Unit – II

What Is Exploratory Data Analysis?

Before you start data analysis or run your data through a machine learning algorithm, you must clean your data and make sure it is in a suitable form. Further, it is essential to know any recurring patterns and significant correlations that might be present in your data. The process of getting to know your data in depth is called Exploratory Data Analysis.

Exploratory Data Analysis is an integral part of working with data.

Machine learning Algorithms

where nearly all manual tasks are being automated, the definition of manual is changing. There are now many different types of Machine Learning algorithms, some of which can help computers play chess, perform surgeries, and get smarter and more personal.

We are living in an era of constant technological progress, and looking at how computing has advanced over the years, we can predict what's to come in the days ahead.

One of the main features of this revolution that stands out is how computing tools and techniques have been democratized. Data scientists have built sophisticated data-crunching machines in the last 5 years by seamlessly executing advanced techniques. The results have been astounding.

The many different types of machine learning algorithms have been designed in such dynamic times to help solve real-world complex problems. The ml algorithms are automated and self-modifying to continue improving over time. Before we delve into the top 10 machine learning algorithms you should know, let's take a look at the different types of machine learning algorithms and how they are classified.

Machine learning algorithms are classified into 4 types:

- Supervised
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning

Below is the list of Top 10 commonly used Machine Learning (ML) Algorithms:

- Linear regression
- Logistic regression
- Decision tree
- SVM algorithm

- Naive Bayes algorithm
- KNN algorithm
- K-means
- Random forest algorithm
- Dimensionality reduction algorithms
- Gradient boosting algorithm and AdaBoosting algorithm

Data Analysis: Regression Modelling

A regression model determines a relationship between an independent variable and a dependent variable, by providing a function. Formulating a regression analysis helps you predict the effects of the independent variable on the dependent one.

Example: we can say that age and height can be described using a linear regression model. Since a person's height increases as age increases, they have a linear relationship.

There are five different types of regression models:

1. **Linear**
2. **Non-linear**
3. **Multiple**
4. **Polynomial**
5. **Logistic**

1. Linear Regression:

To understand the working functionality of this ml algorithm, imagine how you would arrange random logs of wood in increasing order of their weight. There is a catch; however – you cannot weigh each log. You have to guess its weight just by looking at the height and girth of the log (visual analysis) and arranging them using a combination of these visible parameters. This is what linear regression in machine learning is like.
In this process, a relationship is established between independent and dependent variables by fitting them to a line. This line is known as the regression line and is represented by a linear equation Y= a *X + b.
In this equation:

- Y – Dependent Variable
- a – Slope
- X – Independent variable
- b – Intercept

The coefficients a & b are derived by minimizing the sum of the squared difference of distance between data points and the regression line.

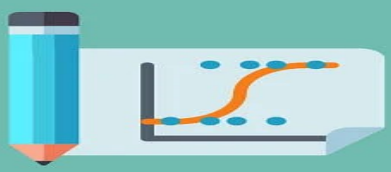
TOP 10 MACHINE LEARNING ALGORITHMS

1

LINEAR REGRESSION

In this process, a relationship is established between independent and dependent variables by fitting them to a line. This line is known as the regression line and represented by a linear equation Y= a *X + b.

2



LOGISTIC REGRESSION

Logistic Regression is used to estimate discrete values (usually binary values like 0/1) from a set of independent variables. It helps predict the probability of an event by fitting data to a logit function.

3

DECISION TREE

This is a supervised learning algorithm that is used for classifying problems. In this algorithm, we split the population into two or more homogeneous sets based on the most significant attributes/ independent variables.

4

SVM ALGORITHM

In SVM (Support Vector Machine) algorithm, we plot raw data as points in an n-dimensional space (n = no. of features you have). The value of each feature is then tied to a particular coordinate, making it easy to classify the data.

5

NAIVE BAYES ALGORITHM

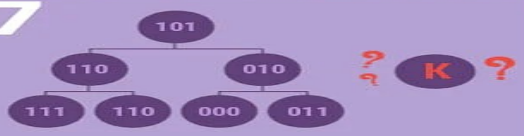
A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

6

KNN ALGORITHM

This algorithm can be applied to both classification and regression problems. It stores all available cases and classifies any new cases by taking a majority vote of its k neighbors. The case is then assigned to the class with which it has the most in common.

7



K-MEANS

In this unsupervised learning algorithm, data sets are classified into a particular number of clusters in such a way that all the data points within a cluster are homogenous and heterogeneous from the data in other clusters.

8

RANDOM FOREST ALGORITHM

A collective of decision trees is called a Random Forest. To classify a new object based on its attributes, each tree is classified, and the tree "votes" for that class. The forest chooses the classification having the most votes.

9

DIMENSIONALITY REDUCTION ALGORITHMS

Dimensionality reduction algorithms like Decision Tree, Factor Analysis, Missing Value Ratio, and Random Forest can help you find relevant details.

10

GRADIENT BOOSTING ALGORITHM AND ADABOOSTING ALGORITHM

These are boosting algorithms used when massive loads of data have to be handled to make predictions with high accuracy.

simplilearn
www.simplilearn.com

TO READ THE FULL ARTICLE GO TO:
https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article

Regression Modelling:

Introduction

Linear and Logistic regressions are usually the first algorithms people learn in data science. Due to their popularity, a lot of analysts even end up thinking that they are the only form of regressions. The ones who are slightly more involved think that they are the most important among all forms of regression analysis.

The truth is that there are innumerable forms of regressions, which can be performed. Each form has its own importance and a specific condition where they are best suited to apply. In this article, I have explained the most commonly used 7 types of regression in data science in a simple manner.

Through this article, I also hope that people develop an idea of the breadth of regressions, instead of just applying linear/logistic regression to every machine learning problem they come across and hoping that they would just fit!

Table of Contents

What is Regression Analysis?

Regression analysis is a form of predictive modelling technique which investigates the relationship between a **dependent** (target) and **independent variable**

(s) (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

Regression analysis is an important tool for modelling and analyzing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized. I'll explain this in more details in coming sections.

Why do we use Regression Analysis?

As mentioned above, regression analysis estimates the relationship between two or more variables. Let's understand this with an easy example:

Let's say, you want to estimate growth in sales of a company based on current economic conditions. You have the recent company data which indicates that the growth in sales is around two and a half times the growth in the economy. Using this insight, we can predict future sales of the company based on current & past information.

There are multiple benefits of using regression analysis. They are as follows:

1. It indicates the **significant relationships** between dependent variable and independent variable.

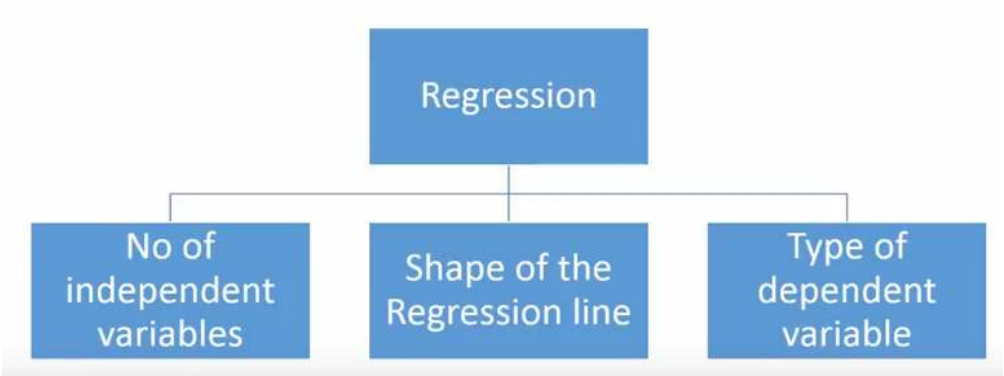
2. It indicates the **strength of impact** of multiple independent variables on a dependent variable.

Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to

eliminate and evaluate the best set of variables to be used for building predictive models.

How many types of regression techniques do we have?

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line). We'll discuss them in detail in the following sections.



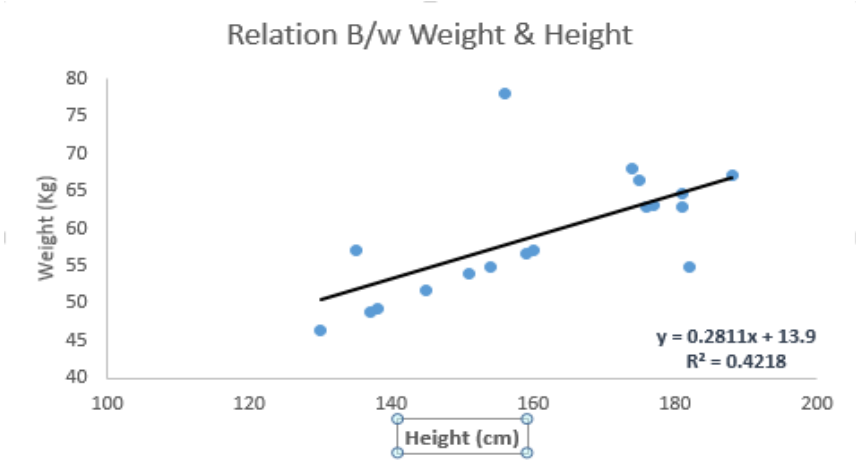
For the creative ones, you can even cook up new regressions, if you feel the need to use a combination of the parameters above, which people haven't used before. But before you start that, let us understand the most commonly used regressions:

1. Linear Regression

It is one of the most widely known modeling technique. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.

Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as regression line).

It is represented by an equation **Y=a+b*X + e**, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

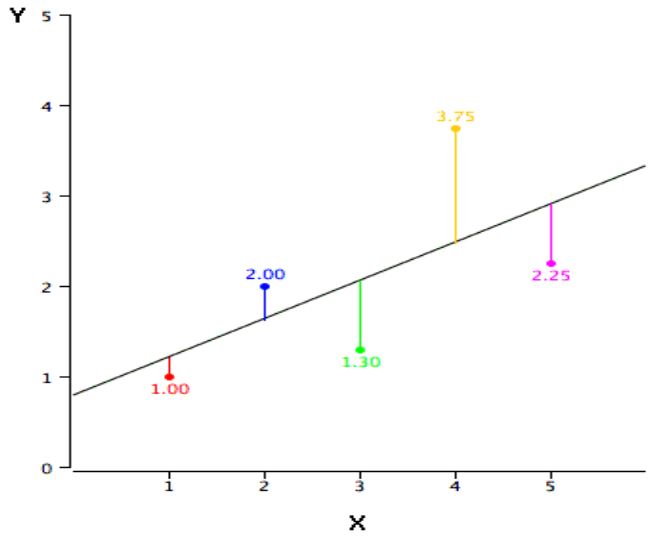


The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable. Now, the question is "How do we obtain best fit line?".

How to obtain best fit line (Value of a and b)?

This task can be easily accomplished by Least Square Method. It is the most common method used for fitting a regression line. It calculates the best-fit line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. Because the deviations are first squared, when added, there is no cancelling out between positive and negative values.

$$\min_{w} ||Xw - y||_2^2$$



We can evaluate the model performance using the metric **R-square**. To know more details about these metrics, you can read: Model Performance metrics Part 1, Part 2 .

Important Points:

- There must be **linear relationship** between independent and dependent variables
- Multiple regression suffers from **multicollinearity, autocorrelation, heteroskedasticity**.
- Linear Regression is very sensitive to **Outliers**. It can terribly affect the regression line and eventually the forecasted values.
- Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable
- In case of multiple independent variables, we can go with **forward selection**, **backward elimination** and **step wise approach** for selection of most significant independent variables.

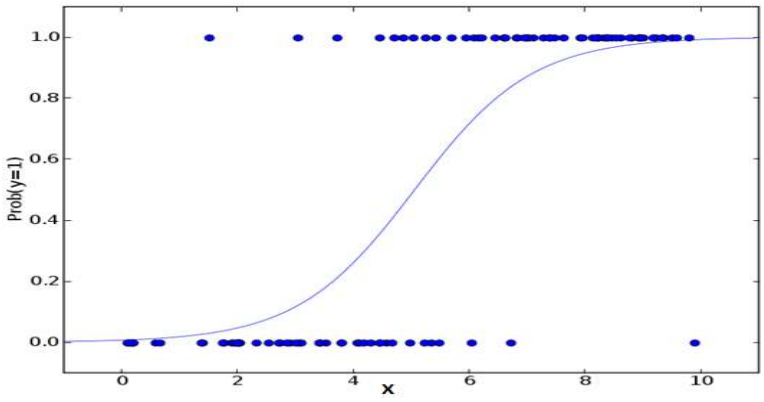
2. Logistic Regression

Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can represented by following equation.

```
odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence
ln(odds) = ln(p/(1-p))
logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk
```

Above, p is the probability of presence of the characteristic of interest. A question that you should ask here is “why have we used log in the equation?”.

Since we are working here with a binomial distribution (dependent variable), we need to choose a link function which is best suited for this distribution. And, it is **logit** function. In the equation above, the parameters are chosen to maximize the likelihood of observing the sample values rather than minimizing the sum of squared errors (like in ordinary regression).



Important Points:

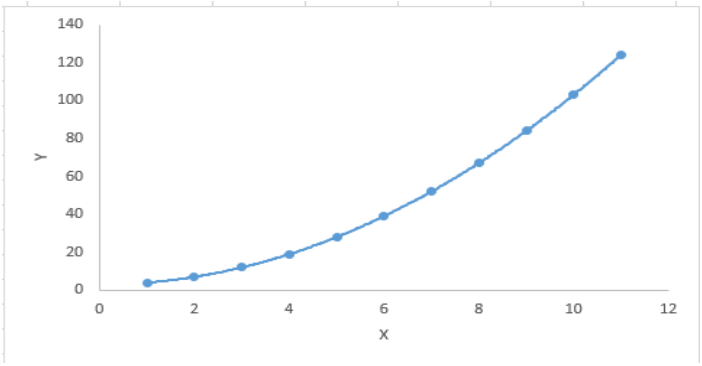
- Logistic regression is widely used for **classification problems**
- Logistic regression doesn't require linear relationship between dependent and independent variables. It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio
- To avoid over fitting and under fitting, we should include all significant variables. A good approach to ensure this practice is to use a step wise method to estimate the logistic regression
- It requires **large sample sizes** because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square
- The independent variables should not be correlated with each other i.e. **no multi collinearity**. However, we have the options to include interaction effects of categorical variables in the analysis and in the model.
- If the values of dependent variable is ordinal, then it is called as **Ordinal logistic regression**
- If dependent variable is multi class then it is known as **Multinomial Logistic regression**.

3. Polynomial Regression

A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation:

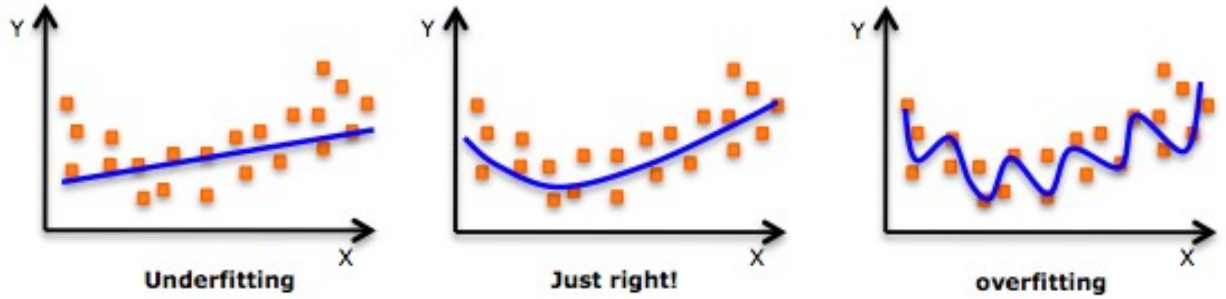
```
y=a+b*x^2
```

In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.



Important Points:

- While there might be a temptation to fit a higher degree polynomial to get lower error, this can result in over-fitting. Always plot the relationships to see the fit and focus on making sure that the curve fits the nature of the problem. Here is an example of how plotting can help:



- Especially look out for curve towards the ends and see whether those shapes and trends make sense. Higher polynomials can end up producing wierd results on extrapolation.

4. Stepwise Regression

This form of regression is used when we deal with multiple independent variables. In this technique, the selection of independent variables is done with the help of an automatic process, which involves *no* human intervention.

This feat is achieved by observing statistical values like R-square, t-stats and AIC metric to discern significant variables. Stepwise regression basically fits the regression model by adding/dropping co-variates one at a time based on a specified criterion. Some of the most commonly used Stepwise regression methods are listed below:

- Standard stepwise regression does two things. It adds and removes predictors as needed for each step.

- Forward selection starts with most significant predictor in the model and adds variable for each step.

- Backward elimination starts with all predictors in the model and removes the least significant variable for each step.

The aim of this modeling technique is to maximize the prediction power with minimum number of predictor variables. It is one of the method to handle higher dimensionality of data set.

5. Ridge Regression

Ridge Regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated). In multicollinearity, even though the least squares estimates (OLS) are unbiased, their variances are large which deviates the observed value far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

Above, we saw the equation for linear regression. Remember? It can be represented as:

y=a+ b*x

This equation also has an error term. The complete equation becomes:

y=a+b*x+e (error term), [error term is the value needed to correct for a prediction error between the observed and predicted value]

=> y=a+y= a+ b1x1+ b2x2+....+e, for multiple independent variables.

In a linear equation, prediction errors can be decomposed into two sub components. First is due to the **biased** and second is due to the **variance**. Prediction error can occur due to any one of these two or both components. Here, we'll discuss about the error caused due to variance.

Ridge regression solves the multicollinearity problem through shrinkage parameter λ (lambda). Look at the equation below.

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

In this equation, we have two components. First one is least square term and other one is lambda of the summation of β2 (beta- square) where β is the coefficient. This is added to least square term in order to shrink the parameter to have a very low variance.

Important Points:

- The assumptions of this regression is same as least squared regression except normality is not to be assumed
- Ridge regression shrinks the value of coefficients but doesn't reaches zero, which suggests no feature selection feature
- This is a regularization method and uses l2 regularization.

6. Lasso Regression

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

Similar to Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) also penalizes the absolute size of the regression coefficients. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models. Look at the equation below: Lasso regression differs from ridge regression in a way that it uses absolute values in the penalty function, instead of squares. This leads to penalizing (or equivalently constraining the sum of the absolute values of the estimates) values which causes some of the parameter estimates to turn out exactly zero. Larger the penalty applied, further the estimates get shrunk towards absolute zero. This results to variable selection out of given n variables.

Important Points:

- The assumptions of lasso regression is same as least squared regression except normality is not to be assumed
- Lasso Regression shrinks coefficients to zero (exactly zero), which certainly helps in feature selection
- Lasso is a regularization method and uses l1 regularization
- If group of predictors are highly correlated, lasso picks only one of them and shrinks the others to zero

7. ElasticNet Regression

ElasticNet is hybrid of Lasso and Ridge Regression techniques. It is trained with L1 and L2 prior as regularizer. Elastic-net is useful when there are multiple features which are correlated. Lasso is likely to pick one of these at random, while elastic-net is likely to pick both.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

A practical advantage of trading-off between Lasso and Ridge is that, it allows Elastic-Net to inherit some of Ridge's stability under rotation.

Important Points:

- It encourages group effect in case of highly correlated variables
- There are no limitations on the number of selected variables
- It can suffer with double shrinkage

Beyond these 7 most commonly used regression techniques, you can also look at other models like Bayesian, Ecological and Robust regression.

How to select the right regression model?

Life is usually simple, when you know only one or two techniques. One of the training institutes I know of tells their students – if the outcome is continuous – apply linear regression. If it is binary – use logistic regression! However, higher the number of options available at our disposal, more difficult it becomes to choose the right one. A similar case happens with regression models.

Within multiple types of regression models, it is important to choose the best suited technique based on type of independent and dependent variables, dimensionality in the data and other essential characteristics of the data. Below are the key factors that you should practice to select the right regression model:

1. Data exploration is an inevitable part of building predictive model. It should be you first step before selecting the right model like identify the relationship and impact of variables

2. To compare the goodness of fit for different models, we can analyse different metrics like statistical significance of parameters, R-square, Adjusted r-square, AIC, BIC and error term. Another one is the Mallow's Cp criterion. This essentially checks for possible bias in your model, by comparing the model with all possible submodels (or a careful selection of them).

3. Cross-validation is the best way to evaluate models used for prediction. Here you divide your data set into two group (train and validate). A simple mean squared difference between the observed and predicted values give you a measure for the prediction accuracy.

4. If your data set has multiple confounding variables, you should not choose automatic model selection method because you do not want to put these in a model at the same time.

5. It'll also depend on your objective. It can occur that a less powerful model is easy to implement as compared to a highly statistically significant model.

6. Regression regularization methods(Lasso, Ridge and ElasticNet) works well in case of high dimensionality and multicollinearity among the variables in the data set.

End Note

By now, I hope you would have got an overview of regression. These regression techniques should be applied considering the conditions of data. One of the best trick to find out which technique to use, is by checking the family of variables i.e. discrete or continuous.

In this article, I discussed about 7 types of regression and some key facts associated with each technique. As somebody who's new in this industry, I'd advise you to learn these techniques and later implement them in your models.

Data analytics is all about looking at various factors to see how they impact certain situations and outcomes. When dealing with data that contains more than two variables, you'll use multivariate analysis. Multivariate analysis isn't just one specific method—rather, it encompasses a whole range of statistical techniques. These techniques allow you to gain a deeper understanding of your data in relation to specific business or real-world scenarios. So, if you're an aspiring data analyst or data scientist, multivariate analysis is an important concept to get to grips with.

In this post, we'll provide a complete introduction to multivariate analysis. We'll delve deeper into defining what multivariate analysis actually is, and we'll introduce some key techniques you can use when analyzing your data. We'll also give some examples of multivariate analysis in action.

What is Multivariate Analysis?

Multivariate Analysis is defined as a process of involving multiple dependent variables resulting in one outcome. This explains that the majority of the problems in the real world are Multivariate. For example, we cannot predict the weather of any year based on the season. There are multiple factors like pollution, humidity, precipitation, etc. Here, we will introduce you to multivariate analysis, its history, and its application in different fields. Also, take up a Multivariate Time Series Forecasting In R to learn more about the concept.

What is multivariate analysis?

In data analytics, we look at different variables (or factors) and how they might impact certain situations or outcomes. For example, in marketing, you might look at how the variable "money spent on advertising" impacts the variable "number of sales." In the healthcare sector, you might want to explore whether there's a correlation between "weekly hours of exercise" and "cholesterol level." This helps us to understand why certain outcomes occur, which in turn allows us to make informed predictions and decisions for the future.
There are three categories of analysis to be aware of:

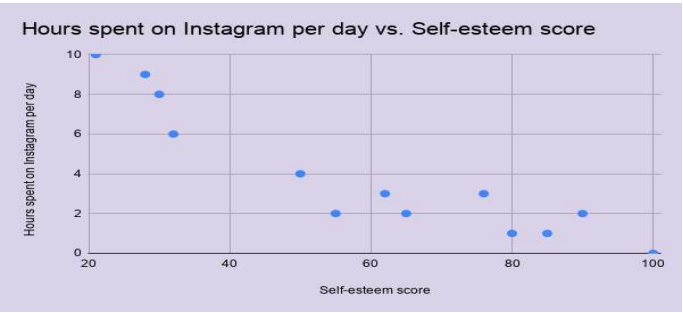
- **Univariate analysis**, which looks at just one variable
- **Bivariate analysis**, which analyzes two variables
- **Multivariate analysis**, which looks at more than two variables

As you can see, multivariate analysis encompasses all statistical techniques that are used to analyze more than two variables at once. The aim is to find patterns and **correlations** between several variables simultaneously—allowing for a much deeper, more complex understanding of a given scenario than you'll get with bivariate analysis.

An example of multivariate analysis

Let's imagine you're interested in the relationship between a person's social media habits and their self-esteem. You could carry out a bivariate analysis, comparing the following two variables:

1. How many hours a day a person spends on Instagram
2. Their self-esteem score (measured using a self-esteem scale)



You may or may not find a relationship between the two variables; however, you know that, in reality, self-esteem is a complex concept. It's likely impacted by many different factors—not just how many hours a person spends on Instagram. You might also want to consider factors such as age, employment status, how often a person exercises, and relationship status (for example). In order to deduce the extent to which each of these variables correlates with self-esteem, and with each other, you'd need to run a multivariate analysis.
So we know that multivariate analysis is used when you want to explore more than two variables at once. Now let's consider some of the different techniques you might use to do this.

3. Multivariate data analysis techniques and examples

There are many different techniques for multivariate analysis, and they can be divided into two categories:

- Dependence techniques
- Interdependence techniques

So what's the difference? Let's take a look.

Multivariate analysis techniques: Dependence vs. interdependence

When we use the terms "dependence" and "interdependence," we're referring to different types of relationships within the data. To give a brief explanation:

Dependence methods

Dependence methods are used when one or some of the variables are dependent on others. Dependence looks at cause and effect; in other words, can the values of two or more independent variables be used to explain, describe, or predict the value of another, dependent variable? To give a simple example, the dependent variable of "weight" might be predicted by independent variables such as "height" and "age."
In machine learning, dependence techniques are used to build predictive models. The analyst enters input data into the model, specifying which variables are independent and which ones are dependent—in other words, which variables they want the model to predict, and which variables they want the model to use to make those predictions.

Interdependence methods

Interdependence methods are used to understand the structural makeup and underlying patterns within a dataset. In this case, no variables are dependent on others, so you're not looking for causal relationships. Rather, interdependence methods seek to give meaning to a set of variables or to group them together in meaningful ways.
So: One is about the effect of certain variables on others, while the other is all about the structure of the dataset.
With that in mind, let's consider some useful multivariate analysis techniques. We'll look at:

- Multiple linear regression
- Multiple logistic regression
- Multivariate analysis of variance (MANOVA)
- Factor analysis

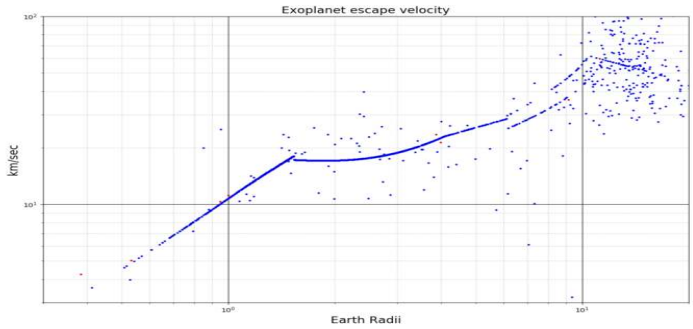
- Cluster analysis

Multiple linear regression:

Multiple linear regression is a dependence method which looks at the relationship between one dependent variable and two or more independent variables. A multiple regression model will tell you the extent to which each independent variable has a linear relationship with the dependent variable. This is useful as it helps you to understand which factors are likely to influence a certain outcome, allowing you to estimate future outcomes.

Example of multiple regression:

As a data analyst, you could use multiple regression to predict crop growth. In this example, crop growth is your dependent variable and you want to see how different factors affect it. Your independent variables could be rainfall, temperature, amount of sunlight, and amount of fertilizer added to the soil. A multiple regression model would show you the proportion of variance in crop growth that each independent variable accounts for.



Multiple logistic regression :

Logistic regression analysis is used to calculate (and predict) the probability of a binary event occurring. A binary outcome is one where there are only two possible outcomes; either the event occurs (1) or it doesn't (0). So, based on a set of independent variables, logistic regression can predict how likely it is that a certain scenario will arise. It is also used for classification. You can learn about **the difference between regression and classification here**.

Example of logistic regression:

Let's imagine you work as an analyst within the insurance sector and you need to predict how likely it is that each potential customer will make a claim. You might enter a range of independent variables into your model, such as age, whether or not they have a serious health condition, their occupation, and so on. Using these variables, a logistic regression analysis will calculate the probability of the event (making a claim) occurring. Another oft-cited example is the filters used to classify email as "spam" or "not spam." You'll find a more detailed explanation in this **complete guide to logistic regression**.

Multivariate analysis of variance (MANOVA):

Multivariate analysis of variance (MANOVA) is used to measure the effect of multiple independent variables on two or more dependent variables. With MANOVA, it's important to note that the independent variables are categorical, while the dependent variables are metric in nature. A categorical variable is a variable that belongs to a distinct category—for example, the variable "employment status" could be categorized into certain units, such as "employed full-time," "employed part-time," "unemployed," and so on. A metric variable is measured quantitatively and takes on a numerical value.
In MANOVA analysis, you're looking at various combinations of the independent variables to compare how they differ in their effects on the dependent variable.

Example of MANOVA:

Let's imagine you work for an engineering company that is on a mission to build a super-fast, eco-friendly rocket. You could use MANOVA to measure the effect that various design combinations have on both the speed of the rocket and the amount of carbon dioxide it emits. In this scenario, your categorical independent variables could be:

- Engine type, categorized as E1, E2, or E3
- Material used for the rocket exterior, categorized as M1, M2, or M3
- Type of fuel used to power the rocket, categorized as F1, F2, or F3

Your metric dependent variables are speed in kilometers per hour, and carbon dioxide measured in parts per million. Using MANOVA, you'd test different combinations (e.g. E1, M1, and F1 vs. E1, M2, and F1, vs. E1, M3, and F1, and so on) to calculate the effect of all the independent variables. This should help you to find the optimal design solution for your rocket.

Factor analysis

Factor analysis is an interdependence technique which seeks to reduce the number of variables in a dataset. If you have too many variables, it can be difficult to find patterns in your data. At the same time, models created using datasets with too many variables are susceptible to overfitting. Overfitting is a modeling error that occurs when a model fits too closely and specifically to a certain dataset, making it less generalizable to future datasets, and thus potentially less accurate in the predictions it makes.
Factor analysis works by detecting sets of variables which correlate highly with each other. These variables may then be condensed into a single variable. Data analysts will often carry out factor analysis to prepare the data for subsequent analyses.

Factor analysis example:

Let's imagine you have a dataset containing data pertaining to a person's income, education level, and occupation. You might find a high degree of correlation among each of these variables, and thus reduce them to the single factor "socioeconomic status." You might also have data on how happy they were with customer service, how much they like a certain product, and how likely they are to recommend the product to a friend. Each of these variables could be grouped into the single factor "customer satisfaction" (as long as they are found to correlate strongly with one another). Even though you've reduced several data points to just one factor,

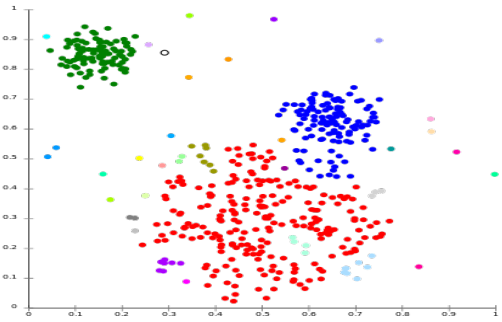
you're not really losing any information—these factors adequately capture and represent the individual variables concerned. With your "streamlined" dataset, you're now ready to carry out further analyses.

Cluster analysis:

Another interdependence technique, cluster analysis is used to group similar items within a dataset into clusters. When grouping data into clusters, the aim is for the variables in one cluster to be more similar to each other than they are to variables in other clusters. This is measured in terms of intracluster and intercluster distance. Intracluster distance looks at the distance between data points within one cluster. This should be small. Intercluster distance looks at the distance between data points in different clusters. This should ideally be large. Cluster analysis helps you to understand how data in your sample is distributed, and to find patterns.

Cluster analysis example:

A prime example of cluster analysis is audience segmentation. If you were working in marketing, you might use cluster analysis to define different customer groups which could benefit from more targeted campaigns. As a **healthcare analyst**, you might use cluster analysis to explore whether certain lifestyle factors or geographical locations are associated with higher or lower cases of certain illnesses. Because it's an interdependence technique, cluster analysis is often carried out in the early stages of data analysis.



More multivariate analysis techniques:

This is just a handful of multivariate analysis techniques used by data analysts and data scientists to understand complex datasets. If you're keen to explore further, check out discriminant analysis, conjoint analysis, canonical correlation analysis, structural equation modeling, and multidimensional scaling.

Related content: An intro to data visualization, as taught by Dr. Humera Noor Minhas, a data analyst with more than 20 year's experience working in the field.

4. What are the advantages of multivariate analysis?

The one major advantage of multivariate analysis is the depth of insight it provides. In exploring multiple variables, you're painting a much more detailed picture of what's occurring—and, as a result, the insights you uncover are much more applicable to the real world.

Remember our self-esteem example back in section one? We could carry out a bivariate analysis, looking at the relationship between self-esteem and just one other factor; and, if we

found a strong correlation between the two variables, we might be inclined to conclude that this particular variable is a strong determinant of self-esteem. However, in reality, we know that self-esteem can't be attributed to one single factor. It's a complex concept; in order to create a model that we could really trust to be accurate, we'd need to take many more factors into account. That's where multivariate analysis really shines; it allows us to analyze many different factors and get closer to the reality of a given situation.

5. **Key takeaways and further reading**

In this post, we've learned that multivariate analysis is used to analyze data containing more than two variables. To recap, here are some key takeaways:

- The aim of multivariate analysis is to find patterns and correlations between several variables simultaneously
- Multivariate analysis is especially useful for analyzing complex datasets, allowing you to gain a deeper understanding of your data and how it relates to real-world scenarios
- There are two types of multivariate analysis techniques: Dependence techniques, which look at cause-and-effect relationships between variables, and interdependence techniques, which explore the structure of a dataset
- Key multivariate analysis techniques include multiple linear regression, multiple logistic regression, MANOVA, factor analysis, and cluster analysis—to name just a few.

Bayesian modeling, inference and Bayesian networks :

Bayesian Statistics continues to remain incomprehensible in the ignited minds of many analysts. Being amazed by the incredible power of machine learning, a lot of us have become unfaithful to statistics. Our focus has narrowed down to exploring machine learning. Isn't it true?
We fail to understand that machine learning is not the only way to solve real world problems. In several situations, it does not help us solve business problems, even though there is data involved in these problems. To say the least, knowledge of statistics will allow you to work on complex analytical problems, irrespective of the size of data.

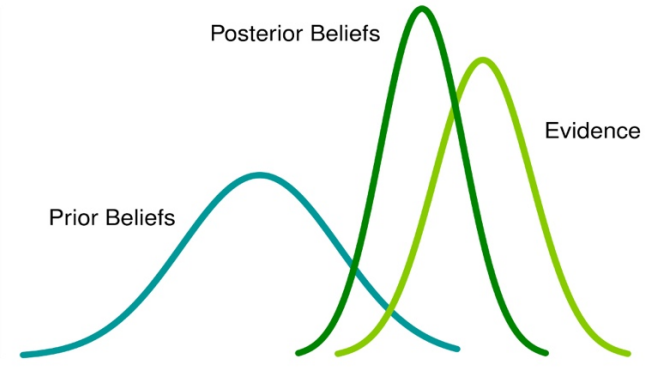


Table of Contents

- Bernoulli likelihood function
- Prior Belief Distribution
- Posterior belief Distribution

5. Test for Significance – Frequentist vs Bayesian
   - p-value
   - Confidence Intervals
   - Bayes Factor
   - High Density Interval (HDI)

Before we actually delve in Bayesian Statistics, let us spend a few minutes understanding *Frequentist Statistics*, the more popular version of statistics most of us come across and the inherent problems in that.

1. Frequentist Statistics

The debate between *frequentist* and *bayesian* have haunted beginners for centuries. Therefore, it is important to understand the difference between the two and how does there exists a thin line of demarcation!

It is the most widely used inferential technique in the statistical world. Infact, generally it is the first school of thought that a person entering into the statistics world comes across.

Frequentist Statistics tests whether an event (hypothesis) occurs or not. It calculates the probability of an event in the long run of the experiment (i.e the experiment is repeated under the same conditions to obtain the outcome).

Here, the sampling distributions of **fixed size** are taken. Then, the experiment is theoretically repeated **infinite number of times** but practically done with a stopping intention. For example, I perform an experiment with a stopping intention in mind that I will stop the experiment when it is repeated 1000 times or I see minimum 300 heads in a coin toss.

Let's go deeper now.

Now, we'll understand *frequentist statistics* using an example of coin toss. The objective is to estimate the fairness of the coin. Below is a table representing the frequency of heads:

no. of tosses	no. of heads	difference
10	4	-1
50	25	0
100	44	-6
500	255	5
1000	502	2
5000	2533	33
10000	5067	67

We know that probability of getting a head on tossing a fair coin is 0.5. No. of heads represents the actual number of heads obtained. Difference is the difference between 0.5*(No. of tosses) - no. of heads.

An important thing is to note that, though the difference between the actual number of heads and expected number of heads(50% of number of tosses) increases as the number of tosses are increased, the proportion of number of heads to total number of tosses approaches 0.5 (for a fair coin).

This experiment presents us with a very common flaw found in frequentist approach i.e. *Dependence of the result of an experiment on the number of times the experiment is repeated.*

To know more about frequentist statistical methods, you can head to this excellent course on inferential statistics.

3. The Inherent Flaws in Frequentist Statistics

Till here, we've seen just one flaw in *frequentist statistics*. Well, it's just the beginning.
20th century saw a massive upsurge in the *frequentist statistics* being applied to numerical models to check whether one sample is different from the other, a parameter is important enough to be kept in the model and variousother manifestations of hypothesis testing. But *frequentist statistics* suffered some great flaws in its design and interpretation which posed a serious concern in all real life problems. For example:
1. p-values measured against a sample (fixed size) statistic with some stopping intention changes with change in intention and sample size. i.e If two persons work on the same data and have different stopping intention, they may get two different p- values for the same data, which is undesirable.
For example: Person A may choose to stop tossing a coin when the total count reaches 100 while B stops at 1000. For different sample sizes, we get different t-scores and different p-values. Similarly, intention to stop may change from fixed number of flips to total duration of flipping. In this case too, we are bound to get different *p-values*.
2- Confidence Interval (C.I) like p-value depends heavily on the sample size. This makes the stopping potential absolutely absurd since no matter how many persons perform the tests on the same data, the results should be consistent.
3- Confidence Intervals (C.I) are not probability distributions therefore they do not provide the most probable value for a parameter and the most probable values.
These three reasons are enough to get you going into thinking about the drawbacks of the *frequentist approach* and why is there a need for *bayesian approach*. Let's find it out.
From here, we'll first understand the basics of Bayesian Statistics.

3. Bayesian Statistics

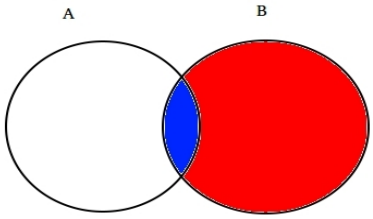
"Bayesian statistics is a mathematical procedure that applies probabilities to statistical problems. It provides people the tools to update their beliefs in the evidence of new data."
You got that? Let me explain it with an example:
Suppose, out of all the 4 championship races (F1) between Niki Lauda and James hunt, Niki won 3 times while James managed only 1.
So, if you were to bet on the winner of next race, who would he be ?
I bet you would say Niki Lauda.
Here's the twist. What if you are told that it rained once when James won and once when Niki won and it is definite that it will rain on the next date. So, who would you bet your money on now ?
By intuition, it is easy to see that chances of winning for James have increased drastically. But the question is: how much ?
To understand the problem at hand, we need to become familiar with some concepts, first of which is conditional probability (explained below).
In addition, there are certain pre-requisites:
Pre-Requisites:

1. Linear Algebra : To refresh your basics, you can check out Khan's Academy Algebra.

2. Probability and Basic Statistics : To refresh your basics, you can check out another course by Khan Academy.

3.1 Conditional Probability

It is defined as the: Probability of an event A given B equals the probability of B and A happening together divided by the probability of B."
For example: Assume two partially intersecting sets A and B as shown below.
Set A represents one set of events and Set B represents another. We wish to calculate the probability of A given B has already happened. Lets represent the happening of event B by shading it with red.



Now since B has happened, the part which now matters for A is the part shaded in blue which is interestingly $A \cap B$. So, the probability of A given B turns out to be:

$$\frac{BlueArea}{RedArea + BlueArea}$$

Therefore, we can write the formula for event B given A has already occurred by:

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

or

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

Now, the second equation can be rewritten as :

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

This is known as **Conditional Probability**.
Let's try to answer a betting problem with this technique.
Suppose, B be the *event of winning of James Hunt*. A be the *event of raining*. Therefore,

1. P(A) =1/2, since it rained twice out of four days.
2. P(B) is 1/4, since James won only one race out of four.
3. P(A|B)=1, since it rained every time when James won.

Substituting the values in the conditional probability formula, we get the probability to be around 50%, which is almost the double of 25% when rain was not taken into account (Solve it at your end).

This further strengthened our belief of James winning in the light of new *evidence* i.e rain. You must be wondering that this formula bears close resemblance to something you might have heard a lot about. Think!
Probably, you guessed it right. It looks like **Bayes Theorem**.
Bayes theorem is built on top of conditional probability and lies in the heart of Bayesian Inference. Let's understand it in detail now.

3.2 Bayes Theorem

Bayes Theorem comes into effect when multiple events A_i form an exhaustive set with another event B. This could be understood with the help of the below diagram.

A1	B	
A2		
A3		

Now, B can be written as

$$B = \sum_{i=1}^{n} B \cap A_i$$

So, probability of B can be written as,

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i)$$

$$P(B \cap A_i) = P(B|A_i) \times P(A_i)$$

But
So, replacing P(B) in the equation of conditional probability we get

$$P(A_i|B) = (P(B|A_i) \times P(A_i))/(\sum_{i=1}^{n}(P(B|A_i) \times P(A_i)))$$

This is the equation of **Bayes Theorem**.

4. Bayesian Inference

There is no point in diving into the theoretical aspect of it. So, we'll learn how it works! Let's take an example of coin tossing to understand the idea behind *bayesian inference*.
An important part of *bayesian inference* is the establishment of *parameters* and *models.*
Models are the mathematical formulation of the observed events. Parameters are the factors in the models affecting the observed data. For example, in tossing a coin, **fairness of coin** may be defined as the parameter of coin denoted by θ. The outcome of the events may be denoted by D.
Answer this now. What is the probability of 4 heads out of 9 tosses(D) given the fairness of coin (θ). i.e P(D|θ)
Wait, did I ask the right question? No.

We should be more interested in knowing : Given an outcome (D) what is the probaility of coin being fair (θ=0.5)
Lets represent it using Bayes Theorem:

$$P(\theta|D)=(P(D|\theta) \text{ X } P(\theta))/P(D)$$

Here, P(θ) is the ***prior*** i.e the strength of our belief in the fairness of coin before the toss. It is perfectly okay to believe that coin can have any degree of fairness between 0 and 1.
P(D|θ) is the likelihood of observing our result given our distribution for θ. If we knew that coin was fair, this gives the probability of observing the number of heads in a particular number of flips.
P(D) is the evidence. This is the probability of data as determined by summing (or integrating) across all possible values of θ, weighted by how strongly we believe in those particular values of θ.
If we had multiple views of what the fairness of the coin is (but didn't know for sure), then this tells us the probability of seeing a certain sequence of flips for all possibilities of our belief in the coin's fairness.
P(θ|D) is the posterior belief of our parameters after observing the evidence i.e the number of heads .
From here, we'll dive deeper into mathematical implications of this concept. Don't worry. Once you understand them, getting to its *mathematics* is pretty easy.
To define our model correctly , we need two mathematical models before hand. One to represent the ***likelihood function P(D|θ)*** and the other for representing the distribution of ***prior beliefs .*** The product of these two gives the ***posterior belief P(θ|D)*** distribution.
Since prior and posterior are both beliefs about the distribution of fairness of coin, intuition tells us that both should have the same mathematical form. Keep this in mind. We will come back to it again.
So, there are several functions which support the existence of bayes theorem. Knowing them is important, hence I have explained them in detail.

4.1. Bernoulli likelihood function

Lets recap what we learned about the likelihood function. So, we learned that:
It is the probability of observing a particular number of heads in a particular number of flips for a given fairness of coin. This means our probability of observing heads/tails depends upon the fairness of coin (θ).
P(y=1|θ)= θ^y [If coin is fair θ=0.5, probability of observing heads (y=1) is 0.5]
P(y=0|θ)= $(1-\theta)^{1-y}$ [If coin is fair θ=0.5, probability of observing tails(y=0) is 0.5]
It is worth noticing that representing 1 as heads and 0 as tails is just a mathematical notation to formulate a model. We can combine the above mathematical definitions into a single definition to represent the probability of both the outcomes.

$$P(y|\theta)=\theta^y.(1-\theta)^{1-y}$$

This is called the **Bernoulli Likelihood Function** and the task of coin flipping is called Bernoulli's trials.

$$y=\{0,1\},\theta=(0,1)$$

And, when we want to see a series of heads or flips, its probability is given by:

$$P(y_1, y_2, \ldots y_n|\theta) = \prod_1^n P(y_i|\theta)$$

Furthermore, if we are interested in the probability of number of heads *z* turning up in *N*number of flips then the probability is given by:

$$P(z, N|\theta) = \theta^z.(1-\theta)^{N-z}$$

4.2. Prior Belief Distribution

This distribution is used to represent our strengths on beliefs about the parameters based on the previous experience.

But, what if one has no previous experience?

Don't worry. Mathematicians have devised methods to mitigate this problem too. It is known as uninformative priors. I would like to inform you beforehand that it is just a misnomer. Every uninformative prior always provides some information event the constant distribution prior.

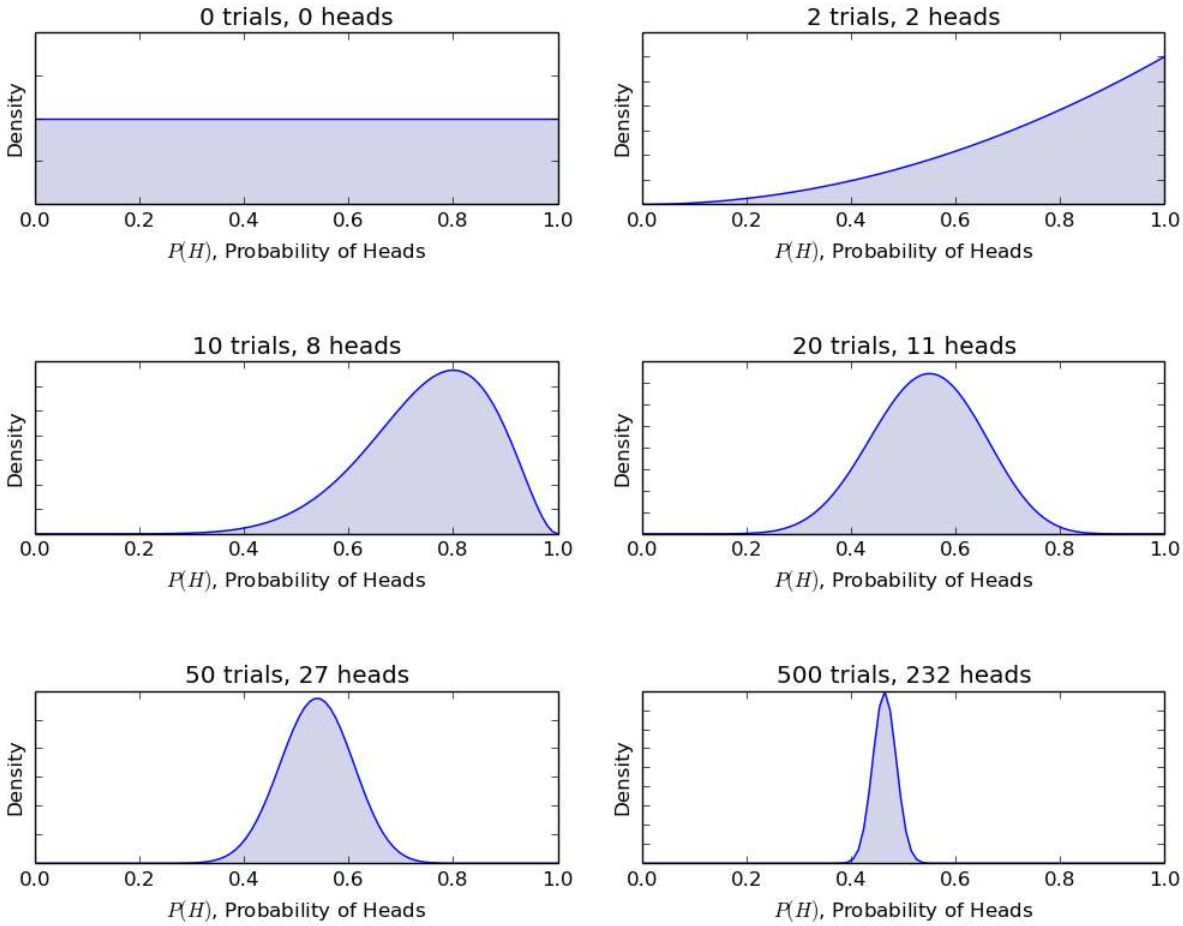
Well, the mathematical function used to represent the prior beliefs is known as *beta distribution***.** It has some very nice mathematical properties which enable us to model our beliefs about a binomial distribution.

Probability density function of beta distribution is of the form :

$$x^{\alpha-1}.(1-x)^{\beta-1}/B(\alpha,\beta)$$

where, our focus stays on numerator. The denominator is there just to ensure that the total probability density function upon integration evaluates to 1.

α and β are called the shape deciding parameters of the density function. Here α is analogous to number of heads in the trials and β corresponds to the number of tails. The diagrams below will help you visualize the beta distributions for different values of α and β



You too can draw the beta distribution for yourself using the following code in R:

```
> library(stats)
> par(mfrow=c(3,2))
> x=seq(0,1,by=o.1)
> alpha=c(0,2,10,20,50,500)
> beta=c(0,2,8,11,27,232)
> for(i in 1:length(alpha)){
      y<-dbeta(x,shape1=alpha[i],shape2=beta[i])
      plot(x,y,type="l")
}
```

Note: α and β are intuitive to understand since they can be calculated by knowing the mean (μ) and standard deviation (σ) of the distribution. In fact, they are related as :

$$\mu = \frac{\alpha}{\alpha + \beta}$$

If mean and standard deviation of a distribution are known , then there shape parameters can be easily calculated.

Inference drawn from graphs above:

1. When there was no toss we believed that every fairness of coin is possible as depicted by the flat line.
2. When there were more number of heads than the tails, the graph showed a peak shifted towards the right side, indicating higher probability of heads and that coin is not fair.

3. As more tosses are done, and heads continue to come in larger proportion the peak narrows increasing our confidence in the fairness of the coin value.

4.3. Posterior Belief Distribution

The reason that we chose prior belief is to obtain a beta distribution. This is because when we multiply it with a likelihood function, posterior distribution yields a form similar to the prior distribution which is much easier to relate to and understand. If this much information whets your appetite, I'm sure you are ready to walk an extra mile.
Let's calculate posterior belief using bayes theorem.
Calculating posterior belief using Bayes Theorem

$$P(\theta|z,N) = P(z,N|\theta)P(\theta)/P(z,N)$$
$$= \theta^z(1-\theta)^{N-z}.\theta^{\alpha-1}(1-\theta)^{\beta-1}/[B(\alpha,\beta)P(z,N)]$$
$$= \theta^{z+\alpha-1}(1-\theta)^{N-z+\beta-1}/[B(z+\alpha,N-z+\beta)]$$

Now, our posterior belief becomes,

$$P(\theta|z+\alpha,N-z+\beta)$$

This is interesting. Just knowing the mean and standard distribution of our belief about the parameter θ and by observing the number of heads in N flips, we can update our belief about the model parameter(θ).
Lets understand this with the help of a simple example:
Suppose, you think that a coin is biased. It has a mean (μ) bias of around 0.6 with standard deviation of 0.1.
Then ,
α= 13.8 , β=9.2
i.e our distribution will be biased on the right side. Suppose, you observed 80 heads (z=80) in 100 flips(N=100). Let's see how our prior and posterior beliefs are going to look:
prior = P(θ|α,β)=P(θ|13.8,9.2)
Posterior = P(θ|z+α,N-z+β)=P(θ|93.8,29.2)
Lets visualize both the beliefs on a graph:

The R code for the above graph is as:

```
> library(stats)
> x=seq(0,1,by=0.1)
> alpha=c(13.8,93.8)
> beta=c(9.2,29.2)
> for(i in 1:length(alpha)){
    y<-dbeta(x,shape1=alpha[i],shape2=beta[i])
    plot(x,y,type="l",xlab = "theta",ylab = "density")
}
```

As more and more flips are made and new data is observed, our beliefs get updated. This is the real power of Bayesian Inference.

5. Test for Significance – Frequentist vs Bayesian
Without going into the rigorous mathematical structures, this section will provide you a quick overview of different approaches of frequentist and bayesian methods to test for significance and difference between groups and which method is most reliable.

5.1. p-value

In this, the t-score for a particular sample from a sampling distribution of *fixed size* is calculated. Then, p-values are predicted. We can interpret p values as (taking an example of p-value as 0.02 for a distribution of mean 100) : There is 2% probability that the sample will have mean equal to 100.
This interpretation suffers from the flaw that for sampling distributions of different sizes, one is bound to get different t-score and hence different p-value. It is completely absurd. A p-value less than 5% does not guarantee that null hypothesis is wrong nor a p-value greater than 5% ensures that null hypothesis is right.

5.2. Confidence Intervals

Confidence Intervals also suffer from the same defect. Moreover since C.I is not a probability distribution , there is no way to know which values are most probable.

5.3. Bayes Factor

Bayes factor is the equivalent of p-value in the bayesian framework. Lets understand it in an comprehensive manner.
The *null hypothesis* in bayesian framework assumes ∞ probability distribution only at a particular value of a parameter (say θ=0.5) and a zero probability else where. (M1)
The *alternative hypothesis* is that all values of θ are possible, hence a flat curve representing the distribution. (M2)
Now, posterior distribution of the new data looks like below.

Bayesian statistics adjusted credibility (probability) of various values of θ. It can be easily seen that the probability distribution has shifted towards M2 with a value higher than M1 i.e M2 is more likely to happen.
Bayes factor does not depend upon the actual distribution values of θ but the magnitude of shift in values of M1 and M2.
In panel A (shown above): left bar (M1) is the prior probability of the null hypothesis.
In panel B (shown), the left bar is the posterior probability of the null hypothesis.
Bayes factor is defined as the ratio of the posterior odds to the prior odds,

$$BF = \frac{P(M = null|z, N)}{P(M = alt|z, N)} \Big/ \frac{P(M = null)}{P(M = alt)}$$

To reject a null hypothesis, a BF <1/10 is preferred.
We can see the immediate benefits of using Bayes Factor instead of p-values since they are independent of intentions and sample size.

5.4. High Density Interval (HDI)

HDI is formed from the posterior distribution after observing the new data. Since HDI is a probability, the 95% HDI gives the 95% most credible values. It is also guaranteed that 95 % values will lie in this interval unlike C.I.
Notice, how the 95% HDI in prior distribution is wider than the 95% posterior distribution. This is because our belief in HDI increases upon observation of new data.

Principal component analysis and neural networks:

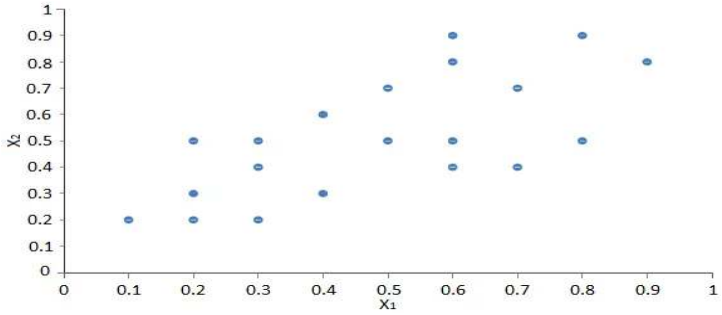
Principal components analysis (PCA) is a statistical technique that allows identifying underlying linear patterns in a data set so it can be expressed in terms of another data set of a significatively lower dimension without much loss of information.

The final data set should explain most of the variance of the original data set by reducing the number of variables. The final variables will be named as principal components.

To illustrate the whole process, we will make use of the following data set, with only 2 dimensions.

Instance	x_1	x_2		Instance	x_1	x_2
1	0.3	0.5		11	0.6	0.8
2	0.4	0.3		12	0.4	0.6
3	0.7	0.4		13	0.3	0.4
4	0.5	0.7		14	0.6	0.5
5	0.3	0.2		15	0.8	0.5
6	0.9	0.8		16	0.8	0.9
7	0.1	0.2		17	0.2	0.3
8	0.2	0.5		18	0.7	0.7
9	0.6	0.9		19	0.5	0.5
10	0.2	0.2		20	0.6	0.4

The next scatter chart shows the values of the variable x_2 against the values of the variable x_1.



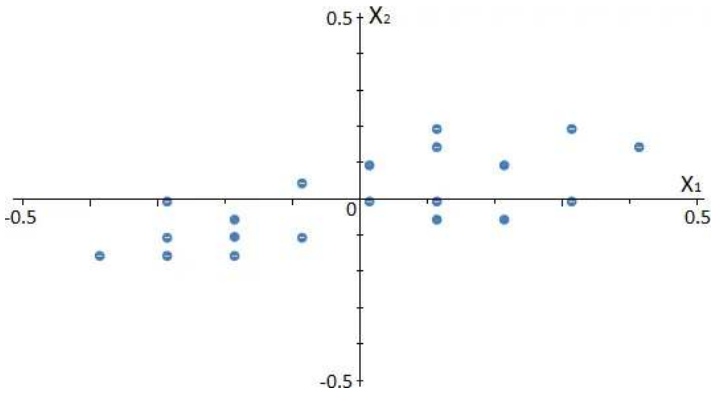
The objective is to convert that data set into a new one of only 1 dimension with minimal information loss.

The steps to perform principal components analysis are the following:

1. Subtract mean.
2. Calculate the covariance matrix.
3. Calculate eigenvectors and eigenvalues.
4. Select principal components.
5. Reduce the data dimension.

1. Subtract mean:

The first step in the principal component analysis is to subtract the mean for each variable of the data set.
The next scatter chart shows how the data is rearranged in our example.



As shown, the subtraction of the mean results in a data translation, which has zero mean.

2. Calculate the covariance matrix:

The covariance of two random variables measures the degree of variation from their means for each other.
The sign of the covariance provides us with information about the relation between them:

- If the covariance is positive, then the two variables increase and decrease together.
- If the covariance is negative, then when one variable increases, the other decreases, and vice versa.

These values determine the linear dependencies between the variables, which will be used to reduce the data set'sdimension.
Back to our example, the covariance matrix is shown next.

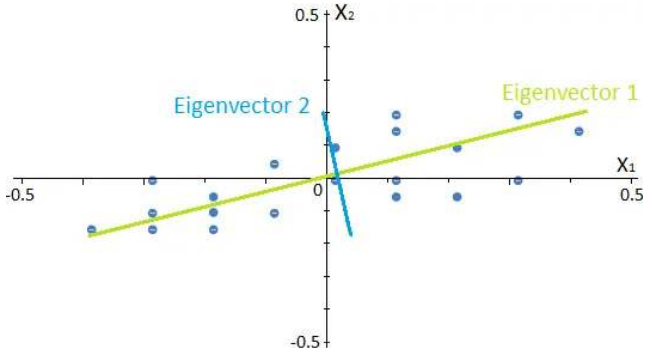
		x_1	x_2
x_1		0.33	0.25
x_2		0.25	0.41

The variance is a measure of how the data is spread from the mean.
The diagonal values show the covariance of each variable and itself, and they equal their variance.
The off-diagonal values show the covariance between the two variables. In this case, these values are positive, which means that both variables increase and decrease together.

3. Calculate eigenvectors and eigenvalues:

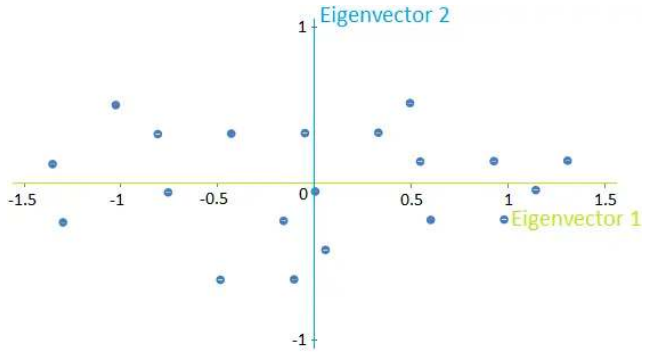
Eigenvectors are defined as those vectors whose directions remain unchanged after any linear transformation has been applied.
However, their length could not remain the same after the transformation, i.e., the result of this transformation is the vector multiplied by a scalar. This scalar is called eigenvalue, and each eigenvector has one associated with it.

The number of eigenvectors or components that we can calculate for each data set is equal to the data set's dimension. In this case, we have a 2-dimensional data set, so the number of eigenvectors will be 2. The next image represents the eigenvectors for our example.



Since they are calculated from the covariance matrix described before, eigenvectors represent the directions in which the data have a higher variance. On the other hand, their respective eigenvalues determine the data set's variance in that direction.

Once we have obtained these new directions, we can plot the data in terms of them, as shown in the following image for our example.



Note that the data has not changed; we are rewriting them in terms of these new directions instead of the previous x_1-x_2 directions.

4. Select principal components:

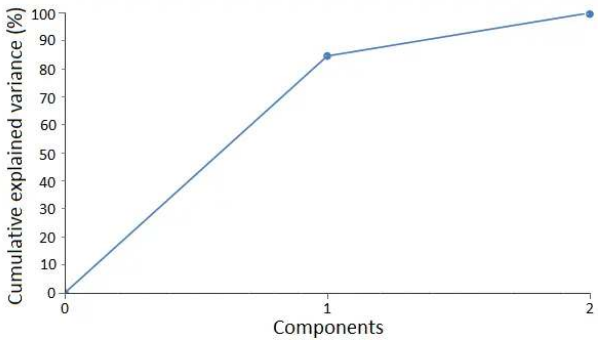
Among the available eigenvectors that were previously calculated, we must select those onto which we project the data. The selected eigenvectors will be called principal components.

To establish a criterion to select the eigenvectors, we must first define the relative variance of each and the total variance of a data set. The relative variance of an eigenvector measures how much information can be attributed to it. The total variance of a data set is the sum of the variance of all the variables.

These two concepts are determined by the eigenvalues. For our example, the following table shows the relative and cumulative variance for each eigenvector.

	Relative variance	**Cumulative variance**
PC_1	84.60	84.60
PC_2	15.40	100

As we can see, the first eigenvector can explain almost 85% of all the data's variance, while the second eigenvector explains around 15% of it. The following graph shows the cumulative variance for the components.



A common way to select the variables is to establish the amount of information that we want the final data set to explain. If this amount of information decreases, the number of principal components that we select will decrease.
In this case, as we want to reduce the 2-dimensional data set into a 1-dimensional data set, we will select the first eigenvector as the principal component.
Consequently, the final reduced data set will explain 85% of the variance of the original one.

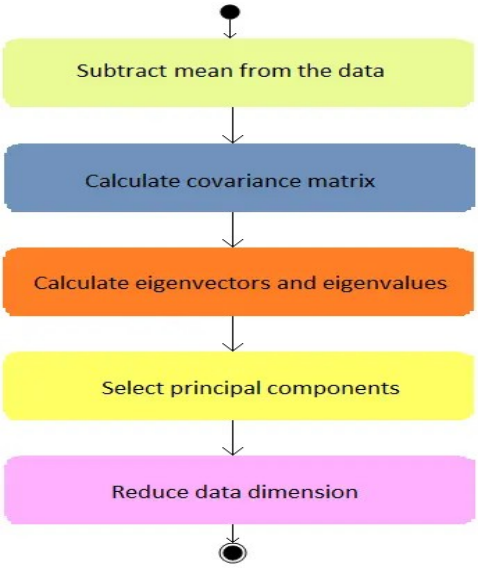
5. Reduce data dimension

Once we have selected the principal components, the data must be projected onto them. The following image shows the result of this projection for our example.

Although this projection can explain most of the variance of the original data, we have lost the information about the variance along with the second component. In general, this process is irreversible, which means that we cannot recover the original data from the projection.

Conclusions:

Principal components analysis is a technique that allows us to identify the underlying dependencies of a data set and to reduce its dimensionality significantly attending to them.
The following diagram summarizes the activities that need to be performed in principal components analysis.



This technique is beneficial for processing data sets with hundreds of variables while maintaining, at the same time, most of the information from the original data set.
Principal components analysis can also be implemented within a neural network. However, as this process is irreversible, the data's reduction should be done only for the inputs and not for the target variables.

Fuzzy Logic

https://www.studocu.com/in/document/university-of-madras/computer-application/unit-3-big-of-the-data-of-

the-analytics-of-the-notes-of-the-unit-of-the-3-of-the-bachelor/29194790

The term fuzzy refers to things that are not clear or are vague. In the real world many times we encounter a situation when we can9t determine whether the

state is true or false, their
fuzzy logic provides very valuable flexibility for reasoning. In this way, we can consider the inaccuracies and uncertainties of any situation.
In the Boolean system truth value, 1.0 represents the absolute truth value and 0.0 represents the absolute false value. But in the fuzzy system, there is no logic for the

absolute truth and absolute false value. But in fuzzy logic, there is an intermediate value too present which is partially true and partially false.

ARCHITECTURE

Its Architecture contains four parts:

→ RULE BASE: It contains the set of rules and the IF-THEN conditions provided by the experts to govern the decision-making system,

on the basis of linguistic information.
Recent developments in fuzzy theory offer several effective methods for the design and
tuning of fuzzy
controllers. Most of these developments reduce the number of fuzzy rules.