0	
P5	5.83	26.25	0

Step 3: Repeat step 2

Now minimum distance is P5-[P3,[P0,P4]] which is 5.83. So, the next distance matrix will be:

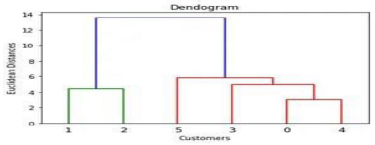
	[P5,[P3,[P0,P4]]]	[P1,P2]
[P5,[P3,[P0,P4]]]	0	
[P1,P2]	13.6	0

Step 3: Repeat step 2
Now there are only two clusters whose distance is 13.6. So, the final distance matrix will be:

Merging of [P1,P2] and [P5,[P3,[P0,P4]]]

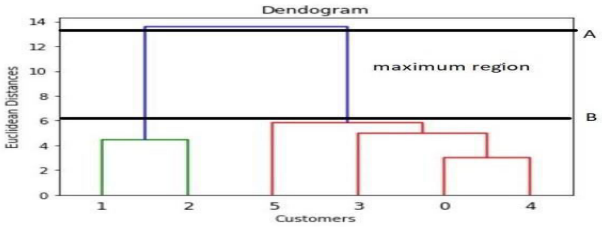
	[[P1,P2],[P5,[P3,[P0,P4]]]]
[[P1,P2],[P5,[P3,[P0,P4]]]]	0

Step 4: Create a Dendrogram to visualize the history of groupings.



As we can see in the dendrogram firstly P0 and P4 are merged, then P1 and P2 are merged, then P3 and [P0,P4] merged, then P5 and [P3,[P0,P4]] and finally [P1,P2] and [P5,[P3,[P0,P4]]].

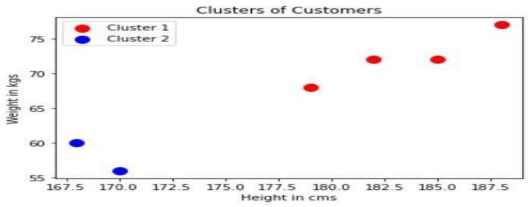
Step 5: Find optimal number of clusters from Dendrogram.



For finding the optimal number of clusters we need to :
1.Determine the largest vertical distance that doesn't intersect any other cluster.
2.Draw two horizontal lines at both extremes like A and B in above figure.
3.The optimal number of cluster = number of vertical lines going through the horizontal lines.

Here, from above Dendrogram we can clearly see that there are 2 vertical lines going through horizontal lines.

Therefore, Optimal number of clusters = 2.



Agglomerative Hierarchical Clustering Solved Example

- Consider the following set of 6 one dimensional data points:
- 18, 22, 25, 42, 27, 43
- Apply the agglomerative hierarchical clustering algorithm to build the hierarchical clustering **dendogram**.

Agglomerative Hierarchical Clustering Solved Example

- Step – 1

	18	22	25	27	42	43
18	0	4	7	9	24	25
22	4	0	3	5	20	21
25	7	3	0	2	17	18
27	9	5	2	0	15	16
42	24	20	17	15	0	1
43	25	21	18	16	1	0

Step 2

Minimum is 1

Merge 43 into 42 remove the row 43 and column 43

	18	22	25	27	42	43
18	0	4	7	9	24	25
22	4	0	3	5	20	21
25	7	3	0	2	17	18
27	9	5	2	0	15	16
42	24	20	17	15	0	1
43	25	21	18	16	1	0

(42, 43)

Minimum is 2 (row 27 & col 25)

(25,27) is a cluster

	18	22	25	27	42, 43
18	0	4	7	9	24
22	4	0	3	5	20
25	7	3	0	2	17
27	9	5	2	0	15
42, 43	24	20	17	15	0

(42, 43), (25, 27)

Remove row 27 and col 27

	18	22	25, 27	42, 43
18	0	4	7	24
22	4	0	3	20
25, 27	7	3	0	15
42, 43	24	20	15	0

Next minimum is 3 . Between (22,(25,27))

	18	22, 25, 27	42, 43
18	0	4	24
22, 25, 27	4	0	15
42, 43	24	15	0

(42, 43), (((25, 27), 22), 18)

	18, 22, 25, 27	42, 43
18, 22, 25, 27	0	15
42, 43	15	0

	18, 22, 25, 27, 42, 43
18, 22, 25, 27, 42, 43	0

	18	22	25, 27	42, 43
18	0	4	7	24
22	4	0	3	20
25, 27	7	3	0	15
42, 43	24	20	15	0

(42, 43), ((25, 27), 22)

Remove row 22 and col 22

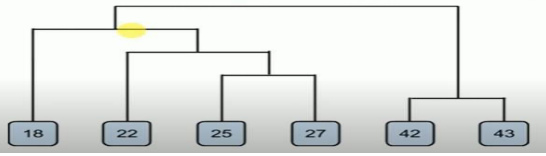
	18	22, 25, 27	42, 43
18	0	4	24
22, 25, 27	4	0	15
42, 43	24	15	0

Next minimum is 4 .Between (18 ,(22,(25,27))

Remove row 18 & col 18

- Dendrogram

((42, 43), (((25, 27), 22), 18))



GRADIENT DESCENT OPTIMIZATION

Optimization

- Optimization is the process of maximizing or minimizing a real function by systematically choosing input values
- It refers to usage of specific methods to determine the best solution from all feasible solutions.
- Three components of an optimization problem: objective function (minimization or maximization), decision variables and constraints

The three pillars of Data Science are Linear Algebra, Statistics and Optimization

Gradient Descent

- An optimization algorithm for finding a local minimum of a differentiable function.
- Used when training a machine learning model.
- ***Used to find the values of a function's parameters (coefficients) that minimize a cost function as far as possible.***

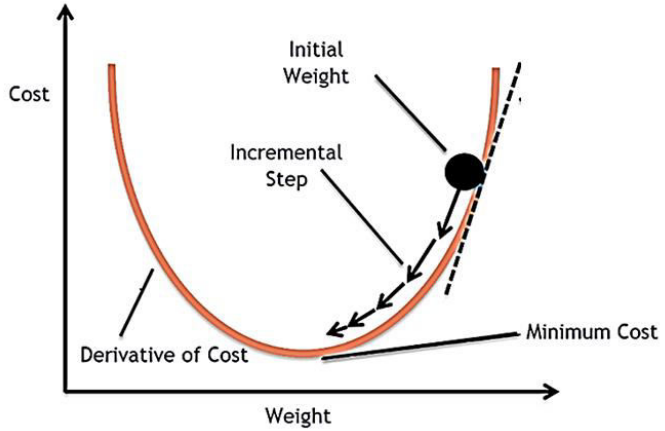
Gradient

- ***Slope of a function***
- ***A gradient measures how much the output of a function changes if you change the inputs a little bit. —Lex Fridman (MIT)***
- ***In ML terms -*** measures the change in all weights with regard to the change in error.
  - The higher the gradient, the steeper the slope and the faster a model can learn.
  - But if the slope is zero, the model stops learning.
  - In mathematical terms, a gradient is a partial derivative with respect to its inputs.

Gradient Descent

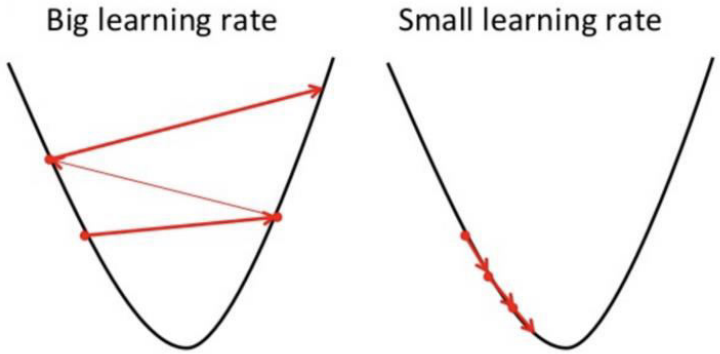
The three parameters associated with Gradient Descent are

- Starting point, first derivative and learning rate are required to determine the solution for a particular iteration.
- The learning rate (α) determines the convergence (i.e. the number of iterations required to reach the local minimum). It should neither be too small nor too large
- Very small α leads to very slow convergence and a very large value leads to oscillations around the minima or may even lead to divergence.



Learning rate

- Should be set to an appropriate value, which is neither too high or too low



Gradient Descent

- Let $f(X)$ denote the objective function and X_0 denote the starting point. In iteration k, the best point is given by

$$X_k = X_{k-1} - \alpha G_{k-1}$$

where α is the learning rate (step length) and
$G_{k-1} = \nabla f(X) = f'(X)$ is the derivative of $f(X)$

- Consider for example, $f(X) = x_1 + 2x_1^2 + 2x_1x_2 + 3x_2^2$, $\alpha = 0.1$ and

$$X_0 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}.$$

In this case,

$$f'(X) = \begin{bmatrix} 1 + 4x_1 + 2x_2 \\ 2x_1 + 6x_2 \end{bmatrix}.$$

Gradient Descent

- In the first iteration, the direction G_0 and the best point X_1 are estimated as follows:

$$G_0 = f'(X_0) = \begin{bmatrix} 4 \\ 4 \end{bmatrix} \quad \text{and} \quad X_1 = X_0 - \alpha G_0 = \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix}.$$

- Similarly, In the next iteration,

$$G_1 = f'(X_1) = \begin{bmatrix} 1.6 \\ 0.8 \end{bmatrix} \quad \text{and} \quad X_2 = X_1 - \alpha G_1 = \begin{bmatrix} -0.06 \\ 0.02 \end{bmatrix}.$$

- The iterations continue till convergence. The parameter α plays a significant role in both convergence and stability.

Gradient Descent

- In the first iteration, the direction G_0 and the best point X_1 are estimated as follows:

$$G_0 = f'(X_0) = \begin{bmatrix} 4 \\ 4 \end{bmatrix} \quad \text{and} \quad X_1 = X_0 - \alpha G_0 = \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix}.$$

- Similarly, In the next iteration,

$$G_1 = f'(X_1) = \begin{bmatrix} 1.6 \\ 0.8 \end{bmatrix} \quad \text{and} \quad X_2 = X_1 - \alpha G_1 = \begin{bmatrix} -0.6 \\ 0.02 \end{bmatrix}.$$

- The iterations continue till convergence. The parameter α plays a significant role in both convergence and stability.

Gradient Descent

Apply gradient descent approach to minimize the function:

$$f(X) = 4\,x_1^2 + 3\,x_1x_2 + 2.5\,x_2^2 - 5.5\,x_1 - 4\,x_2.$$

Assume the step size is 0.135 and the starting point is

$$X_0 = \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}.$$

Let the stopping criteria be *the absolute difference between the function values in successive iterations less than 0.005.*

Gradient Descent

Solution:

Given: $f(X) = 4\,x_1^2 + 3\,x_1x_2 + 2.5\,x_2^2 - 5.5x_1 - 4x_2$, $\alpha = 0.135$ and $X_0 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$

The stopping condition is

$$|f(X_k) - f(X_{k-1})| < 0.005$$

The search direction is the gradient of $f(X)$:

$$f'(X) = \begin{bmatrix} 8\,x_1 + 3\,x_2 - 5.5 \\ 3\,x_1 + 5\,x_2 - 4 \end{bmatrix}$$

The update equation is

$$X_k = X_{k-1} - \alpha\, f'(X_{k-1})$$

Solution:

Iteration Count k	X_{k-1}	$f'(X_{k-1})$	X_k	$f(X_k)$
1	$\begin{bmatrix}2\\2\end{bmatrix}$	$\begin{bmatrix}16.50\\12.00\end{bmatrix}$	$\begin{bmatrix}-0.2275\\0.3800\end{bmatrix}$	0.039925
2	$\begin{bmatrix}-0.2275\\0.3800\end{bmatrix}$	$\begin{bmatrix}-6.1800\\-2.7825\end{bmatrix}$	$\begin{bmatrix}0.6068\\0.7556\end{bmatrix}$	-2.0841
3	$\begin{bmatrix}0.6068\\0.7556\end{bmatrix}$	$\begin{bmatrix}1.6213\\1.5986\end{bmatrix}$	$\begin{bmatrix}0.3879\\0.5398\end{bmatrix}$	-2.3342
4	$\begin{bmatrix}0.3879\\0.5398\end{bmatrix}$	$\begin{bmatrix}-0.7771\\-0.1371\end{bmatrix}$	$\begin{bmatrix}0.4928\\0.5583\end{bmatrix}$	-2.3675
5	$\begin{bmatrix}0.4928\\0.5583\end{bmatrix}$	$\begin{bmatrix}0.1177\\0.2702\end{bmatrix}$	$\begin{bmatrix}0.4769\\0.5219\end{bmatrix}$	-2.3732
6	$\begin{bmatrix}0.4769\\0.5219\end{bmatrix}$	$\begin{bmatrix}-0.1188\\0.0401\end{bmatrix}$	$\begin{bmatrix}0.4930\\0.5164\end{bmatrix}$	-2.3745

X(k-1)		f'(k-1)		alpha*f'(k-1)		X_k=x(k-1)-alpha*f1(k-1)		$f(x_k)$	Diff
2	2	16.5	12	2.2275	1.62	-0.2275	0.38	0.03992	
-0.2275	0.38	-6.18	-2.7825	-0.8343	-0.3756375	0.6068	0.7556	-2.08409	-2.1240
0.6068	0.7556	1.62131	1.5986	0.2189	0.2158	0.3879	0.5398	-2.3341	-0.2501
0.3879	0.5398	-0.77713	-0.13709	-0.1049	-0.01851	0.4928	0.5583	-2.367541	-0.03336
0.4928	0.5583	0.11769	0.27018	0.01589	0.03647	0.4769	0.5219	-2.373191	-0.00565
0.4769	0.5219	-0.1188	0.04014	-0.0160	0.00542	0.492990702	0.5164	-2.37447	-0.00128

Solution:

Comparing the function values obtained in 5th and 6th iterations, their absolute difference is less than 0.005. Therefore, the optimal solution is

$$X_{opt} = \begin{bmatrix}0.4930\\0.5164\end{bmatrix}.$$

Additional information: If either the stopping condition threshold is lesser (say 0.0001) or the number of iterations is more (say 10), the optimal solution will be

$$X_{opt} = \begin{bmatrix}0.5\\0.5\end{bmatrix}.$$

Variants of Gradient Descent:

- There are three variants of gradient descent based on the amount of data (samples) considered for computing the gradient at each iteration
  1. Batch Gradient Descent
  2. Mini-Batch Gradient Descent
  3. Stochastic Gradient Descent

Types of gradient descent

- Batch gradient descent
  - Also called vanilla gradient descent
  - Calculates the error for each example within the training dataset, but only after all training examples have been evaluated does the model get updated.
  - This whole process is like a cycle and it's called a training epoch.
- Advantages
  - Computationally efficient
  - Produces a stable convergence
- Disadvantages
  - Requires the entire training dataset be in memory and available to the algorithm

Stochastic Gradient Descent (SGD)

- SGD does the update for each training example within the dataset, meaning it updates the parameters for each training example one by one.
- Advantages:
  - frequent updates allow a detailed rate of improvement.
- Disadvantages:
  - more computationally expensive than the batch gradient descent approach
  - Frequent updates may result in noisy gradients, which may cause the error rate to jump around instead of slowly decreasing.

Mini-Batch Gradient Descent (SGD)

- combination of the concepts of SGD and batch gradient descent.
- splits the training dataset into small batches and performs an update for each of those batches.
- The go-to algorithm used when training a neural network
- The most common type of gradient descent within deep learning.

Momentum:
Momentum is a variant of gradient descent that incorporates information from the previous weight updates to help the algorithm converge more quickly to the optimal solution. Momentum adds a term to the weight update that is proportional to the running average of the past gradients, allowing the algorithm to move more quickly in the direction of the optimal solution.

Gradient Descent in Linear Regression

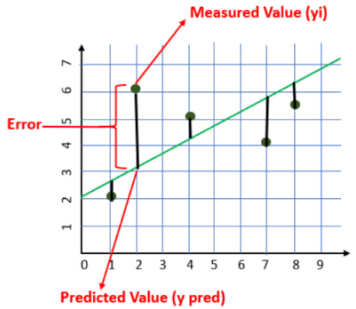
A linear regression model attempts to explain the relationship between a dependent (output variables) variable and one or more independent (predictor variable) variables using a straight line.
This straight line is represented using the following formula:

y=mx+c

Where, y: dependent variable
x: independent variable
m: Slope of the line (For a unit increase in the quantity of X, Y increases by m.1 = m units.)
c: y intercept

Cost Function

The cost is the error in our predicted value. We will use the Mean Squared Error function to calculate the cost.



$$Cost\ Function(MSE) = \frac{1}{n}\sum_{i=0}^{n}(y_i - y_{i\,pred})^2$$

Replace $y_{i\,pred}$ with $mx_i + c$

$$Cost\ Function(MSE) = \frac{1}{n}\sum_{i=0}^{n}(y_i - (mx_i + c))^2$$

Our goal is to minimize the cost as much as possible in order to find the best fit line. We are not going to try all the permutation and combination of m and c (inefficient way) to find the best-fit line. For that, we will use Gradient Descent Algorithm.

Step by Step Algorithm:

1. Let m = 0 and c = 0. Let L be our learning rate. It could be a small value like 0.01 for good accuracy.

2. Calculate the partial derivative of the Cost function with respect to m. Let partial derivative of the Cost function with respect to m be D_m (With little change in m how much Cost function changes).

$$D_m = \frac{\partial(Cost\ Function)}{\partial m} = \frac{\partial}{\partial m}\left(\frac{1}{n}\sum_{i=0}^{n}(y_i - y_{i\,pred})^2\right)$$

$$= \frac{1}{n}\frac{\partial}{\partial m}\left(\sum_{i=0}^{n}(y_i - (mx_i + c))^2\right)$$

$$= \frac{-2}{n}\sum_{i=0}^{n}x_i(y_i - (mx_i + c))$$

$$= \frac{-2}{n}\sum_{i=0}^{n}x_i(y_i - y_{i\,pred})$$

Step by Step Algorithm:

Similarly, let's find the partial derivative with respect to c. Let partial derivative of the Cost function with respect to c be D_c (With little change in c how much Cost function changes).

$$D_c = \frac{\partial(Cost\ Function)}{\partial c} = \frac{\partial}{\partial c}\left(\frac{1}{n}\sum_{i=0}^{n}(y_i - y_{i\,pred})^2\right)$$

$$= \frac{1}{n}\frac{\partial}{\partial c}\left(\sum_{i=0}^{n}(y_i - (mx_i + c))^2\right)$$

$$= \frac{-2}{n}\sum_{i=0}^{n}(y_i - (mx_i + c))$$

$$\frac{-2}{n}\sum_{i=0}^{n}(y_i - y_{i\,pred})$$

Step by Step Algorithm:

3. Now update the current values of m and c using the following equation:

$$m = m - LD_m$$

$$c = c - LDc$$

4. We will repeat this process until our Cost function is very small (ideally 0).

Gradient Descent Algorithm gives optimum values of m and c of the linear regression equation. With these values of m and c, we will get the equation of the best-fit line and ready to make predictions.

Problems in gradient descent

Vanishing Gradient
Exploding Gradient
Saddle Point

Vanishing Gradient

If the range of the initial values for the weights is not carefully chosen, and the range of the values of the weights during training is not controlled, a vanishing gradient would occur which is the main hurdle to learning.

The reason for vanishing gradient is that, during backpropagation, the gradients of inceptive layers are obtained by multiplying the gradients of concluding layers. So, let's say if the gradients of the concluding layers are less than one, their multiplication vanishes very fast.

Exploding Gradient

Exploding gradient occurs when the derivatives increase as we go backward with every layer during backpropagation. This is the exact opposite of the vanishing gradients.

The architecture becomes unstable due to large changes in loss at each update step. The weights grow exponentially and become very large, and derivatives stabilize.

Saddle Point

Saddle point on the surface of loss function is a point where, from one perspective, that critical point looks like a local minima, while from another perspective, it looks like a local maxima.

Saddle point injects confusion into the learning process. Model learning stops (or becomes extremely slow) at this point, thinking that “minimum” has been achieved since the slope becomes zero

The approaches to handle these problems:

Changing the architecture
Careful Weight Initialization

References

- https://builtin.com/data-science/gradient-descent
- https://www.youtube.com/watch?v=rIVLE3condE
- https://www.analyticsvidhya.com/blog/2021/04/gradient-descent-in-linear-regression/
- https://ruder.io/optimizing-gradient-descent/index.html#momentum
- https://distill.pub/2017/momentum/
- https://medium.com/optimization-algorithms-for-deep-neural-networks/gradient-descent-with-momentum-dce805cd8de8
- https://blog.paperspace.com/intro-to-optimization-momentum-rmsprop-adam/

Managing Health and Safety

Objectives

- To be aware of various hazards that they may come across at workplace
- To prevent/handle any accidents or emergencies
- To know about the defined health, safety and security measures that should be followed at the time of occurrence of such unpredictable events.

Performance Criteria

- PC1. Comply with your organization's current health, safety and security policies and procedures
- PC2. Report any identified breaches in health, safety, and security policies and procedures to the designated person
- PC3. Identify and correct any hazards that you can deal with safely, competently and within the limits of your authority
- PC4. Report any hazards that you are not competent to deal with to the relevant person in line with organizational procedures and warn other people who may be affected
- PC5. Follow your organization's emergency procedures promptly, calmly, and efficiently
- PC6. Identify and recommend opportunities for improving health, safety, and security to the designated person
- PC7. Complete any health and safety records legibly and accurately

Topics

- 1. Workplace safety
- 2. Reporting accidents and emergencies
- 3. Protecting health and safety as you work

Workplace Safety

- Workplace safety refers to the working environment at a company and encompasses all factors that impact the safety, health, and well-being of employees.
  - Environmental hazards, unsafe working conditions or processes, drug and alcohol abuse, and workplace violence.

https://www.inc.com/encyclopedia/workplace-safety.html

Why is it important?

- A safe work environment is a productive one.
  - Safety measures protect employees as well as equipment and business property.
  - Avoiding or minimizing injuries and damage to equipment and facilities will result in fewer expenses and more profit for a business.

https://smallbusiness.chron.com/workplace-safety-43459.html

Workplace Safety Guidelines

- **Fire Safety**
  Employees should be aware of all emergency exits, including fire escape routes, of the office building and also the locations of fire extinguishers and alarms.
- **Falls and Slips**
  To avoid falls and slips, all things must be arranged properly. Any spilt liquid, food or other items such as paints must be immediately cleaned to avoid any accidents. Make sure there is proper lighting and all damaged equipment, stairways and light fixtures are repaired immediately.
- **First Aid**
  Employees should know about the location of first-aid kits in the office. First-aid kits should be kept in places that can be reached quickly. These kits should contain all the important items for first aid, for example, all the things required to deal with common problems such as cuts, burns, headaches, muscle cramps, etc.
- **Security**
  Employees should make sure that they keep their personal things in a safe place.
- **Electrical Safety**
  Employees must be provided basic knowledge of using electrical equipment and common problems. Employees must also be provided instructions about electrical safety such as keeping water and food items away from electrical equipment. Electrical staff and engineers should carry out routine inspections of all wiring to make sure there are no damaged or broken wires.

Report Accidents and Emergencies

- Accident
  - An unplanned, uncontrolled, or unforeseen event resulting in injury or harm to people and damages to goods.
  - Eg: a person falling down and getting injured or a glassware item that broke upon being knocked over.
- Emergency
  - A serious or crisis situation that needs immediate attention and action.
  - Eg: a customer having a heart attack or sudden outbreak of fire in your organization needs immediate attention.

Guidelines - Reporting Accidents and Emergencies

- **Notice and correctly identify accidents and emergencies**:
  - need to be aware of what constitutes an emergency and what constitutes an accident in an organization
  - also be aware of the procedures to tackle each form of accident and emergency

Guidelines - Reporting Accidents and Emergencies

- **Get help promptly and in the most suitable way**
  - Follow the procedure for handling a particular type of accident and emergency.
  - Promptly act as per the guidelines
  - Do not act outside the guidelines and policies laid down for your role even if your actions are motivated by the best intention.
  - Remember that only properly trained and certified professionals may be authorized to take decisions beyond the organization's policies and guidelines, if the situation requires.

Guidelines - Reporting Accidents and Emergencies

- **Follow company policies and procedures for preventing further injury while waiting for help to arrive**:
  - If someone is injured, do not act as per your impulse or gut feeling.
  - Go as per the procedures laid down by your organization's policy for tackling injuries.
  - You need to stay calm and follow the prescribed procedures.
  - If you panic or act outside the prescribed guidelines, you may end up further aggravating the emergency situation or putting the injured person into further danger. You may even end up injuring yourself.

Guidelines - Reporting Accidents and Emergencies

- **Act within the limits of your responsibility and authority when accidents and emergencies arise**
  - Provide help and support within your authorized limit.
  - Provide medical help to the injured only if you are certified to provide the necessary aid. Otherwise, wait for the professionals to arrive and give necessary help.
  - In case of emergencies also, act within your authorized limits and let the professionals do the task allocated to them.
  - Do not attempt to handle any emergency situation for which you do not have formal training or authority. You may end up harming yourself and the people around you.

Guidelines - Reporting Accidents and Emergencies

- **Promptly follow instructions given by senior staff and the emergency services**
  - Provide necessary services as described by the organization's policy for your role.
  - Follow the instructions of senior staff that are trained to handle particular situations. Work under their supervision when handling accidents and emergencies.

Types of Accidents

- Trip and Fall
- Slip and Fall
- Injuries caused due to escalators or elevators (or lifts)
- Accidents due to falling of goods
- Accidents due to moving objects (Eg: trolleys)

Handling Accidents

- **Attend to the injured person immediately** Depending on the level and seriousness of the injury, see that the injured person receives first aid or medical help at the earliest.
- **Inform your supervisor** about the accident giving details about the probable cause of accident and a description of the injury
- **Assist your supervisor** in investigating and finding out the actual cause of the accident.

Types of Emergencies

- Medical emergencies, such as heart attack or an expectant mother in labor
- Substance emergencies, such as fire, chemical spills, and explosions
- Structural emergencies, such as loss of power or collapsing of walls
- Security emergencies, such as armed robberies, intruders, and mob attacks or civil disorder
- Natural disaster emergencies, such as floods and earthquakes

Handling Emergencies

- you should be aware of:
  - What is the evacuation plan and procedure to follow in case of an emergency?
  - Who all should you notify within the organization?
  - Which external agencies, such as police or ambulance, you should notify in which emergency?
  - What all services and equipment should you shut down during which emergency?
-

Handling Emergencies

- Keep a list of numbers to call during emergency, such as those of police, fire brigade, security, ambulance etc. Ensure that these numbers are fed into the organizations telephone program and hard copies of the numbers are placed at strategic locations in the organization.
- Regularly check that all emergency handling equipments are in working condition, such as the fire extinguisher and fire alarm system.
- Ensure that emergency exits are not obstructed and keys to such exists are easily accessible. Never place any objects near the emergency doors or windows

Protect Health and Safety as you work

- Hazard
  - any source of potential harm or danger to someone or any adverse health effect produced under certain condition
  - A hazard can harm an individual or an organization.
    - For example, hazard to an organization include loss of property or equipment while hazard to an individual involve harm to health or body.

Some potential sources of hazards

- **Using computers**
  - poor sitting postures
  - excessive duration of sitting in one position
  - glare from the computer screen

  Stretching up at regular intervals or doing some simple yoga in your seat only can mitigate such hazards.

Some potential sources of hazards

- **Handling office equipment**
  - Improper handling of office equipment can result in injuries.
  - Eg: sharp-edged equipment if not handled properly can cause cuts.
  - Staff members should be trained to handle equipment properly. Relevant manual should be made available by administration on handling equipment.

Some potential sources of hazards

- **Handling objects**
  - Lifting or moving heavy items without proper procedure or techniques can be a source of potential hazard.
  - Always follow approved procedure and proper posture for lifting or moving objects.

Some potential sources of hazards

- **Stress at work**
  - Long working hours can be stressful
  - Aggressive conflicts or arguments with colleagues

  Always look for ways for conflict resolution with colleagues.

  Have some relaxing hobbies for stress against long working hours.

Some potential sources of hazards

- **Working environment**
  - Poor ventilation, inappropriate height chairs and tables, stiffness of furniture, poor lighting,
  - staff unaware of emergency procedures, or poor housekeeping.
  - Hazards may also include physical or emotional intimidation, such as bullying or ganging up against someone.

How do you identify health and safety problems?

- Do you or your co-workers have injuries or health complaints? If so, what types?
- Who has been hurt or is having symptoms?
- When do you or your co-workers feel these symptoms?
- Where in the workplace are safety or health problems occurring?
- What are the conditions that are causing problems?

CAUTION: Health Hazards

Common types of health hazards in the workplace are:

- **Chemical** (asbestos, solvents, chlorine)
- **Biological** (tuberculosis, HIV, hepatitis, molds)
- **Physical** (noise, heat and cold, radiation, vibration)
- **Ergonomics or Repetitive Strain Injuries** (carpal tunnel syndrome, back injuries)
- **Psychological** (stress)

How health hazards enter your body:

- Breathing (**inhalation**)
- Swallowing (**ingestion**)
- Skin (**absorption**)
- Cuts (**injection**)

Harm caused by health hazards depends on:

- Strength, or potency, of the agent.
- Amount of the agent that is present.
- How long you are exposed to the agent.
- Part of your body that is exposed.

Types of health effects:

- **Acute:** the effect shows up right away.
- **Chronic:** problems show up after a long period of exposure and/or long after the exposure ends.
- **Local:** only the part of the body that was exposed is affected.
- **Systemic:** an agent enters the body and affects other parts of the body.

Cancer

- Cancer is a term for many diseases in different parts of the body.
- Carcinogens are agents that cause cancer.
- There is no totally safe level of exposure to something that causes cancer.
- Cancer from a workplace exposure may develop 10, 20 or more years after exposure.

Sensitization

- You may become allergic or sensitive to some agents you work with. Sensitization can develop over time.
- For example, a health care worker may develop a serious allergic reaction to latex used in gloves.

Reproductive effects

- Both men and women can be affected by reproductive hazards at work.
- Reproductive hazards cause miscarriages and birth defects.

CAUTION: Safety Hazards

Common types of safety hazards in the workplace are:

- Slips, trips and falls
- Being caught in or struck by moving machinery or other objects
- Fire and explosions
- Transportation and vehicle-related accidents
- Confined spaces
- Violence

Slips, Trips and Falls

- Bad housekeeping and poor drainage can make floors and other walking surfaces wet and slippery.
- Electrical wires along the floor pose a tripping hazard.
- You can fall if you are not provided with fall protection equipment, guardrails, and safe ladders.

Caught In or Struck By Moving Machinery/Objects

- Machinery can cause injuries in different ways:
- You can get parts of your body caught in or struck by exposed moving parts if machines are not properly guarded, or not locked out when being repaired.
- You can be struck by flying objects from machines without protective guards.

Fire and Explosions

- Improper labeling, handling or storage of certain materials can pose a risk of fire or explosion.
- Every workplace should have an evacuation plan for getting people out of a building in case of fire and an alarm or alert system to quickly inform employees of an emergency.
- Every worker should be trained on what to do in case of an emergency.

Transportation and Vehicle-Related Accidents

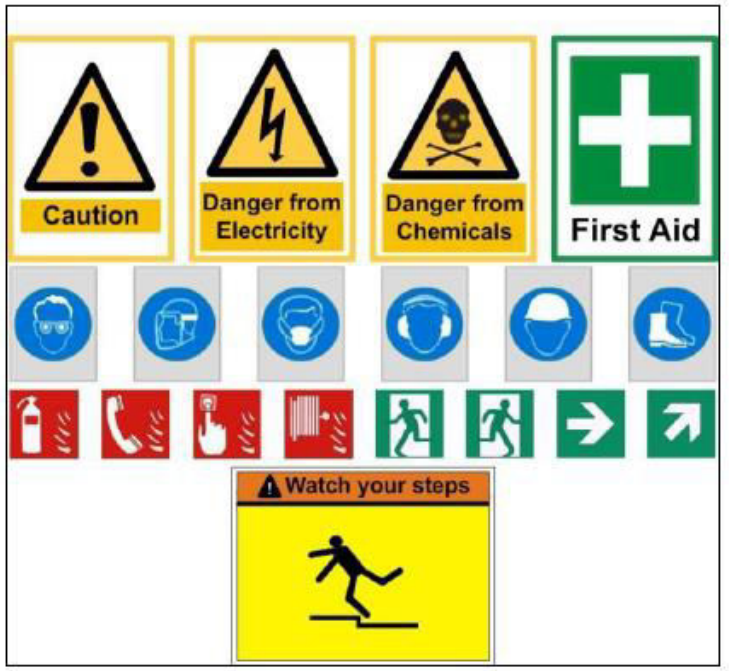
- Operators of vehicles and equipment can be injured or cause injury to pedestrians if equipment is unsafe or if adequate training has not been provided.
- You can be seriously injured or killed after being hit by a vehicle while repairing roads or doing other work in traffic zones. This danger exists when traffic is not properly routed and/or adequate barriers are not placed between the workers and the traffic.

Confined Spaces

- A confined space is an area with small openings for a worker to enter and exit and is not designed for regular work. Examples of confined spaces include manholes, sewer digestors and silos. There are many hazards in confined spaces.
- Workers can become unconscious and die from a lack of oxygen.
- There may be too much oxygen, or other chemicals that can catch fire or explode.
- Poisonous gases and vapors, such as hydrogen sulfide or carbon monoxide, may also build up in a confined space.
- Confined spaces can also pose physical hazards. They can be very hot or cold, very loud, or slippery and wet.
- Grain, sand or gravel can bury a worker.

Violence

- Violence on the job is a growing problem.
- Homicides are the second leading cause of workplace fatalities. Workplace violence includes physical assault as well as near misses, verbal abuse and sexual harassment.



Healthy Living

References

- https://www.osha.gov/dte/outreach/intro_osha/7SafetyHealthProbsWorkplace.pdf
- https://www.firesafe.org.uk/health-and-safety-safety-signs-and-signals-regulations-1996/

Data and Information Management

Objectives

- To analyze data and present it in a suitable format, as is suitable for the given process or organization
- To understand the process of standardized reporting and the nuances of a publishing a report with a specified end objective in mind

Performance Criteria

- PC1. Establish and agree with appropriate people the data/information you need to provide, the formats in which you need to provide it, and when you need to provide it
- PC2. Obtain the data/information from reliable sources
- PC3. Check that the data/information is accurate, complete and up-to-date
- PC4. Obtain advice or guidance from appropriate people where there are problems with the data/information
- PC5. Carry out rule-based analysis of the data/information, if required
- PC6. Insert the data/information into the agreed formats
- PC7. Check the accuracy of your work, involving colleagues where required
- PC8. Report any unresolved anomalies in the data/information to appropriate people
- PC9. Provide complete, accurate and up-to-date data/information to the appropriate people in the required formats on time

Topics

1. Knowledge Management
2. Standardized reporting and compliance
3. Decision Models

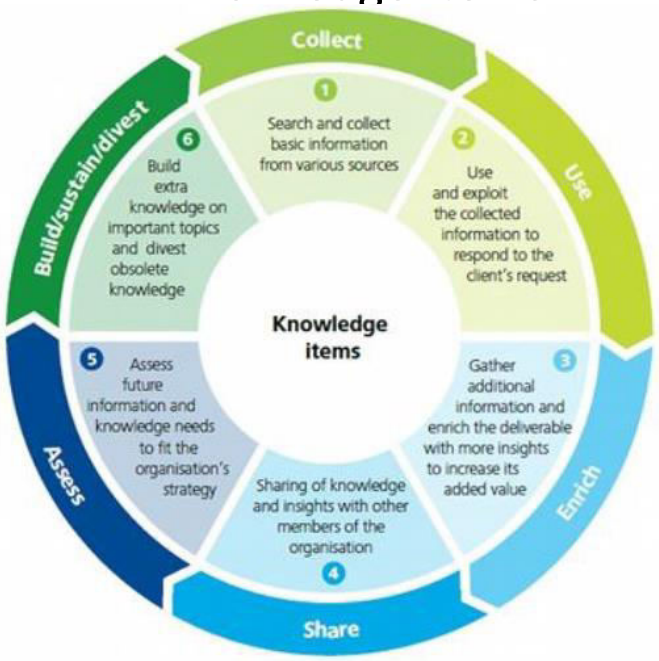
Knowledge Management

- **Knowledge management** (KM) is the process of capturing, developing, sharing, and effectively using organizational **knowledge**.
- It refers to a multi-disciplinary approach to achieve organizational objectives by making the best use of **knowledge**.

Knowledge Management

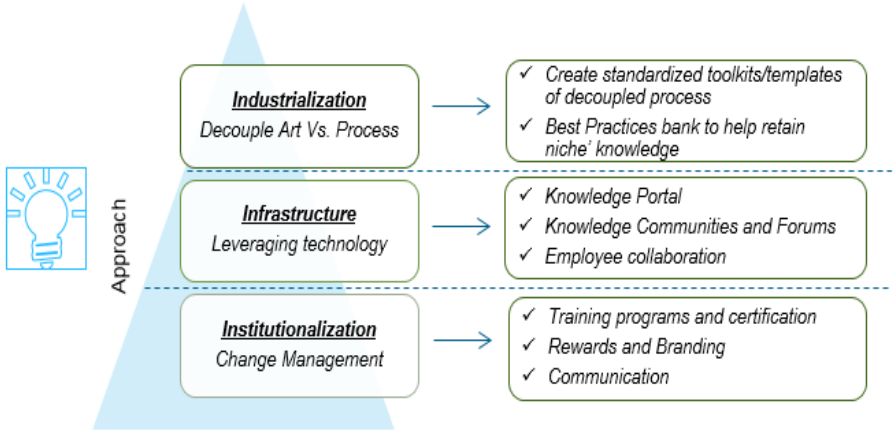
- KM to support an organization's growth engine :
  - Capture uniqueness of each project in new growth and improvise existing work
  - Decouple the Art with Process and make complex work scalable
  - Reduce people dependencies
  - Decrease time to innovate/deliver through faster knowledge distribution
-

Knowledge Items

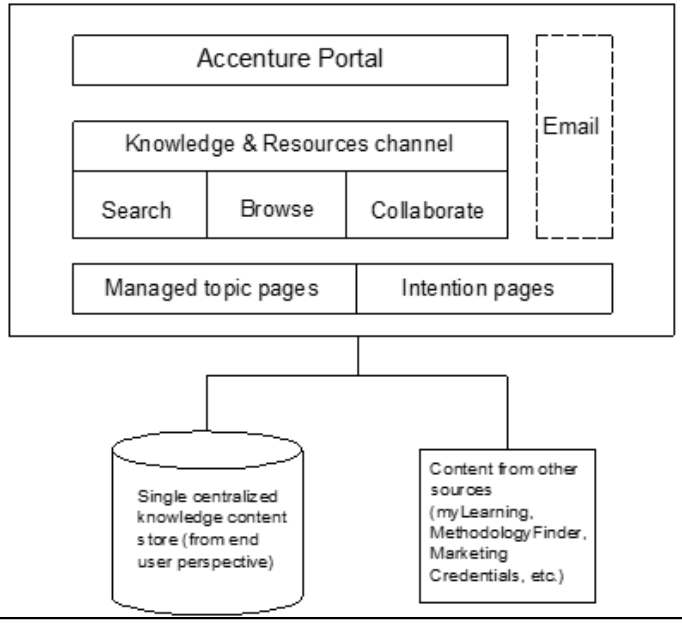


Trends	Examples/Implications
Crowdsourcing and social media playing a big role in collaborative product development	• Fiat Mio is the world's first crowd-sourced car, developed based on new ideas sourced from users via a social media platform. • While this trend will change the way products will be designed in the future, organizations will also have to collate, classify, and assess large amount of data, and feed it back into the product development process.
Big data analytics close looping with knowledge base	• Many product development organizations, including auto OEMs, are critically analyzing Big Data (product failure data, service data, warranty data, historical design data, materials data etc.) to extract information patterns • CAD/CAE/PLM technologies now have capabilities to integrate various extracted knowledge elements into the product design process.
Internal and external collaboration for knowledge sharing	• Many leading corporates in North America and Europe have developed collaboration platforms with features including blogs, wikis, chats, communities of practice, expert corners and alerts among others to allow employees to share knowledge through discussion threads, idea voting etc. • While collaboration platforms are becoming extremely popular within organizations, in some cases they are also being extended to external users with appropriate security features. • Organizations will need to have business processes and supporting infrastructure in place to enable such collaboration platforms and process the information shared on these platforms.
Use of telematics in assimilating knowledge for product development	• Sensing devices integrated with remote monitoring devices are being used to gather intelligence on product performance, health diagnostics, usage patterns etc. • Manufacturing organizations must work on integrating the knowledge and data generated in this way into the core product development process to enable faster and right decision making.
Baking knowledge into the product development process	• Manufacturing organizations the world over are realizing that tacit knowledge available among current employees should be formalized and captured for posterity. • Global auto OEMs are gearing up to put business processes and the required infrastructure in place to capture this tacit knowledge in terms of best practices/lessons learned etc., and close-loop the knowledge management process by integrating these best practices

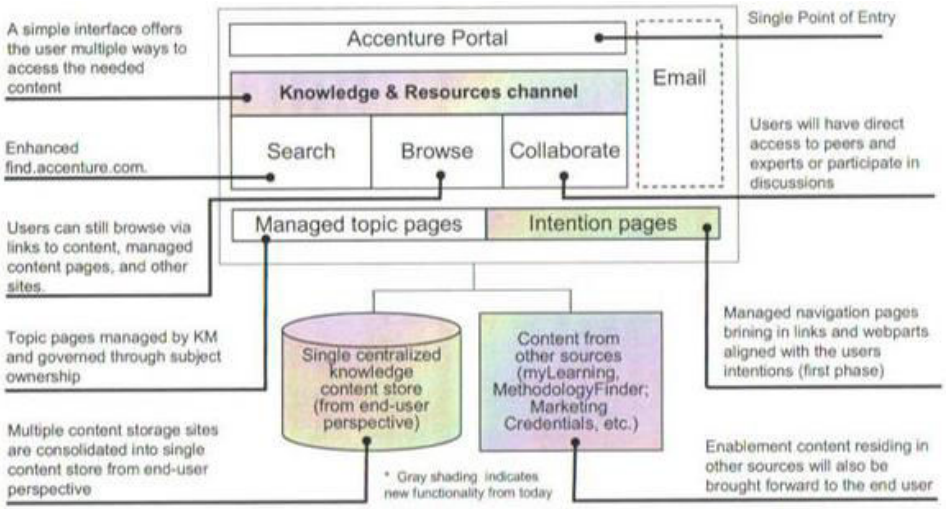
Knowledge Management Process



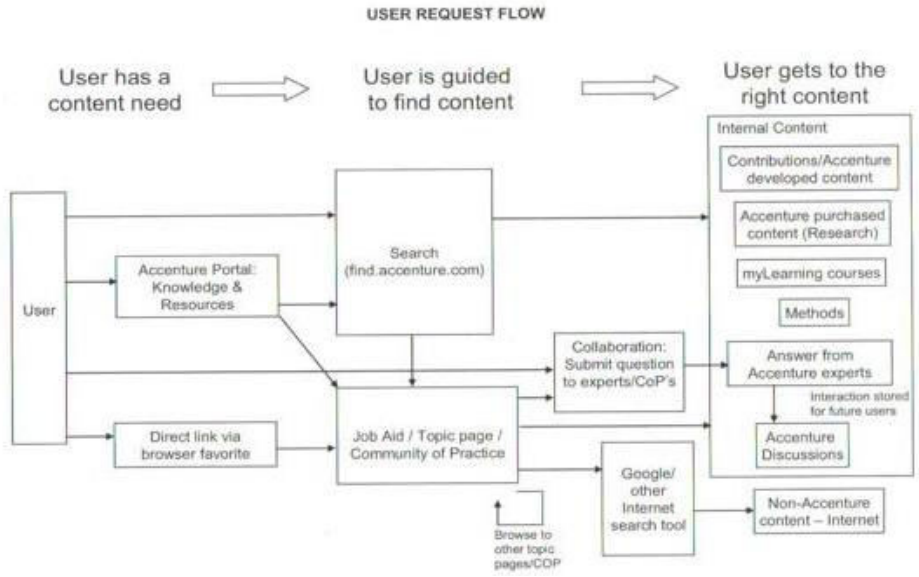
Accenture KM Solution Structure



Accenture KM portal



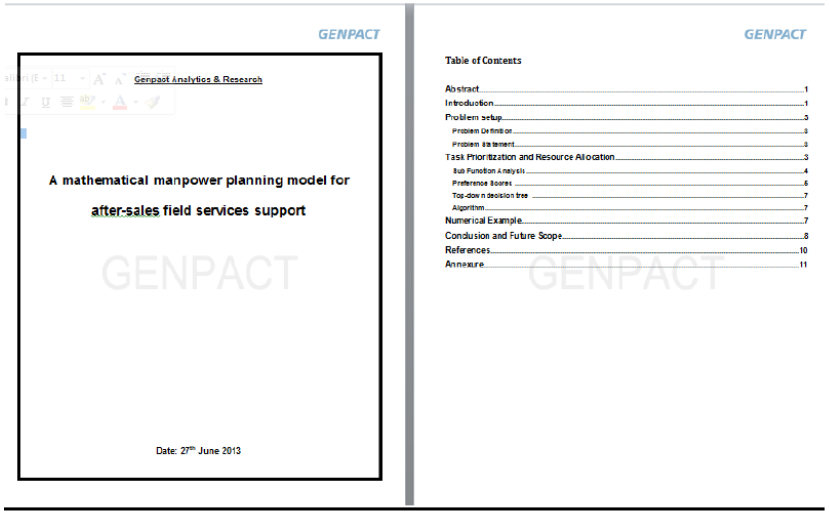
Accenture KM Report Extraction



Standardized Reporting and Compliance

- Template?
  - Pre-created structures based on which reports are to be created
- Types
  - Financial reporting templates
  - Marketing and sales reporting templates
  - Data entry templates
  - Research templates
  - Pricing and product costing templates
  - Any other reporting or data presentation requirements

Research Template


GENPACT

Genpact Analytics & Research

A mathematical manpower planning model for after-sales field services support

Date: 27th June 2013

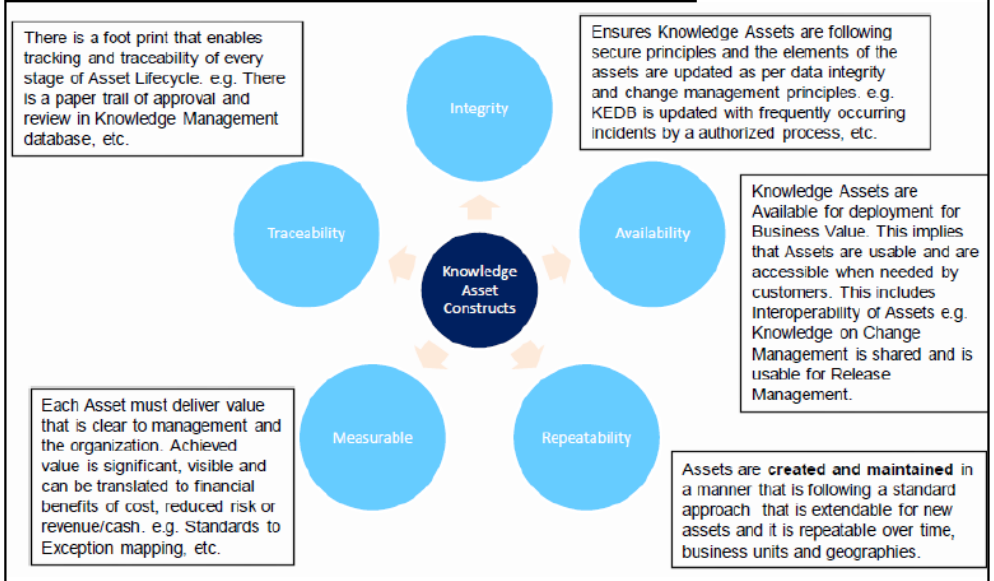
GENPACT

Table of Contents

Organizing data/information

- The five major mechanisms leading to high quality knowledge content are:
  - Standardized content formats
  - A clearly specified knowledge content production process
  - Informal or formal peer review assuring that the document knowledge is valid & relevant
  - Information quality criteria
  - Guidelines – specifying minimal requirements in terms of document content, style, size & ownership and format
-

Policies and procedures for recording and sharing information



What is Compliance?

- Conforming to a rule
  - such as a specification, policy, standard or law.

- Regulatory compliance describes the goal that organizations aspire to achieve in their efforts to ensure that they are aware of and take steps to comply with relevant laws and regulations.

Why is Compliance important?

- Reduce many of the company's greatest risks
- Reduce the severity of claims and penalties when violations of law occur despite the program
- Enhance company performance and profitability

Decision Models

- The **Decision Model** is an intellectual template for perceiving, organizing, and managing the business logic behind a business **decision**.
  - Business Logic - set of business rules represented as atomic elements of conditions leading to conclusions
- Decision Models are used to model a decision being made once as well as to model a repeatable decision-making approach that will be used over and over again.

The Vroom-Yetton-Jago Decision Model

- Decision Making is affected by three main factors:
  - Decision quality
    - How important is to have right solution?
    - Higher the quality, more the people involved
  - Subordinate commitment
    - How important is your team and others involve
    - Should increase participation levels
  - Time constraints
    - How much do you have to make the decision?
    - More time, more you have luxury of including others

https://en.wikipedia.org/wiki/Vroom%E2%80%93Yetton_decision_model#:~:text=The%20Vroom%E2%80%93Yetton%20contingency%20model,is%20contingent%20to%20the%20situation.

Leadership styles

- **Autocratic** - you make the decision and inform others of it.
- Process:
  - A1 -you use the information you already have and make the decision
  - A2 - you ask team members for specific information and once you have it, you make the decision.

Leadership styles

- **Consultative** - you gather information from the team and other and then make the decision.
- Process:
  - C1- you inform team members of what you're doing and may individually ask opinions, however, the group is not brought together for discussion. You make the decision.
  - C2 -you are responsible for making the decision, however, you get together as a group to discuss the situation, hear other perspectives, and solicit suggestions.

Leadership styles

- **Collaborative** - you and your team work together to reach a consensus.
- Process:
  - The team makes a decision together. Your role is mostly facilitative and you help the team come to a final decision that everyone agrees on.

The Kepner-Tregoe Approach

- The goal is not to make the perfect choice, or the choice that has no defects. So the decision maker must accept some risk.
- Kepner-Tregoe Matrix helps to evaluate and mitigate the risks of your decision.

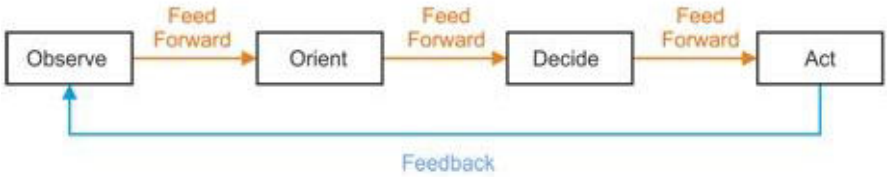
The Kepner-Tregoe Matrix

- Comprises of 4 steps
  - Situation Appraisal – identify concerns and outline the priorities.
  - Problem Analysis – describe the exact problem or issue by identifying and evaluating the causes
  - Decision Analysis – identify and evaluate alternatives by performing a risk analysis for each and then make a final decision.
  - Potential Problem Analysis – evaluate the final decision for risk and identify the contingencies and preventive actions necessary to minimize that risk.

https://www.valuebasedmanagement.net/methods_kepner-tregoe_matrix.html#:~:text=The%20Kepner%2DTregoe%20Matrix%20is,analyzing%20potential%20risks%20and%20opportunities.

OODA Loops

- **Observe** – collect current information from as many sources as practically possible.
- **Orient** – analyze this information, and use it to update your current reality.
- **Decide** – determine a course of action.
- **Act** – follow through on your decision.



https://en.wikipedia.org/wiki/OODA_loop

Stage 1 - Observe

- What's happening in the environment that directly affects me?
- What's happening that indirectly affects me?
- What's happening that may have residual affects later on?
- Were my predictions accurate?
- Are there any areas where prediction and reality differ significantly?

Stage 2 - Orient

- Cultural traditions.
- Genetic heritage.
- The ability to analyze and synthesize.
- Previous experience.
- New information coming in.

Stage 3 - Decide

- As you keep on cycling through the OODA Loop, and new suggestions keep arriving, these can trigger changes to your decisions and subsequent actions – essentially, you're learning as you continue to cycle through the steps. The results of your learning are brought in during the Orient phase, which in turn influences the rest of the decision making process.

Stage 4 - Act

- The Act stage is where you implement your decision. You then cycle back to the Observe stage, as you judge the effects of your action. This is where actions influence the rest of the cycle, and it's important to keep learning from what you, and your opponents, are doing.

Learning and Self Development

Objectives

- To give a overview on how skills and competency can be enhanced in a professional environment.
- To understand the need of skills improvement for personal and organizational growth.

Performance Criteria

- PC1. obtain advice and guidance from appropriate people to develop the knowledge, skills and competence
- PC2. identify accurately the knowledge and skills you need for their job role
- PC3. identify accurately the current level of knowledge, skills and competence and any learning and development needs
- PC4. agree with appropriate people a plan of learning and development activities to address the learning needs
- PC5. undertake learning and development activities in line with the plan
- PC6. apply the new knowledge and skills in the workplace, under supervision
- PC7. obtain feedback from appropriate people on the knowledge and skills and how effectively you apply them
- PC8. review the knowledge, skills and competence regularly and take appropriate action

Topics

- Knowledge, skills and competencies
- Training and Development
- Learning and Development policies and record keeping

Knowledge, skills and competencies

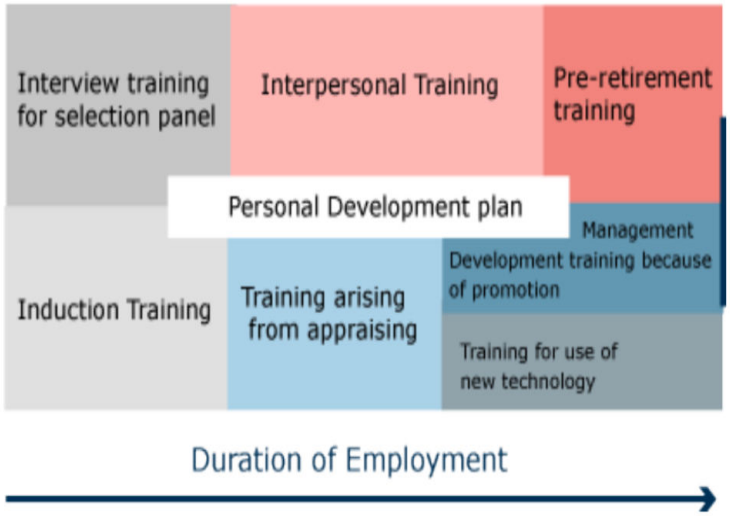
- Knowledge
  - Mastery of facts, range of information in subject matter area
- Skills
  - Proficiency, expertise, or competence in given area; e.g., science, art, crafts and engineering
- Competency
  - Demonstrated performance to use knowledge and skills when needed

What are the Skill sets you need?

Important Skills

- Interpersonal Skill
- Team Skill
- Communications
- Planning and Organizing
- Organizational Knowledge and Competence
- Problem Solving and Analytical Ability
- Judgment
- Direction and Motivation
- Decisiveness
- Self-Development
- Flexibility
- Leadership

Systematic Learning for the Individual



Importance of Knowledge, Skills & Competence (KSC)

- The primary purpose of KSC is to measure those qualities that will set one candidate apart from the others.
- KSC identify the better candidates from a group of persons basically qualified for a position.
  - How well an applicant can show that he or she matches the position's defined KSAs determines whether that person will be seriously considered for the job.

Importance of developing skills

- Many job roles are requiring formal training qualifications. Why?
  - Legislative requirements or
  - To meet the requirements of specific employers
- Developing your skills through further training provides significant benefits including
  - Increased career development opportunities
  - Personal growth
  - Increase your knowledge and understanding of the industry

Importance of developing skills
Benefits

- Increased career development opportunities
  - Experience alone, not sufficient for promotion
  - Further training and acquiring new skills is needed
- Personal Growth
  - Advantages of training includes building your networking, time management, communication and negotiation skills.
- Increase your knowledge and understanding of the industry
  - To know about current industry trends & a better perspective to approach industry problems to move to next level

Training and Development

I. Identifying Training Needs
II. Evaluation/Review of Trainings
III. Feedback

I. Identifying Training Needs

- Methods used by the organization to review skills and knowledge
  - Training need analysis
  - Skills need analysis
  - Performance appraisals

Training need analysis

- Training needs analysis involves:
  - monitoring current performance using techniques such as observation, interviews and questionnaires
  - anticipating future shortfalls or problems
  - identifying the type and level of training required and analysing how this can best be provided.

Work / Task Analysis

- Develop an understanding of what employees need to know in order to perform their jobs by conducting interviews with SME, supervisor, manager etc.
  - What tasks are performed?
  - How frequently are they performed?
  - How important is each task?
  - What knowledge is needed to perform the task?
  - How difficult is each task?
  - What kinds of training are available?

Work / Task Analysis

- Organize the identified tasks. Develop a sequence of tasks. Or list the tasks by importance.
- Observe the employee performing the job. Document the tasks being performed.
- Are there differences between high and low performing employees on specific work tasks?

Performance Analysis

- PA is used to identify which employees need the training.
- Review performance appraisals.
- Look for performance measures such as benchmarks and goals.

Performance Analysis

Sources of performance data

- Performance Appraisals
- Quotas met (un-met)
- Performance Measures
- Turnover
- Shrinkage / Leakage / Spoilage / Losses
- Accidents / Safety Incidents / Grievances
- Absenteeism
- Units per Day / Units per Week/ Returns
- Customer Complaints

II. Evaluation/Review of Trainings

- Evaluation of the impact of learning interventions may be carried out at a number of levels and involve a variety of factors:
  - Reaction
  - Learning
  - Transfer
  - Results
  - Return on Investment

III. Feedback

- Feedback is an essential mean to understand and identify the right trainings & knowledge needed for

Sl no	Questions	Agree	Neutral	Disagree
1	The objectives of the training were clearly defined			
2	Participation and interaction were encouraged			
3	The topics covered were relevant to me			
4	The content was organized and easy to follow			
5	The materials distributed were helpful			
6	This training experience will be useful in my work			
7	The trainer was knowledgeable about the training topics			
8	The trainer was well prepared			
9	The training objectives were met			
10	The time allotted for the training was sufficient.			
11	The meeting room and facilities were adequate and comfortable.			

Advantages of Training

- Training is an investment
- Trainings help you keep your skills up to date, and prepare you for greater responsibilities.
- It can boost your confidence, strengthen your professional credibility and help you become more creative in tackling new challenges.
- Trainings makes your working life more interesting and can significantly increase your job satisfaction.
- It can accelerate your career development and is an important part of upgrading to chartered membership.

Learning and Development policies and record keeping

- Learning and Development policy for each organization differs.
- It includes
  - significant investments in developing in-house capabilities in many training areas, both technical and non-technical
  - Partnership with several leading training providers, in order to ensure best-in-class training for their employees.
  - Training needs identification is done for new recruits. Mandatory New Hire Orientation program for newly joined
  - Training needs for new process/new role.
  - Training covers business/process understanding, technical capabilities, communication, interpersonal skills, leadership skills etc.

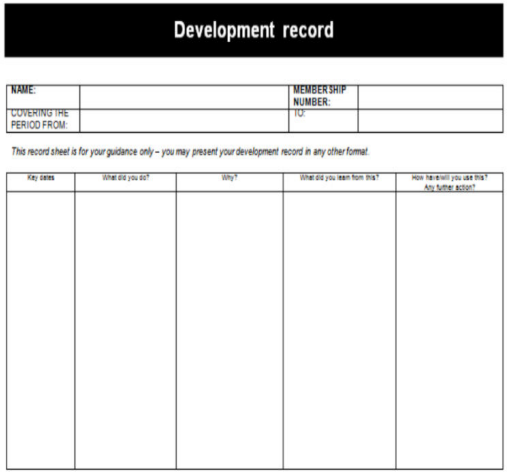
Learning and Development policies and record keeping

- Need for Record Keeping
  - Keep your skills and knowledge up to date
  - Review your learning over the previous 12 months, and set your development objectives for the coming year.
  - Reflecting on the past and planning for the future in this way makes your development more methodical and easier to measure. This is a particularly useful exercise prior to your annual appraisal!
  - Without regularly recording your experience, it difficult to review your learning and learning needs

Maintaining a personal portfolio

- It helps to provide documented evidence of your commitment to your chosen profession; and of your continued competence
- It will act as an excellent reference, both in the updating of your Curriculum Vitae and in recalling details of topics you have studied
- It will be a most useful aid in your career development, providing a means by which you can plan, record and review your relevant activities

Sample Development Record



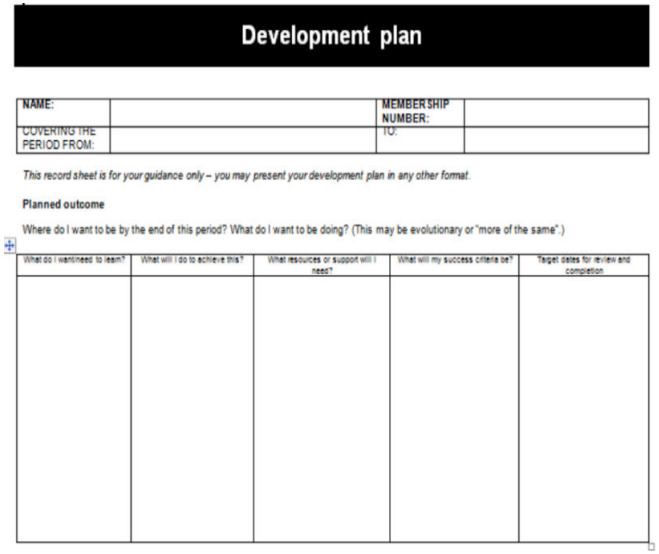
Development record

NAME:		MEMBERSHIP NUMBER:	
COVERING THE PERIOD FROM:		TO:	

This record sheet is for your guidance only – you may present your development record in any other format.

Key dates	What did you do?	Why?	What did you learn from this?	How have/will you use this? Any further action?

Sample Development Plan



Development plan

NAME:		MEMBERSHIP NUMBER:	
COVERING THE PERIOD FROM:		TO:	

This record sheet is for your guidance only – you may present your development plan in any other format.

Planned outcome

Where do I want to be by the end of this period? What do I want to be doing? (This may be evolutionary or "more of the same".)

What do I want/need to learn?	What will I do to achieve this?	What resources or support will I need?	What will my success criteria be?	Target dates for review and completion

Continuous Professional Development (CPD)

- The term is generally used to mean a physical folder or portfolio documenting your development as a professional.
- CPD refers to the process of tracking and documenting the skills, knowledge and experience that you gain both formally and informally as you work, beyond any initial training.
- It's a record of what you experience, learn and then apply.
- It helps you to reflect, review and document your learning and to develop and update your professional knowledge and skills.

Why CPD?

- provides an overview of your professional development to date
- reminds you of your achievements and how far you've progressed
- directs your career and helps you keep your eye on your goals
- uncovers gaps in your skills and capabilities and opens up further development needs
- demonstrates your professional standing to clients and employers and for career development or career change

Ask yourself

- Where am I now?
- Where do I want to be?
- What do I have to do to get there? -

Thank you