Data Viewpoints: Web Based Mathematical Software for Data Analysis

PROJECT ID: 9387

B.Tech. Project Report
submitted in partial fulfillment of
the requirements for the
Degree of Bachelor of Technology
Under Biju Patnaik University of Technology

By

Bhawnesh Dipu **Roll No. CSE 201316665**

Debadeepa Rath **Roll No. CSE 201311643**

2016- 2017

Under the guidance of

Dr. Motahar Reza

NATIONAL INSTITUTE OF SCIENCE & TECHNOLOGY
Palur Hills, Berhampur- 761008, Odisha, India

ABSTRACT

Data visualization is increasingly an essential element of business intelligence (BI). No longer restricted to specialized applications, data visualization in the form of charts, maps, and other graphical representations is enabling business users to better understand data and use it to achieve tactical and strategic objectives. Moreover, data visualization is prompting a cultural shift toward more analytic, data-driven business and operations by empowering users to explore, in a graphically inviting medium, data that was previously available only in tabular reports.

In WebMetica , following operations will be implemented: - Functions plotting , Graph Plotting, parametric plotting, Artificial Neural Network for Stock Prediction, Regression ,Interpolation ,Clustering visualization and 3D Plotting.

ACKNOWLEDGEMENT

We would like to take this opportunity to thanks all those individuals whose invaluable contribution in a direct or indirect manner has gone into the making of this project report is a tremendous learning experience for us.

We give our sincere thanks to **Dr. Motahar Reza, Project Advisor** for giving us the opportunity and motivating us to complete the project within stipulated period of time and providing a helping environment.

We give our sincere thanks to **Dr. Sandipan Mallik, Project Coordinator**, for helping us throughout our project and encouraging us to complete this project.

We acknowledge with immense pleasure the sustained interest, encouraging attitude and constant inspiration rendered by **Prof. Sangram Mudali** (Director) &**Prof. Geetika Mudali** (Placement Director)N.I.S.T. Their continued drive for better quality in everything that happens at N.I.S.T. and selfless inspiration has always helped us to move ahead.

Bhawnesh Dipu

Debadeepa Rath

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. With interactive visualization, you can take the concept a step further by using technology to drill down into charts and graphs.

It's technology, however, that truly lit the fire under data visualization. Computers made it possible to process large amounts of data at lightning-fast speeds. Today, data visualization has become a rapidly evolving blend of science and art that is certain to change the corporate landscape over the next few years.

Why is data visualization important?

The way the human brain processes information, using charts or graphs to visualize large amounts of complex data is easier than poring over spreadsheets or reports. Data visualization is a quick, easy way to convey concepts in a universal manner – and you can experiment with different scenarios by making slight adjustments.

Data visualization can also:

- Identify areas that need attention or improvement.
- Clarify which factors influence customer behavior.
- Help you understand which products to place where.
- Predict sales volumes.

Data visualization is going to change the way our analysts work with data. They are going to be expected to respond to issues more rapidly. And they will need to be able to dig for more insights – look at data differently, more imaginatively. Data visualization will promote that creative data exploration.

CHAPTER 2
BASIC METHODS

2.1 Line Plotting

A line plot is a graphical display of data along a number line with Xs or dots recorded above the responses to indicate the number of occurrences a response appears in the data set. The Xs or dots represent the frequency. A line plot will have an outlier. An outlier is a number that is much greater or much less than the other numbers in the data set. A line plot consists of a horizontal line which is the x-axis with equal intervals. It is important for a line to plot to have a title and a label of the x-axis to provide the reader an overview of what is being displayed. Also, line plots must have legends to explain what is being measured.

2.1.1 Examples of a Line Plot

Table 2.1 Data for Line Plot

Horizontal label	Vertical label
0	1
1	2
2	3
3	4
4	5
5	6

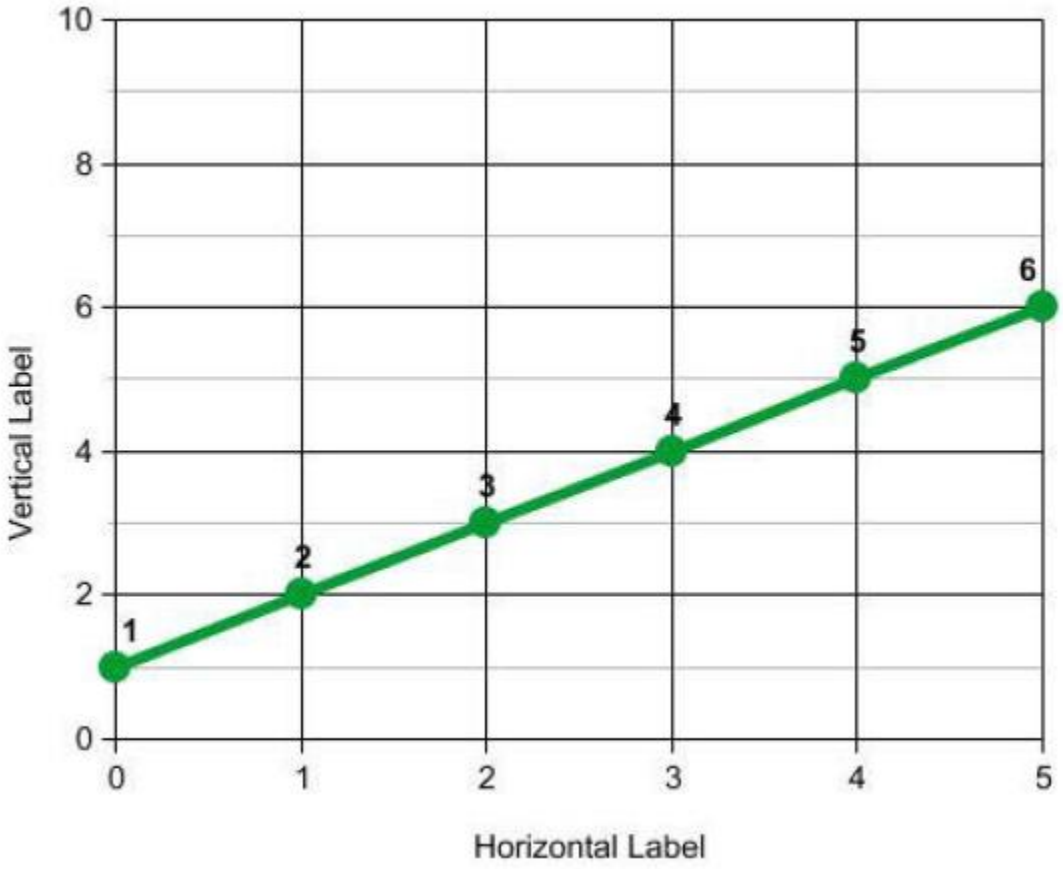


Figure 2.1 Horizontal vs Vertical Label

2.2 Scatter Plotting

A scatter plot is used to graphically represent the relationship between two variables. Explore the relationship between scatterplots and correlations, the different types of correlations, how to interpret scatterplots, and more. A scatter plot (also called a scatter graph, scatter chart, scattergram, or scatter diagram)[3] is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are colorcoded, one additional variable can be displayed. The data is displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.[4]

A scatter plot can be used either when one continuous variable that is under the control of the experimenter and the other depends on it or when both continuous variables are

independent. If a parameter exists that is systematically incremented and/or decremented by the other, it is called the control parameter or independent variable and is customarily plotted along the horizontal axis. The measured or dependent variable is customarily plotted along the vertical axis. If no dependent variable exists, either type of variable can be plotted on either axis and a scatter plot will illustrate only the degree of correlation (not causation) between two variables.

One of the most powerful aspects of a scatter plot, however, is its ability to show nonlinear relationships between variables. The ability to do this can be enhanced by adding a smooth line such as LOESS.[5] Furthermore, if the data are represented by a mixture model of simple relationships, these relationships will be visually evident as superimposed patterns. The scatter diagram is one of the seven basic tools of quality control.[6] Scatter charts can be built in the form of bubble, marker, or/and line charts.[7]

2.2.1 Scatter plot

Imagine that you are interested in studying patterns in individuals with children under the age of 10. You collect data from 25 individuals who have at least one child. After you have collected your data, you enter it into a table.

One variable is plotted on each axis. Scatterplots are made up of marks; each mark represents one study participant's measures on the variables that are on the x-axis and yaxis of the scatterplot.

Most scatterplots contain a line of best fit, which is a straight line drawn through the center of the data points that best represents the trend of the data. Scatterplots provide a visual representation of the correlation, or relationship between the two variables.

2.3 Bar Plotting

A bar chart or bar graph is a chart or graph that presents grouped data with rectangular bars with lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a Line graph.

A bar graph is a chart that uses either horizontal or vertical bars to show comparisons among categories. One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value. Some bar graphs present bars clustered in groups of more than one.

A bar graph is a chart that uses bars to show comparisons between categories of data. The bars can be either horizontal or vertical. Bar graphs with vertical bars are sometimes called vertical bar graphs. A bar graph will have two axes. One axis will describe the types of categories being compared, and the other will have numerical values that represent the values of the data. It does not matter which axis is which, but it will determine what bar graph is shown. If the descriptions are on the horizontal axis, the bars will be oriented vertically, and if the values are along the horizontal axis, the bars will be oriented horizontally.

2.3.1 Types of Bar Graphs

There are many different types of bar graphs. They are not always interchangeable. Each type will work best with a different type of comparison. The comparison you want to make will help determine which type of bar graph to use. First we'll discuss some simple bar graphs.

A simple vertical bar graph is best when you have to compare between two or more independent variables. Each variable will relate to a fixed value. The values are positive and therefore, can be fixed to the horizontal value.

2.3.2 Vertical Bar Graph

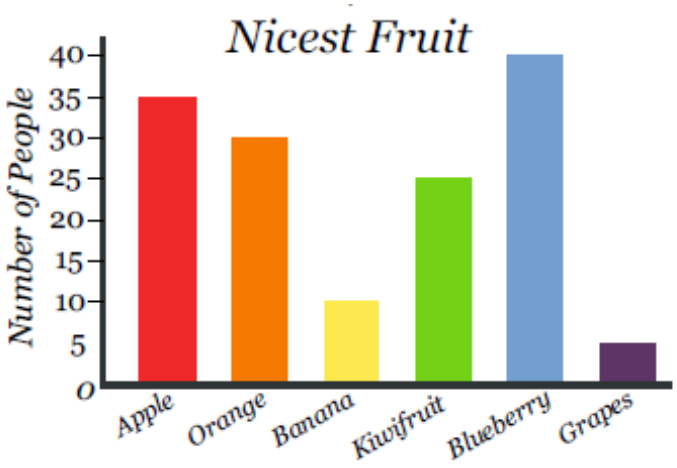


Figure 2.2 Vertical Bar Graph

If your data has negative and positive values but is still a comparison between two or more fixed independent variables, it is best suited for a horizontal bar graph. The vertical axis can be oriented in the middle of the horizontal axis, allowing for negative and positive values to be represented.

2.3.3 Horizontal bar

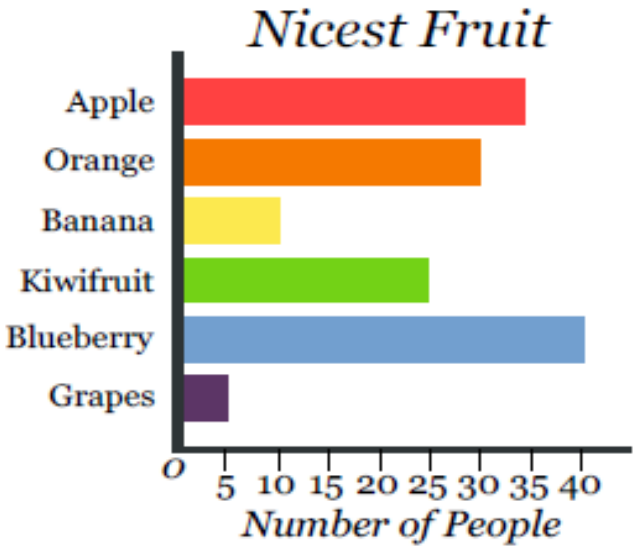


Figure 2.3 Horizontal Bar Graph

A range bar graph represents a range of data for each independent variable. Temperature ranges or price ranges are common sets of data for range graphs. Unlike the above graphs, the data do not start from a common zero pole.

2.4 Pie Chart

A pie chart (or a circle chart) is a circular statistical graphic which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents. While it is named for its resemblance to a pie which has been sliced, there are variations on the way it can be presented. The earliest known pie chart is generally credited to William Playfair's *Statistical Breviary* of 1801.[1][2]

Pie charts are very widely used in the business world and the mass media.[3] However, they have been criticized,[4] and many experts recommend avoiding them,[5][6][7][8] pointing out that research has shown it is difficult to compare different sections of a given pie chart, or to compare data across different pie charts. Pie charts can be replaced in most cases by other plots such as the bar chart, box plot or dot plots.

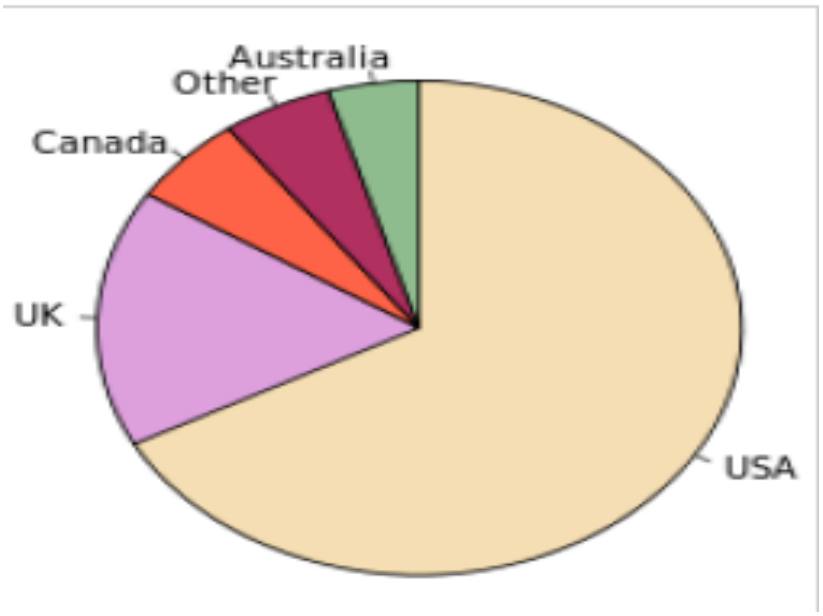
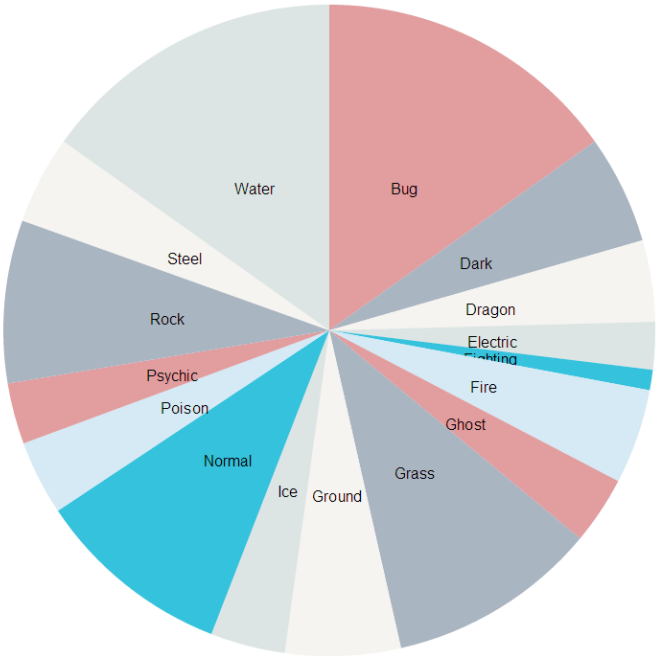


Figure 2.4 Pie Chart of Population of English Native Speakers



CHAPTER 3
STATISTICAL METHOD

3.1 Histogram

A histogram is a graphical representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable (quantitative variable) and was first introduced by Karl Pearson.[1] It is a kind of bar graph. To construct a histogram, the first step is to "bin" the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval. The bins are usually specified as consecutive, nonoverlapping intervals of a variable. The bins (intervals) must be adjacent, and are often (but are not required to be) of equal size.[2]

If the bins are of equal size, a rectangle is erected over the bin with height proportional to the frequency — the number of cases in each bin. A histogram may also be normalized to display "relative" frequencies. It then shows the proportion of cases that fall into each of several categories, with the sum of the heights equaling 1.

However, bins need not be of equal width; in that case, the erected rectangle is defined to have its *area* proportional to the frequency of cases in the bin.[3] The vertical axis is then not the frequency but *frequency density* — the number of cases per unit of the variable on the horizontal axis. Examples of variable bin width are displayed on Census bureau data below.

As the adjacent bins leave no gaps, the rectangles of a histogram touch each other to indicate that the original variable is continuous.[4]

Histograms are sometimes confused with bar charts. A histogram is used for continuous data, where the bins represent ranges of data, while a bar chart is a plot of categorical variables. Some authors recommend that bar charts have gaps between the rectangles to clarify the distinction.

EXAMPLE:

Table 3.1 Bin Count

BIN	COUNT
-3.5	23
-2.5	32
-1.5	109
-0.5	180
1.5	34
2.5	4
3.5	90

The words used to describe the patterns in a histogram are: "symmetric", "skewed left" or "right", "unimodal", "bimodal" or "multimodal".

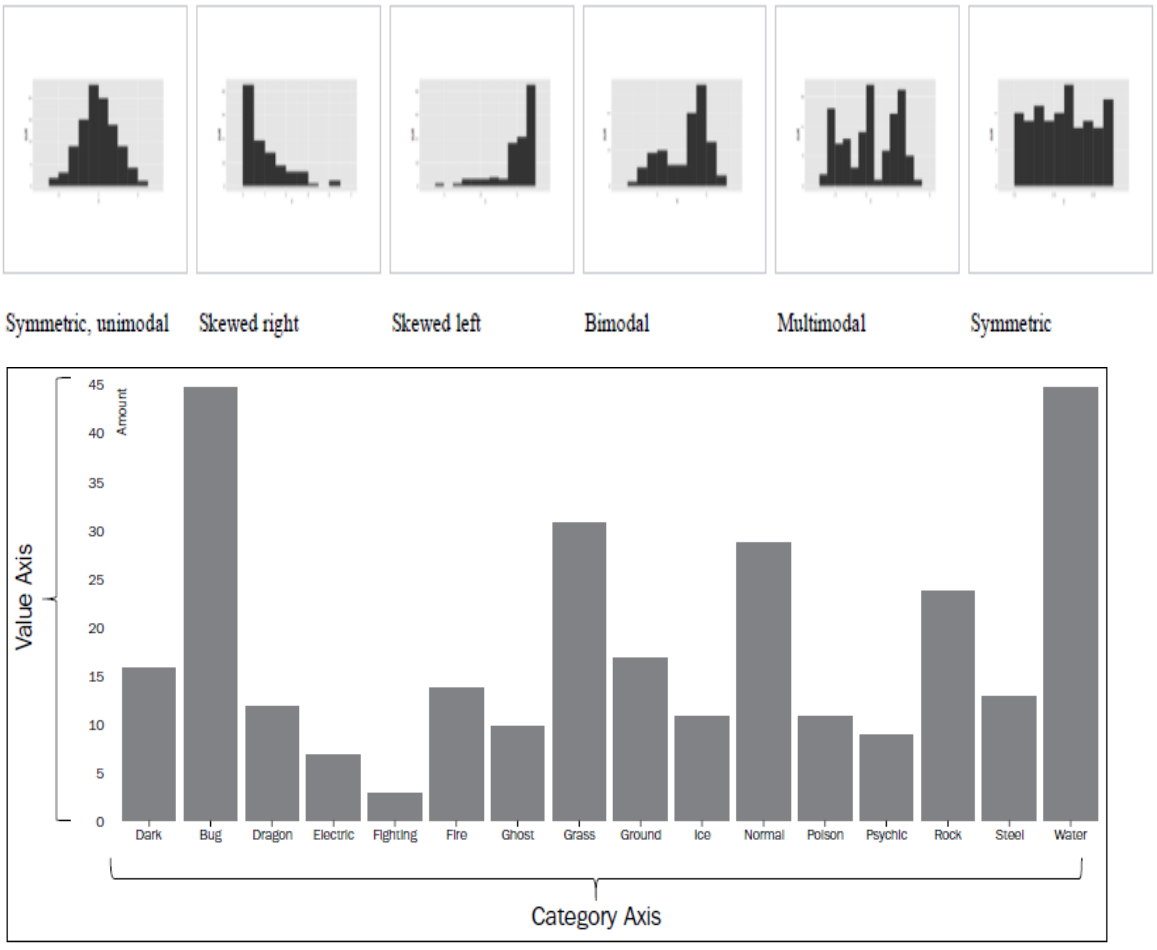


Figure 3.1 Histogra

CHAPTER 4
SCIENTIFIC METHODS

4.1 Contour Plotting

A contour line (also isoline, isopleth, or isarithm) of a function of two variables is a curve along which the function has a constant value.[1][2] It is a crosssection of the three dimensional graph of the function $f(x, y)$ parallel to the x, y plane. In cartography, a contour line (often just called a "contour") joins points of equal elevation (height) above a given level, such as mean sea level.[3] A contour map is a map illustrated with contour lines, for example a topographic map, which thus shows valleys and hills, and the steepness or gentleness of slopes.[4] The contour interval of a contour map is the difference in elevation between successive contour lines.[5]

More generally, a contour line for a function of two variables is a curve connecting points where the function has the same particular value.[2]

The gradient of the function is always perpendicular to the contour lines. When the lines are close together the magnitude of the gradient is large: the variation is steep. A level set is a generalization of a contour line for functions of any number of variables.

Contour lines are curved, straight or a mixture of both lines on a map describing the intersection of a real or hypothetical surface with one or more horizontal planes. The configuration of these contours allows map readers to infer relative gradient of a parameter and estimate that parameter at specific places. Contour lines may be either traced on a visible three dimensional model of the surface, as when a photogrammetrist viewing a stereomodel plots elevation contours, or interpolated from estimated surface elevations, as when a computer program threads contours through a network of observation points of area centroids. In the latter case, the method of interpolation affects the reliability of individual isolines and their portrayal of slope, pits and peaks.

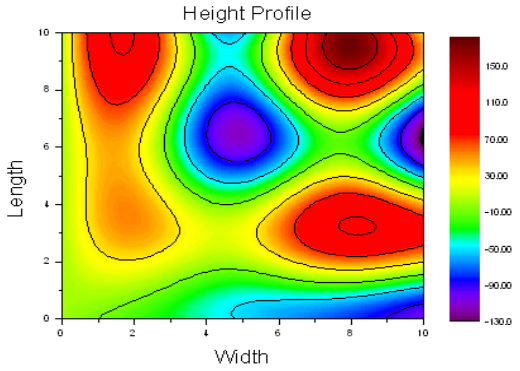


Figure 4.1 Contour Plotting

4.2 Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi objective optimization that involves trial and failure. It is often necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

.

4.2.1 Definition

Typical cluster models include:

- *Connectivity models*: for example, hierarchical clustering builds models based on distance connectivity.
- *Centroid models*: for example, the kmeans algorithm represents each cluster by a single mean vector.
- *Distribution models*: clusters are modeled using statistical distributions, such as multivariate normal distributions used by the Expectation maximization algorithm.
- *Density models*: for example, DBSCAN and OPTICS defines clusters as connected dense regions in the data space.
- *Subspace models*: in Biclustering (also known as Coclustering or twomodeclustering), clusters are modeled with both cluster members and relevant attributes.
- *Group models*: some algorithms do not provide a refined model for their results and just provide the grouping information.
- *Graphbased models*: a clique, that is, a subset of nodes in a graph such that every two nodes in the subset are connected by an edge can be considered as a prototypical form of cluster. Relaxations of the complete connectivity requirement (a fraction of the edges can be missing) are known as quasicliques, as in the HCS clustering algorithm.

A "clustering" is essentially a set of such clusters, usually containing all objects in the data set. Additionally, it may specify the relationship of the clusters to each other, for example, a hierarchy of clusters embedded in each other.

Clusterings can be roughly distinguished as:

- *Hard clustering*: each object belongs to a cluster or not.
- *Soft clustering* (also: *fuzzy clustering*): each object belongs to each cluster to a certain degree (for example, a likelihood of belonging to the cluster).

There are also finer distinctions possible, for example:

- *Strict partitioning clustering*: each object belongs to exactly one cluster
- *Strict partitioning clustering with outliers*: objects can also belong to no cluster, and are considered outliers.
- *Overlapping clustering* (also: *alternative clustering, multiview clustering*): objects may belong to more than one cluster; usually involving hard clusters.
- *Hierarchical clustering*: objects that belong to a child cluster also belong to the parent cluster.
- *Subspace clustering*: while an overlapping clustering, within a uniquely defined subspace, clusters are not expected to overlap.

4.2.2 Connectivity based clustering (hierarchical clustering)

Connectivity based clustering, also known as *hierarchical clustering*, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the yaxis marks the distance at which the clusters merge, while the objects are placed along the xaxis such that the clusters don't mix.

Connectivity based clustering is a whole family of methods that differ by the way distances are computed. Apart from the usual choice of distance functions, the user also needs to decide on the linkage criterion (since a cluster consists of multiple objects, there are multiple candidates to compute the distance to) to use. Popular choices are known as singlelinkage clustering (the minimum of object distances), complete linkage clustering (the maximum of object distances) or UPGMA ("Unweighted Pair Group Method with Arithmetic Mean", also known as average linkage clustering). Furthermore, hierarchical clustering can be agglomerative (starting with single elements and aggregating them into clusters) or divisive (starting with the complete data set and dividing it into partitions).

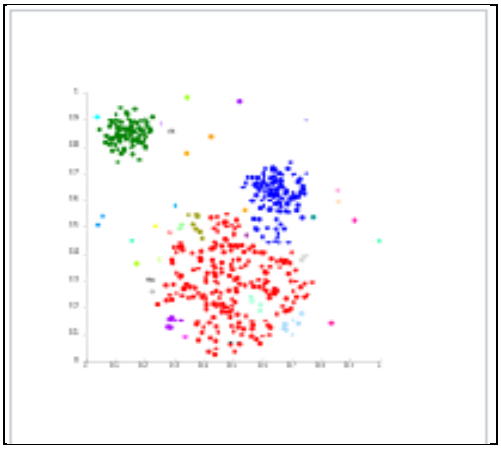


Figure 4.2 Single linkage on Gaussian Data

Single linkage on Gaussian data. At 35 clusters, the biggest cluster starts fragmenting into smaller parts, while before it was still connected to the second largest due to the single link effect.

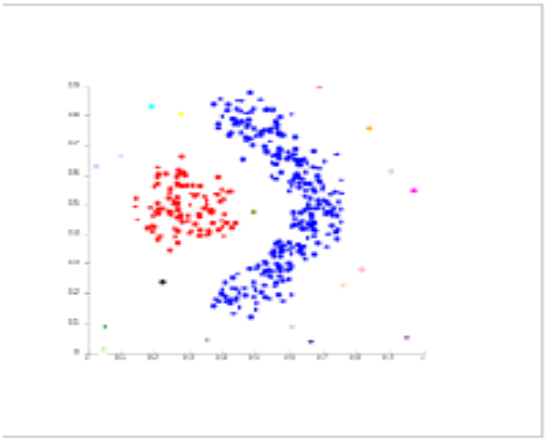


Figure 4.3 Single Linkage on Density based Clusters

Single linkage on density based clusters. 20 clusters extracted, most of which contain single elements, since linkage clustering does not have a notion of "noise".

4.2.3 Centroid based clustering

In centroidbased clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, kmeans clustering gives a formal definition as an optimization problem: find the cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized. The optimization problem itself is known to be NPhard, and thus the common approach is to search only for approximate solutions. A particularly

well known approximative method is Lloyd's algorithm,[8] often actually referred to as "kmeans algorithm". It does however only find a local optimum, and is commonly run multiple times with different random initializations. Variations of kmeans often include such optimizations as choosing the best of multiple runs, but also restricting the centroids to members of the data set (kmedoids), choosing medians (kmedians clustering), choosing the initial centers less randomly (Kmeans++) or allowing a fuzzy cluster assignment (Fuzzy cmeans).

4.4 K-Means Clustering Examples

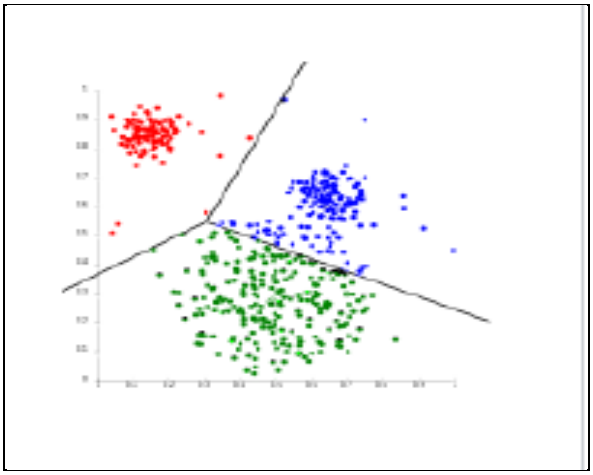


Figure 4.4 K-Means Algorithm

Kmeans separates data into Voronoicells, which assumes equalsized clusters (not adequate here).

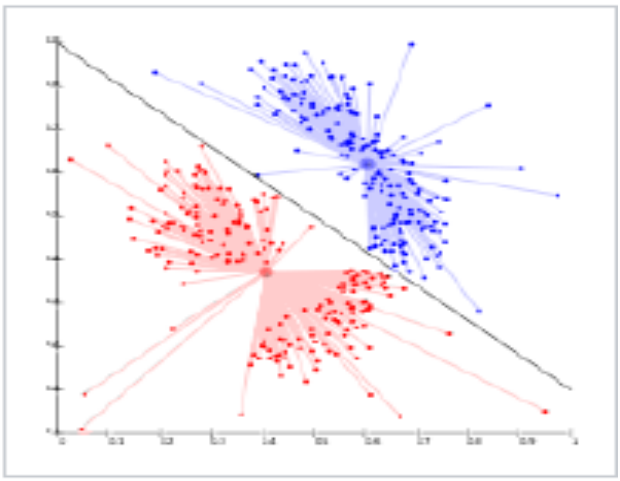


Figure 4.5 K means cannot represent density based clusters

4.4.1 Distribution based Clustering

The clustering model most closely related to statistics is based on distribution models. Clusters can then easily be defined as objects belonging most likely to the same distribution. A convenient property of this approach is that this closely resembles the way artificial data sets are generated: by sampling random objects from a distribution.

While the theoretical foundation of these methods is excellent, they suffer from one key problem known as overfitting, unless constraints are put on the model complexity. A more complex model will usually be able to explain the data better, which makes choosing the appropriate model complexity inherently difficult.

One prominent method is known as Gaussian mixture models (using the expectationmaximization algorithm). Here, the data set is usually modelled with a fixed (to avoid overfitting) number of Gaussian distributions that are initialized randomly and whose parameters are iteratively optimized to fit better to the data set. This will converge to a local optimum, so multiple runs may produce different results. In order to obtain a hard clustering, objects are often then assigned to the Gaussian distribution they most likely belong to; for soft clusterings, this is not necessary.

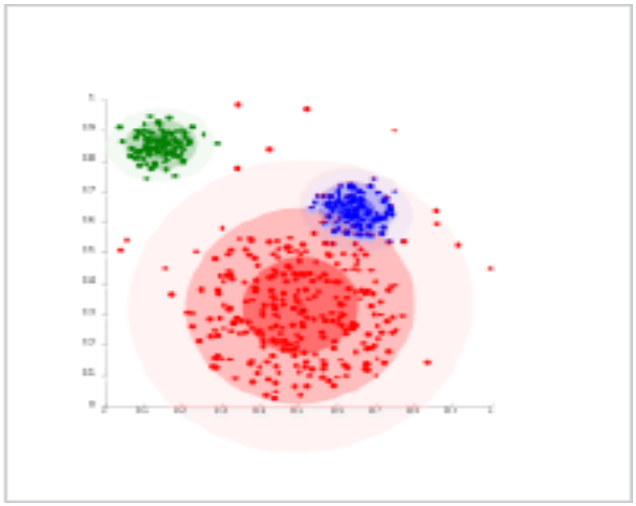


Figure 4.6 On Gaussian distributed data, EM works well, since it uses Gaussians for modelling clusters

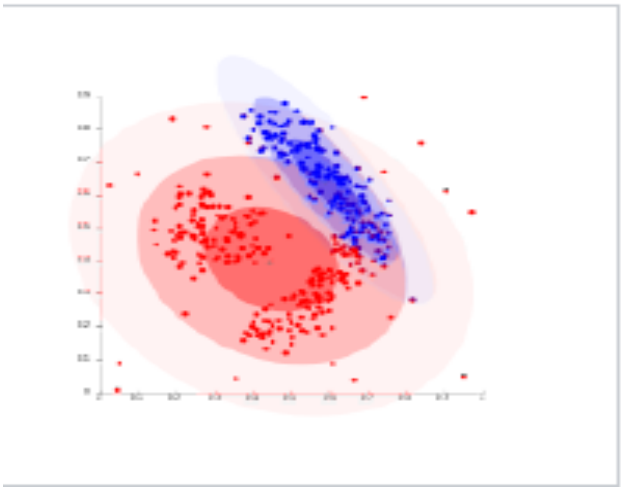


Figure 4.7 Density based clusters cannot be modeled using Gaussian distributions

4.4.2 Density based clustering

In density based clustering,[9] clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas that are required to separate clusters are usually considered to be noise and border points.

The most popular[10] density based clustering method is DBSCAN.[11] In contrast to many newer methods, it features a well defined cluster model called "density reachability".

Similar to linkage based clustering, it is based on connecting points within certain distance thresholds. However, it only connects points that satisfy a density criterion, in the original variant defined as a minimum number of other objects within this radius. A cluster consists of all

Density connected objects (which can form a cluster of an arbitrary shape, in contrast to many other methods) plus all objects that are within these objects' range. Another interesting property of DBSCAN is that its complexity is fairly low it requires a linear number of range queries on the database and that it will discover essentially the same results (it is deterministic for core and noise points, but not for border points) in each run, therefore there is no need to run it multiple times. OPTICS[12] is a generalization of DBSCAN that removes the need to choose an appropriate value for the range parameter , and produces a hierarchical result related to that of linkage clustering. DeLiClu,[13] Density Link Clustering combines ideas from single linkage clustering and OPTICS, eliminating the parameter entirely and offering performance improvements over OPTICS by using an Rtree index.

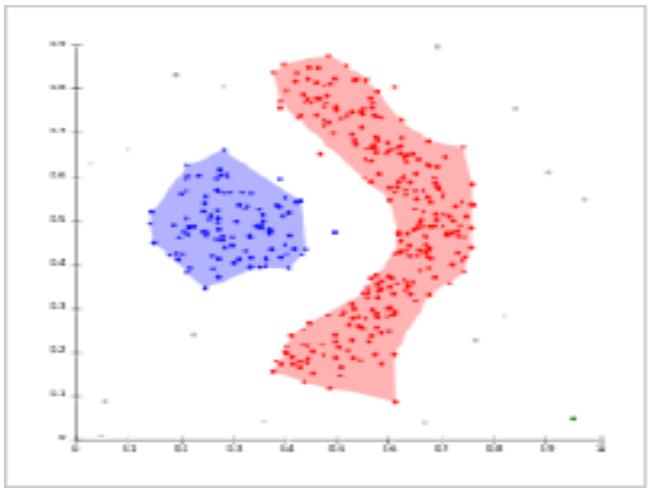


Figure 4.8 Density based clustering with DBSCAN.

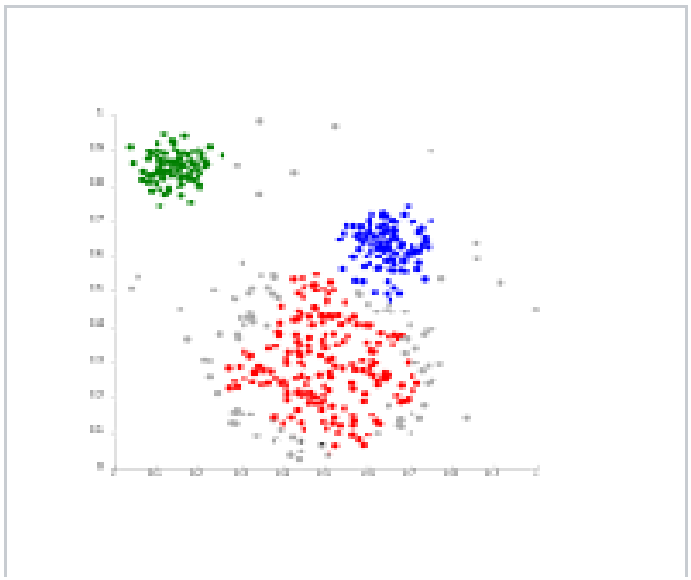


Figure 4.9 DBSCAN assumes clusters of similar density, and may have problems separating nearby clusters

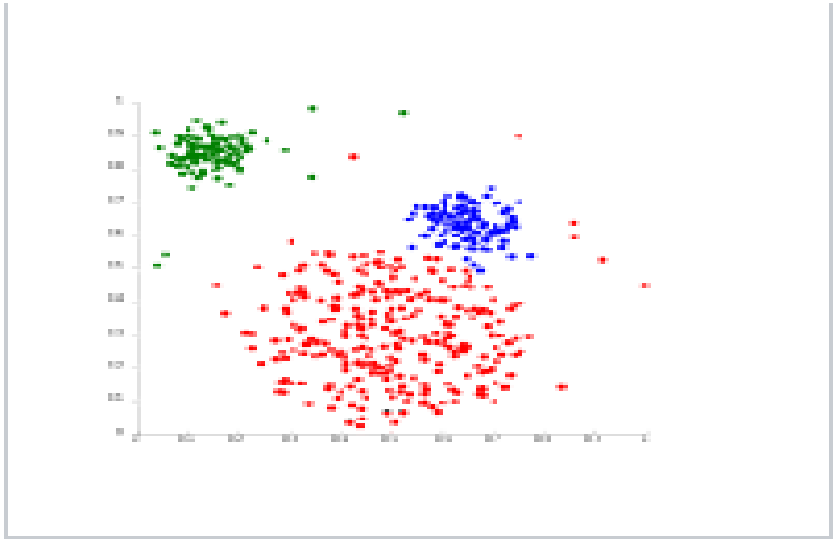


Figure 4.10 OPTICS is a DBSCAN variant that handles different densities much better.

CHAPTER 5
INTERPOLATION

In the mathematical field of numerical analysis, interpolation is a method of constructing new data points within the range of a discrete set of known data points. In engineering and science, one often has a number of data points, obtained by sampling or experimentation, which represent the values of a function for a limited number of values of the independent variable. It is often required to interpolate (i.e. estimate) the value of that function for an intermediate value of the independent variable. A different problem which is closely related to interpolation is the approximation of a complicated function by a simple function. Suppose the formula for some given function is known, but too complex to evaluate efficiently. A few known data points from the original function can be used to create an interpolation based on a simpler function. Of course, when a simple function is used to estimate data points from the original, interpolation errors are usually present; however, depending on the problem domain and the interpolation method used, the gain in simplicity may be of greater value than the resultant loss in precision. In the examples below if we consider x as a topological space and the function f forms a different kind of Banach spaces then the problem is treated as "interpolation of operators". The classical results about interpolation of operators are the Riesz–Thorin theorem and the Marcinkiewicz theorem. There are also many other subsequent results.

5.1 Linear interpolation

One of the simplest methods is linear interpolation (sometimes known as lerp).

Generally, linear interpolation takes two data points, say (*xa*,*ya*) and (*xb*,*yb*), and the interpolant is given by:

$Y= y_a+(y_b-y_a) (x-x_a)/(x_b-x_a)$ at the point(x,y).

$y-y_a / y_{b-}y_a = x-x_a / x_b-x_a$

$y-y_a / x-x_a = y_b-ya / x_b-x_a$

This previous equation states that the slope of the new line between$(x_{a,}x_{b)}$ and is the same as the slope of the line between and(x,y).

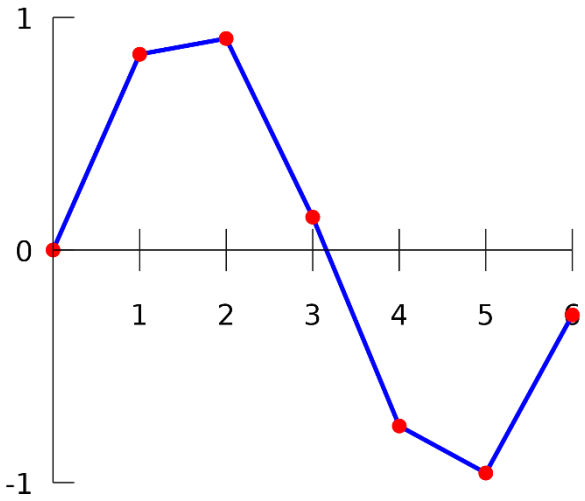


Figure 5.1 Linear Interpolation

5.2 Polynomial interpolation

Polynomial interpolation is a generalization of linear interpolation. Generally, if we have *n* data points, there is exactly one polynomial of degree at most *n*−1 going through all the data points. The interpolation error is proportional to the distance between the data points to the power *n*. Furthermore, the interpolant is a polynomial and thus infinitely differentiable. So, we see that polynomial interpolation overcomes most of the problems of linear interpolation.

However, polynomial interpolation also has some disadvantages. Calculating the interpolating polynomial is computationally expensive (see computational complexity) compared to linear interpolation. Furthermore, polynomial interpolation may exhibit oscillatory artifacts, especially at the end points.

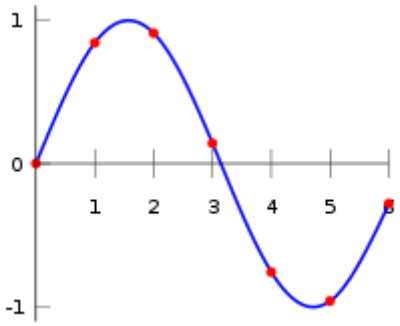


Figure 5.2 Polynomial Interpolation

CHAPTER 6
LINEAR REGRESSION

In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable *y* and one or more explanatory variables (or independent variables) denoted *X*. The case of one explanatory variable is called *simple linear regression*. For more than one explanatory variable, the process is called *multiple linear regression*.[1] (This term is distinct from *multivariate linear regression*, where multiple correlated dependent variables are predicted, rather than a single scalar variable.)[2]

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called *linear models*.[3] Most commonly, the conditional mean of *y* given the value of *X* is assumed to be an affine function of *X*; less commonly, the median or some other quantile of the conditional distribution of *y* given *X* is expressed as a linear function of *X*. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of *y* given *X*, rather than on the joint probability distribution of *y* and *X*, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.[4] This is because models which depend linearly on their unknown parameters are easier to fit than models which are nonlinearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

- If the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of *y* and *X* values. After developing such a model, if an additional value of *X* is then given without its accompanying value of *y*, the fitted model can be used to make a prediction of the value of *y*.

- Given a variable *y* and a number of variables *X*1, ..., *Xp* that may be related to *y*, linear regression analysis can be applied to quantify the strength of the relationship between *y* and the *Xj*, to assess which *Xj* may have no relationship with *y* at all, and to identify which subsets of the *Xj* contain redundant information about *y*.

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares loss function as in ridge regression (*L*2norm penalty) and lasso (*L*1norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

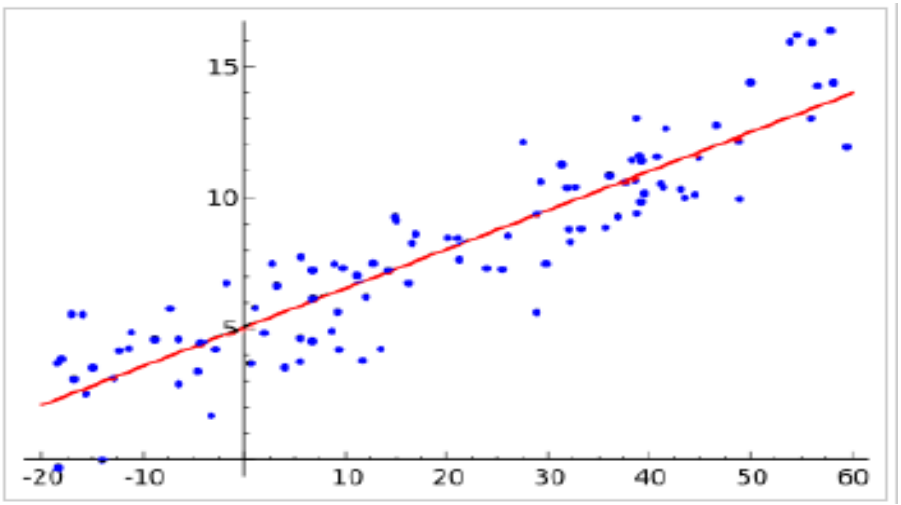


Figure 6.1 Example of Simple Linear Regression, which has one independent Variable

CHAPTER 7
POLYNOMIAL REGRESSION

In statistics, polynomial regression is a form of linear regression in which the relationship between the independent variable *x* and the dependent variable *y* is modelled as an *n*th degree polynomial in *x*. Polynomial regression fits a nonlinear relationship between the value of *x* and the corresponding conditional mean of *y*, denoted E(*y* |*x*), and has been used to describe nonlinear phenomena such as the growth rate of tissues,[1] the distribution of carbon isotopes in lake sediments,[2] and the progression of disease epidemics.[3] Although *polynomial regression* fits a nonlinear model to the data, as a statistical estimation problem it is linear, in the sense that the regression function E(*y* | *x*) is linear in the unknown parameters that are estimated from the data. For this reason, polynomial regression is considered to be a special case of multiple linear regression.

The polynomial regression model

$Y_i = a_0+a_1x_i+a_2x_i^m+\ldots\ldots+a_mx_i^m+\varepsilon_i (i=1,2,3\ldots..n)$

can be expressed in matrix form in terms of a design matrix , a response vector , a parameter vector , and a vector of random errors. The *i*th row of X and will contain the *x* and *y* value for the *i*th data sample. Then the modelcan be written as a system of linear equations:

- Second order polynomial in one variable

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

- Second order polynomial in two variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

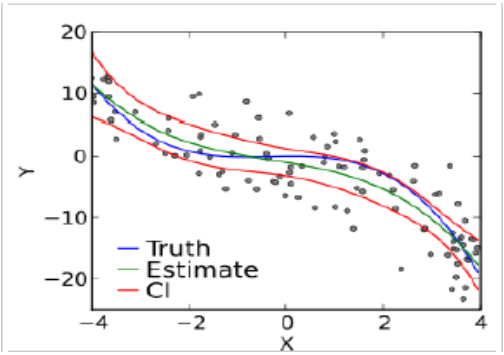


Figure 7.1 Polynomial Regression

CHAPTER 8
CORRELATION

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. For example, height and weight are related; taller people tend to be heavier than shorter people. The relationship isn't perfect. People of the same height vary in weight, and you can easily think of two people you know where the shorter one is heavier than the taller one. Nonetheless, the average weight of people 5'5" is less than the average weight of people 5'6", and their average weight is less than that of people 5'7", etc. Correlation can tell you just how much of the variation in peoples' weights is related to their heights.

Although this correlation is fairly obvious your data may contain unsuspected correlations. You may also suspect there are correlations, but don't know which are the strongest. An intelligent correlation analysis can lead to a greater understanding of your data.

8.2 Correlation Coefficient

The main result of a correlation is called the correlation coefficient (or "r"). It ranges from 1.0 To +1.0. The closer r is to +1 or 1, the more closely the two variables are related.

If r is close to 0, it means there is no relationship between the variables. If r is positive, it means that as one variable gets larger the other gets large r. If r is negative it means that as one gets larger, the other gets smaller (often called an "inverse" correlation).

While correlation coefficients are normally reported as r = (a value between 1 and +1), squaring them makes then easier to understand. The square of the coefficient (or r square) is equal to the percent of the variation in one variable that is related to the variation in the other. After squaring r, ignore the decimal point. An r of .5 means 25% of the variation is

related (.5 squared =.25). An r value of .7 means 49% of the variance is related (.7 squared = .49).

A key thing to remember when working with correlations is never to assume a correlation means that a change in one variable causes a change in another. Sales of personal computers and athletic shoes have both risen strongly in the last several years and there is a high correlation between them, but you cannot assume that buying computers causes people to buy athletic shoes (or vice versa).

8.3 Types of Correlation

All correlations have two properties: strength and direction. The strength of a correlation is determined by its numerical value. The direction of the correlation is determined by whether the correlation is positive or negative.

- Positive correlation: Both variables move in the same direction. In other words, as one variable increases, the other variable also increases. As one variable decreases, the other variable also decreases.I.e., years of education and yearly salary are positively correlated.
- Negative correlation: The variables move in opposite directions. As one variable increases, the other variable decreases. As one variable decreases, the other variable increases. I.e., hours spent sleeping and hours spent awake are negatively correlated.

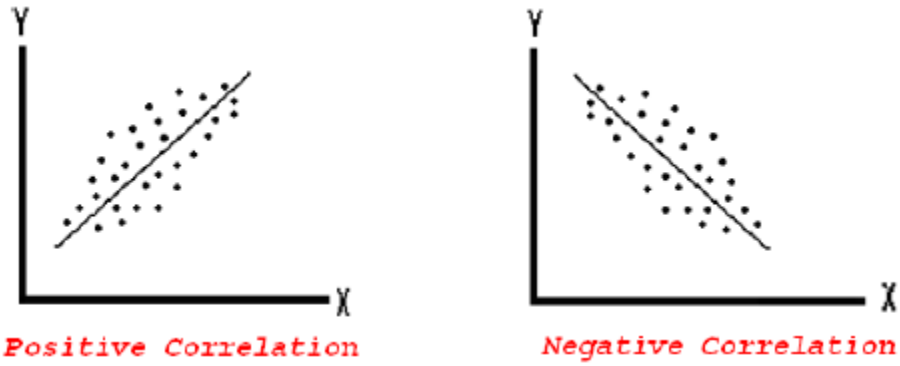


Figure 8.1 Positive and Negative Correlation

All positive correlations have scatterplots that move in the same direction as the positive correlation in the image above. All negative correlations have scatterplots that move in the same direction as the negative correlation in the image above.

8.4 No Correlations

What does it mean to say that two variables have no correlation? It means that there is no apparent relationship between the two variables. For example, there is no correlation between shoe size and salary. This means that high scores on shoe size are just as likely to occur with high scores on salary as they are with low scores on salary.

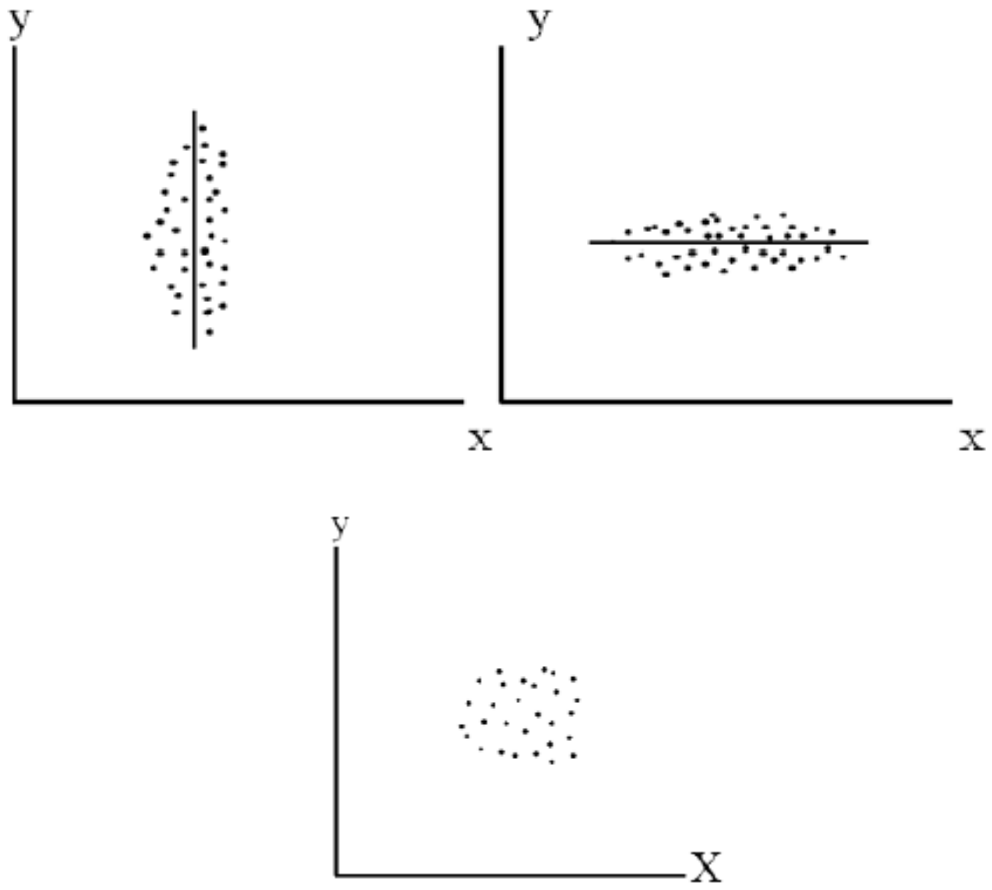


Figure 8.2 No Correlation

If your line of best fit is horizontal or vertical like the scatterplots on the top row, or if you are unable to drawa line of best fit because there is no pattern in the data points, then there is little or no correlation.

CHAPTER 9
HEAT MAP

A heat map (or heatmap) is a graphical representation of data where the individual values contained in a matrix are represented as colors. The term 'heat map' was originally coined and trademarked by software designer Cormac Kinney in 1991, to describe a 2D display depicting financial market information,[1] though similar plots such as shading matrices have existed for over a century.[2]

9.1 Types

There are many different color schemes that can be used to illustrate the heatmap, with perceptual advantages and disadvantages for each. Rainbow colormaps are often used, as humans can perceive more shades of color than they can of gray, and this would purportedly increase the amount of detail perceivable in the image. However, this is discouraged by many in the scientific community, for the following reasons:[6][7][8][9][10]

The colors lack the natural perceptual ordering found in grayscale or blackbody spectrum colormaps. Common colormaps (like the "jet" colormap used as the default in many visualization software packages) have uncontrolled changes in luminance that prevent meaningful conversion to grayscale for display or printing. This also distracts from the actual data, arbitrarily making yellow and cyan regions appear more prominent than the regions of the data that are actually most important.

CHAPTER 10

ARTIFICIAL NEURAL NETWORK

Artificial neural networks (ANNs) or connectionist systems are a computational model used in computer science and other research disciplines, which is based on a large collection of simple neural units (artificial neurons), loosely analogous to the observed behavior of a biological brain's axons. Each neural unit is connected with many others, and links can enhance or inhibit the activation state of adjoining neural units.

Each individual neural unit computes using summation function. There may be a threshold function or limiting function on each connection and on the unit itself, such that the signal must surpass the limit before propagating to other neurons. These systems are selflearning and trained, rather than explicitly programmed, and excel in areas where the solution or feature detection is difficult to express in a traditional computer program.

The goal of the neural network is to solve problems in the same way that the human brain would, although several neural networks are more abstract. Modern neural network projects typically work with a few thousand to a few million neural units and millions of connections, which is still several orders of magnitude less complex than the human brain and closer to the computing power of a worm.

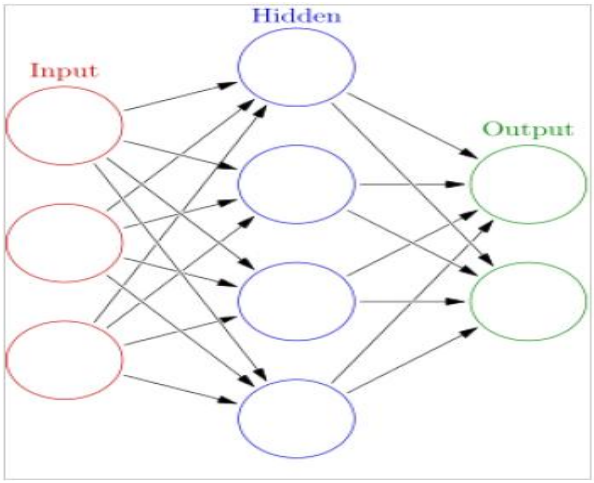


Figure 10.1 Artificial Neural Network

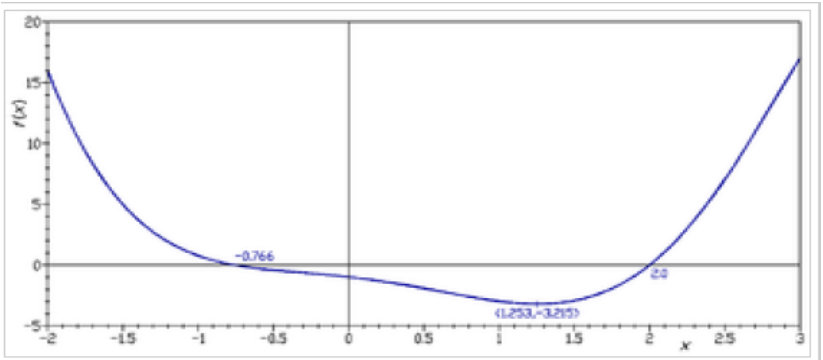
CHAPTER 11

FUNCTION PLOT

In mathematics, the graph of a function f is the collection of all ordered pairs $(x, f(x))$. If the function input x is a scalar, the graph is a twodimensional graph, and for a continuous function is a curve. If the function input x is an ordered pair (x_1, x_2) of real numbers, the graph is the collection of all ordered triples $(x_1, x_2, f(x_1, x_2))$, and for a continuous function is a surface.

Informally, if x is a real number and f is a real function, *graph* may mean the graphical representation of this collection, in the form of a line chart: a curve on a Cartesian plane, together with Cartesian axes, etc. Graphing on a Cartesian plane is sometimes referred to as *curve sketching*. The graph of a function on real numbers may be mapped directly to the graphic representation of the function. For general functions, a graphic representation cannot necessarily be found and the formal definition of the graph of a function suits the need of mathematical statements, e.g., the closed graph theorem in functional analysis.

The concept of the graph of a function is generalized to the graph of a relation. Note that although a function is always identified with its graph, they are not the same because it will happen that two functions with different codomain could have the same graph. For example, the cubic polynomial mentioned below is a surjection if its codomain is the real numbers but it is not if its codomain is the complex field.



Graph of the function $f(x) = x^4 - 4^x$ over the interval $[-2,+3]$. Also shown are its two real roots and global minimum over the same interval.

Figure 11.1 Function Plotting

CHAPTER 12

PARAMETRIC PLOT

In mathematics, parametric equations define a group of quantities as functions of one or more independent variables called parameters.[1] Parametric equations are commonly used to express the coordinates of the points that make up a geometric object such as a curve or surface, in which case the equations are collectively called a parametric representation or parameterization of the object.

In addition to curves and surfaces, parametric equations can describe manifolds and algebraic varieties of higher dimension, with the number of parameters being equal to the dimension of the manifold or variety, and the number of equations being equal to the dimension of the space in which the manifold or variety is considered (for curves the dimension is *one* and *one* parameter is used, for surfaces dimension *two* and *two* parameters, etc.).

Parametric equations are commonly used in kinematics, where the trajectory of an object is represented by equations depending on time as the parameter. Because of this

application, a single parameter is often labeled *t*; however, parameters can represent other physical quantities (such as geometric variables) or can be selected arbitrarily for convenience. Parameterizations are nonunique; more than one set of parametric equations can specify the same curve.

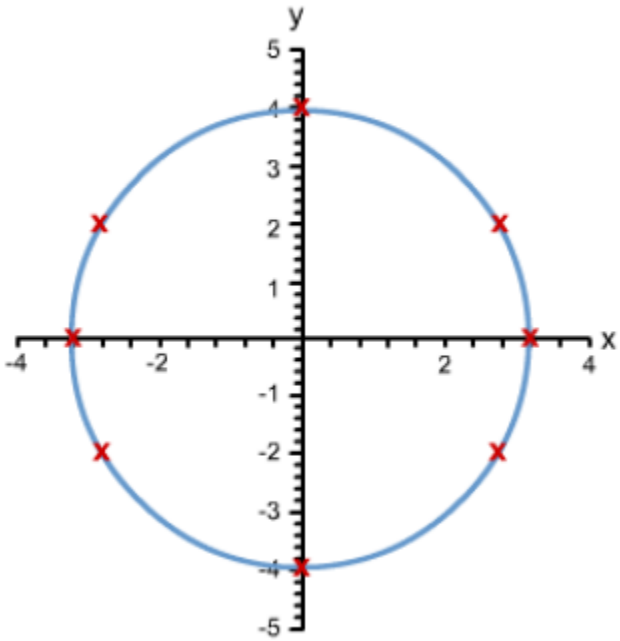


Figure 12.1 Parametric Plotting

CHAPTER 13
CONCLUSION

We have studied about the process of data analysis and how to make a data analysis software using various tools in python. In future we will try to deploy the software on cloud.

REFERENCES

[1] Practical Data Analysis - by Hector Cuesta (Author)
[2] Python for Data Analysis - by Wes Mckinney (Author)