Data Analytics (UE18CS312)

Unit 1

Aronya Baksy

August 2020

1 Introduction

Definition 1. Interpretation and exploration of past data is known as **Analysis** of Data

Definition 2. Data **Analytics** involves analysis (as defined above) along with predictive modelling of the future from the derived insights

Data Analytics is a set of statistical and operation research techniques, AI, IT and management strategies used to frame a problem, collect data and use the data to generate insghts that create value for organizations.

1.1 Types of Data Analytics

1. **Descriptive:** Answers the question "What is happening?". Using visualizations effectively, build them using comprehensive, accurate and live data.
2. **Diagnostic:** Answers the question "Why is it happening?". Identify the root cause and isolate confounding factors
3. **Predictive:** Answers the question "What will happen?". Predict the outcome of a certain action given the historical data.
4. **Prescriptive:** Answers the question "What do I need to do?". Recommend strategies and actions based on outcomes of prediction and testing various strategies.

1.2 Steps in Data Analytics

1. Data Collection
2. Data Preprocessing
3. EDA
4. Insights (ML/DL)
5. Visual Report

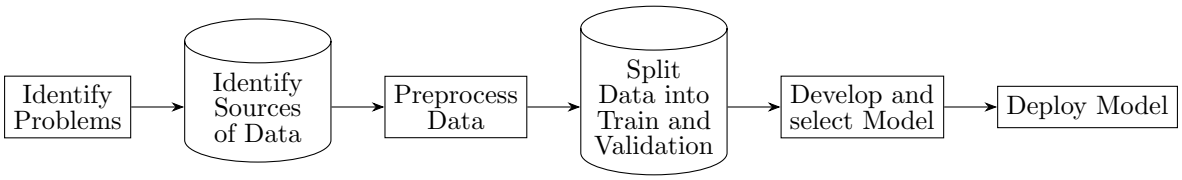

Figure 1: Data Analytics Pipeline

1.3 Costs in Business Decision Making

- **Decision Cost:** Cost of producing a decision, using the help of a decision maker or a fixed procedure.
- **Implementation Cost:** Cost of actions based on the decision taken
- **Failure Cost:** Cost of organization's inability to produce and implement the decision.

The aim of data analytics is to reduce these costs.

1.4 Business use cases for Data Analytics

- **Process Improvement:** Reduce cheque clearance time in banks, reduce patient discharge time in hospitals, reduce wastage in manufacturing, reduce time to deliver customer orders in e-com.
- **Problem Solving:** Reduce the impact of NPAs in financial investments, predict financial fraud, reduce inventory management cost in manufacturing, space allocation and route planning in e-com.
- **Decision Making:** Approve loan decisions, decide interest rates, introduce new products, mark-down pricing and promotions in e-com.

2 Data Sources and Representation

2.1 What is Data

- Collection of **objects** and their **attributes**.
- Attributes are the descriptive properties/characteristics of an object.
- A single object can also be referred to as a point, an entity, a record, an instance, a sample or a case.
- Attribute values are the actual values assigned to an attribute for a particular object. They could be alphanumeric, ordinal, symbols, etc.
- The same attribute can be mapped to different attribute values. The set of all unique values that an attribute can take is called the **domain** of that attribute.
- Different attributes may be mapped to the same attribute value, but with different meanings and properties. (eg: ID number and age are both integers but mean entirely different things).

2.2 Types of Data

- Based on **measurement level**:
  1. **Nominal:** Also known as categorical attributes, represent the names of different states or groups. There is no order or rank among these categories. (eg: ID numbers, names, ZIP codes, eye colours)
  2. **Ordinal:** The values have an inherent ranking but no absolute value of their own. (eg: Grades, height (tall/medium/short), taste of chips on a scale of 1-10)
  3. **Interval:** The distance between attribute values is important and carries meaning to the attribute, but their ratio does not mean anything. (eg: Calendar dates, temperatures in C/F)
  4. **Ratio:** The ratios between different values carries meaning, and there is an absolute zero (eg: length, counts, temperature in K, elapsed time)
- Based on **continuity**:
  1. **Discrete:** Finite/countably infinite set of values, can be represented as integers.
  2. **Continuous:** Real numbers as attribute values, represented as float point numbers.

2.3 Data Representations

2.3.1 Structured and Unstructured Data

- **Structured** data is data that is represented in rectangular, or matrix form with labelled rows and columns. (eg: Relational Databases, spreadsheets)
- **Unstructured** data has no rectangular or matrix representation (eg: text, photos, videos, audio, real-time instrument data, webpages)
- **Semi-Structured** data has no rectangular form, but follows some fixed format. (eg: XML files, NoSQL databases, e-mails)

2.3.2 Data Cubes

- Data cubes are a method of displaying 3 dimensional data efficiently. They are also called **OLAP** cubes.
- They support different analytical operations, called **OLAP** operations (OnLine Analytical Processing).

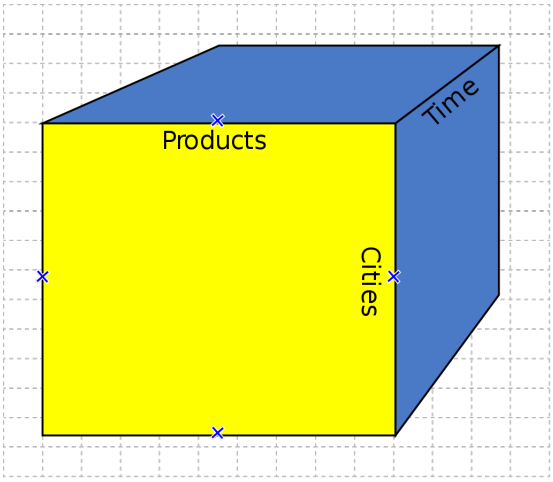

Figure 2: A data cube

2.4 OLAP Operations

1. **Roll-up:** Summarize data by climbing hierarchy or reducing dimension (eg: change cities to countries by grouping all the cities)
2. **Drill-down:** Increase level of granularity along one axis of the cube. (eg: change time in quarters (Q1, Q2 etc.) to time in months)
3. **Slice/Dice:** Projection and Selection (A **slice** represents the result of a single boolean condition on the cube, eg: `city == "Toronto"`, while a **dice** represents the result of a conjunction of 2 or more boolean expressions, eg: `(city == "Toronto") and (time=="Q1" or "Q2") and (product=="phone" or "laptop")`
4. **Pivot:** Change the axes of the cube by rotating them.

3 Data Exploration

- Data Exploration is used to give basic insights on the patterns present in the data, and its overall shape and distribution.

- Exploratory Data Analysis (EDA) consists of using **summary statistics** and **data visualization** to derive this information.
- The **population** is defined as the set of all possible observations in a given problem context for which conclusions are to be drawn.
- The **sample** is a subset of the population which is of interest to the observer, and based on whom conclusions about the entire population will be drawn.

3.1 Summarizing Data

3.1.1 Measures of Central Tendency

- **Mean**: Arithmetic average of all the data values.

$$\overline{x} = \frac{x_1 + x + _2 + x_3 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

  The symbol $\overline{x}$ is used to denote the sample mean, and the symbol μ is used to denote the population mean.

- **Mode**: The most frequently occurring data value. It can be applied to categorical data as well, unlike mean and median.
- **Median**: The value that divides the data into two equal parts, with half the data below and the other half above the median.

3.1.2 Measures of Spread

- **Quartiles**: The first (Q_1), second (Q_2) and third (Q_3) quartiles divide the data into 4 equal parts. The second quartile is the median.
  The kth quartile can be calculated using the value at the index i given by

$$i = k\frac{n+1}{4} \tag{2}$$

  The **IQR** (interquartile range) is defined as the part of the data where the middle 50% of the data lies. It is calculated as $Q_3 - Q_1$

- **Percentiles**: This is a further generalization of the quartile, where the data is broken into 100 equal parts. The quartiles are the 25th, 50th and 75th percentiles respectively.
  The kth percentile can be calculated as the data value at the index i given by

$$i = k\frac{n+1}{100} \tag{3}$$

- **Variance**: Measure of variability of the data points from the mean value. It is calculated by

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{4}$$

  The above formula denotes the population variance. The sample variance is calculated as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x})^2 \tag{5}$$

  The replacement of n with $n-1$ for the sample variance formula is called the Bessel's correction, and is used to correct for the bias that occurs when taking a sample from the population.

- **Standard Deviation**: The amount of dispersion or variance in a dataset is denoted by its standard deviation σ or s (population and sample values). It is calculated by taking the **square root** of the **variance**.

3.1.3 Degrees of Freedom

The number of independent variables in a model, or independent observations in a data set is called the degrees of freedom.
Given a sample mean $\overline{x}$ and $n-1$ values, the last value can have only one fixed value, and hence is dependent on the rest of the values. Hence here there are $n-1$ degrees of freedom.
If there are n items in the sample and k parameters estimated from the sample, then the number of degrees of freedom is $n-k$

3.1.4 Chebyshev's Inequality

The probability of finding a randomly selected value X from the data in the interval $\mu \pm k\sigma$ is given as:

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2} \tag{6}$$

3.1.5 Skewness and Kurtosis

- **Skewness** is a measure of the symmetry of the distribution. For data to be symmetrical, there must be an equal proportion of data in the intervals $(\mu - k\sigma, \mu)$ and $(\mu, \mu + k\sigma)$, where k is a positive integer.
- The skewness is the third moment of the centered distribution of X about the mean μ.

$$g_1 = \mathbb{E}\left[\left(\frac{X - \overline{X}}{\sigma}\right)^3\right] = \frac{1}{\sigma^3}\frac{\sum_{i=1}^{n}(X_i - \overline{X})^3}{n} \tag{7}$$

- For $g_1 > 0$, the tail of the data is longer on the right (+ve dir), for $g_1 < 0$ the tail of the data is longer on the left (-ve dir), while for $g_1 = 0$ both tails are equal.

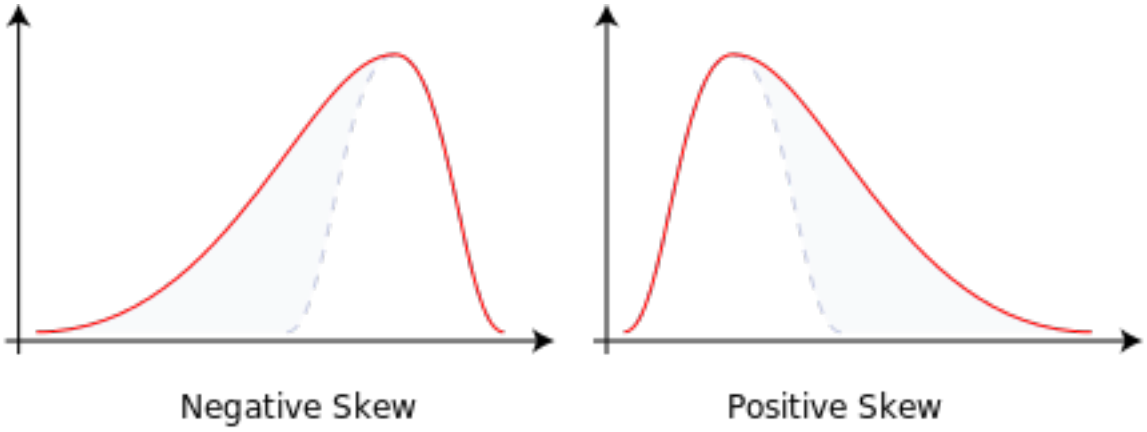

Figure 3: Positively and Negatively Skewed Distributions

- **Kurtosis** is a measure of how heavy the tails of a distribution are. It is given by

$$kurtosis = \mathbb{E}\left[\left(\frac{X - \overline{X}}{\sigma}\right)^4\right] = \frac{1}{\sigma^4}\frac{\sum_{i=1}^{n}(X_i - \overline{X})^4}{n} \tag{8}$$

- A kurtosis value of < 3 indicates a platykurtic disttribution, > 3 indicates a leptokurtic distribution while a mesokurtic distribution (eg: the standard normal) has a kurtosis value of 3.
- The deviation of a distribution from the standard normal is reflected in the *excess kurtosis*, given by $kurtosis - 3$.

4 Sampling techniques

4.1 Probabilistic Sampling

- **Systematic Sampling**: Picking every kth sample, where k is called the sampling rate. (eg: time series analysis, sampling video frames, pixels in an image)
- **Simple Random Sampling**: Can be done with replacement (for finite population) or without replacement (infinite or very large population).
- **Stratified Sampling**: Divide the data into homogenous sub-groups (called strata) and perform simple random sampling on each sub group. Division into strata is based on some characteristic (eg: income level, address, etc.)
- **Cluster Sampling**: Divide the population into clusters, and sample these clusters. Clusters must be internally as heterogenous as possible but two clusters must be similar in their structure. (eg: geographical surveys)

4.2 Non-Probabilistic Sampling

- **Convenience Sampling**: Pick samples that are convenient to the one running the experiment
- **Judgement Sampling**: Pick best samples as per the **judgement** of the person running the experiment.

5 Data Visualizations

5.1 Histogram

- Used to visualize the frequency distribution of continuous data.
- It is the frequency distribution of the data arranged in consecutive non-overlapping intervals.
- The area of each rectangle indicates its frequency. If all the bins have same width then height is proportional to frequency.
- Histograms can effectively show both skew and kurtosis.
- Choosing the appropriate bin size is important. Too small a bin size leads to too much noise, while too large a bin size leads to details being overlooked.
- The histogram of cumulative probabilites represents the **ogive** curve of that variable.

5.2 Bar Chart

- Frequency distribution for categorical/qualitative variables.
- The most and least frequently occurring classes can be identified from a bar chart.

5.3 Strip charts/Scatter Plots

- They indicate the correlation between the two variables under consideration (positive, negative or none at all).
- Using different symbols/colours for each class, they can be used to depict the relative frequencies of classes occurring in the data.
- The correlation is measured in terms of **Pearson's Coefficient of Correlation**, given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{9}$$

- $r > 0$ indicates positive correlation, $r < 0$ indicates negative correlation and $r = 0$ indicates no correlation. $r = \pm 1$ indicates perfect positive or negative correlation.

5.4 Box and Whisker Plot

- It depicts the **outliers** in the data.
- It is constructed using the values of the quartiles of the data.
- Observations greater than $Q_3 + 1.5IQR$ and less than $Q_1 - 1.5IQR$ are considered to be outliers and represented as circles or outliers.
- The box represents the IQR which contains 50% of the data.

5.5 Pie Chart

- For categorical data, they depict the frequency of the class as a percentage of the whole.
- These percentages are scaled appropriately and depicted as sectors on a circle with the enclosed angle proportional to the relative frequency.

5.6 Coxcomb Plot

- Also called the polar extension chart or the rose chart, popularized by Florence Nightingale.
- Each sector has the same enclosed angle, but different radius. The area is proportional to the magnitude of the frequency of that class.

5.7 Bivariate Gaussian Distributions

- A vector $\mathbf{x} = (x_1, x_2, x_3, .., x_k)^T$ follows the Gaussian distribution, whose PDF is given as

$$f_X(x_1, x_2, .., x_n) = \frac{1}{\sqrt{(2\pi)^k |\mathbf{\Sigma}|}} exp(\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)) \tag{10}$$

- The matrix $\mathbf{\Sigma}$ is called the **covariance matrix**, and it is defined as

$$\Sigma = \mathbb{E}((\mathbf{x} - \mu)(\mathbf{x} - \mu)^T) \tag{11}$$

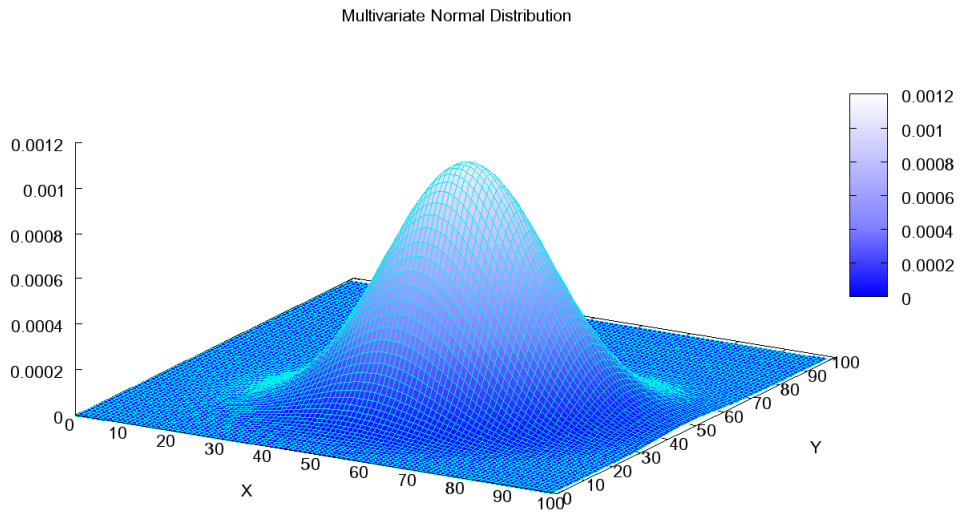

Figure 4: Joint Probability Distribution of 2 Gaussian variables

- Bimodality tests:

1. **Necessary Condition**: $kurtosis - skewness^2 \leq 1$. The equality holds in the extreme case where the distribution is just two straight lines at 2 points and flat everywhere else (double Dirac).
2. **Hartigan's dip test**: $p < 0.05$ implies significant multimodality, $0.05 < p < 0.1$ implies multimodality with marginal significance.

6 Data Cleaning

6.1 Measures of Data Quality

- **Accuracy**: Matches with real-world observations
- **Completeness**: No missing values
- **Consistency**: Appropriate formats, correct spellings, etc.
- **Timeliness**: Update data as time passes
- **Believability**: Trust in data source
- **Interpretability**: Ease of understanding data

6.2 Types of Unclean data

- **Incomplete**:
  - Lacking in attribute values, lacking values of interest or containing only aggregate data
  - Can be handled by **ignoring** the tuple (not feasible if the % of missing values per attr varies), **filling manually** (tedious), filling it automatically with some placeholder value, or (for numerical data) filling with the mean (classwise or global).
- **Noisy**: Containing noise, outliers or errors
  - Caused by instrument error, data entry/transmission problems, tech. limitations
  - Noise can be filtered out using filters that use wavelet/Fourier transforms
- **Inconsistent**: Non-matching data formats, discrepancies in codes or names
- **Intentional**: Filled in due to lack of data, disguised missing data, placeholder values

6.3 Types of Missing Data

- **MCAR** (Missing Completely At Random):
  - The fact that a value is missing is independent of its attribute.
  - Causes unbiased estimates
  - Caused by instrument errors
  - Fill in values based on the attribute (eg: mean), is not realistic
- **MAR** (Missing At Random):
  - Missingness is related to other variables
  - Produces bias in the analysis almost always
  - Fill in values based on other values
- **MNAR** (Missing Not At Random):
  - Missingness is related to unobserved measurements
  - Informative or non-ignorable, need to find more data or find the cause of missing data.
  - eg: Censored data, instruments start to wear out

6.3.1 Solutions to Missing Data

- **MCAR**:
  - Delete rows/columns (if they comprise an appropriately small fraction of the data)
  - Mean imputation (only for MCAR)
  - Pairwise deletion: compute mean, variance and covariance and only retain pairs with high covariance
- **MAR**:
  - Regression Imputation, Stochastic regression Imputation
  - LOCF, BOCF, WOCF (Last/Baseline/Worst Observation Carried Forward)
- **MNAR**; Model the missing values separately

6.4 Noisy Data

- Noise can be defined as random error or variance in a statistical variable.
- Boxplots and scatterplots can be used to identify outliers, but numerical techniques are required to smooth out the noise.

6.4.1 Smoothing Techniques

- **Binning**: Smoothing by dividing data into bins. This can be done by **equal-width** binning (split data into bins of equal width), binning by **bin boundaries** (for each value in a bin, replace it with the boundary value that it is closer to) and smoothing by **bin means** or **bin medians** (replace each value with the bin mean/median).
- **Regression** and **Outlier Analysis** using clustering

7 Data Integration

- Merging of data from multiple sources into a single, coherent data store is known as data integration.
- This must be done while avoiding repeated (ie. redundant) data and inconsistent data. Schemas of all data sources must also match.
- **Entity Identification**: matching different names to the same entity across different datasets (eg: MS Dhoni and Mahendra Singh Dhoni are the same person)
- **Data value conflicts**: caused by different systems of units (Imperial vs metric), different scales, different representations
- **Derived data**: Data in one set may have been derived from attributes in the other set (eg: age from DOB), these can be removed using correlation and covariance analysis.

7.1 Correlation Analysis for Nominal Data: The χ^2 test

- Given the observed dataset containing values o_{ij}, the expected distribution can be calculated as:

$$e_{ij} = \frac{rowsum_i \times colsum_j}{N} \tag{12}$$

- If the dataset has m rows and n columns, the test statistic χ^2 is calculated as:

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{13}$$

  And the number of degrees of freedoms is calculated

$$f = (m-1)(n-1) \tag{14}$$

- Larger values of χ^2 imply a larger extent of correlation
- **Important: Correlation does not imply causation**. There may be a third factor that links two variables, causing them to be correlated.

7.2 Correlation Analysis for Numerical Data

- **Pearson's Correlation Coefficient** is given as

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \overline{A})(b_i - \overline{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\overline{AB}}{(n-1)\sigma_A\sigma_B} \tag{15}$$

- The **covariance** of 2 variables A and B is given by

$$Cov(A,B) = \mathbb{E}((A - \overline{A})(B - \overline{B})) = \frac{\sum_{i=1}^{n}(a_i - \overline{A})(b_i - \overline{B})}{n} \tag{16}$$

- The correlation and covariance are related as

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A\sigma_B} \tag{17}$$

  - $Cov(A,B) > 0$: Both A and B are larger than their expected values
  - $Cov(A,B) < 0$: A is larger than its expected value, B is likely to be smaller than its expected value
  - $Cov(A,B) = 0$: For 2 independent variables, the covariance is 0 but the converse is not true. (ie. not all pairs of variables with Cov 0 will be independent).

7.2.1 Hypothesis test for Pearson's Correlation Coefficient

- If ρ is the population correlation coefficient, then the **null Hypothesis** H_0 is that $\rho = 0$, ie. there is no correlation while the **alternate hypothesis** H_1 is that $\rho \neq 0$, ie there is a positive/negative correlation.
- The test statistic is taken from the Student's-T distribution with $n-2$ degrees of freedom, where n is the number of points in the sample.
- The test statistic is given by

$$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}} \tag{18}$$

7.3 Correlation for ordinal attributes

- The **Spearman's rank coefficient** is used to determine the correlation between ordinal attributes
- The coefficient r_s is given as

$$r_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)} \tag{19}$$

  Where D_i is the difference between the ranks of the variables in the i_{th} case.

7.4 Correlation between continuous and binary attributes

- The **point-biserial correlation** coefficient is used when one attribute is continuous and the other one is binary. It is given by

$$r_b = \frac{\overline{X_1} - \overline{X_0}}{S_X}\sqrt{\frac{n_0 n_1}{n(n-1)}} \tag{20}$$

  Where $\overline{X_i}$ is the mean of all X when Y = i, and S_X is the sample standard deviation of X. n_i is the number of points where Y = i

7.5 Correlation between binary attributes

- The ϕ coefficient is used when both attributes (say X and Y) are binary, given as

$$\phi = \frac{N_{11}N_{00} - N_{01}N_{10}}{\sqrt{N_{X0}N_{X1}N_{Y0}N_{Y1}}} \tag{21}$$

  Here N_{ij} represents the number of points where $X = i$ and $Y = j$

8 Data Reduction

- Obtaining a representation of the data that is smaller in volume but retains maximum information and produces the same analytical results.
- This is done to reduce the time needed for data analysis over large amounts of data.
- Approaches to data reduction:
  - **Dimensionality Reduction**: PCA, Wavelet Transforms, feature subset selection, feature creation
  - **Numerosity Reduction**: Regression models, log-linear models, histograms, clustering, sampling, data cube aggregation
  - **Data Compression**: For continuous data like images, audio, video
- **Curse of Dimensionality**: As the dimension of data increases, data becomes increasingly sparse. The density and distance between points become less meaningful, hence clustering and regression also become less useful.

8.1 Wavelet Transform

- Decomposes a signal into frequency sub-bands.
- The wavelet transform operates such that the relative distance between objects is preserved at different resolutions. Hence natural clusters become more visible.
- Applied in image compression (JPEG2000 standard)

8.1.1 Method

- The input sample X, of length L where L is a power of 2 (can be zero padded if needed)
- Each transform has two functions, smoothing (represented as g) and difference (represented as h). They are defined as

$$g(i) = \frac{1}{2}(X[i] + X[i+1]) \tag{22}$$

  and

$$h(i) = |X[i] - X[i+1]| \tag{23}$$

- The new vectors g and h are then downsampled by a factor of 2, ie. every alternate (2nd, 4th, etc.) sample is dropped
- Wavelet transforms are **insensitive to noise and input order**, **independent of scale and resolution** and are highly **efficient** (O(n) operation).

8.2 Principal Component Analysis

This involves the projection of the high-dimensional data into a k- dimensional space whose basis consists of the first k Eigenvectors of the covariance matrix of this data (where $k < n$ is the new, reduced dimension of the data).

8.2.1 Method

- Normalize the input data, to bring all columns within the same range
- The principal component vectors are computed from the Singular Value Decomposition of the Covariance matrix of the input data.
- Sort the columns of the normalized input data in decreasing order of the eigenvalues. The eigenvalues represent the portion of variance that is retained in that attribute.
- Select the first k columns as the reduced dataset.

8.3 Attribute Subset Selection

- Eliminate **redundant attributes**, such as duplicate or derived data.
- Eliminate **irrelevant attributes**, which do not contribute to the analysis at hand.
- **Heuristic Search** can take many different forms:
  - Best single attribute selected based on a statistical test
  - Step-wise Forward Selection: pick the best single attribute first, and then conditional on the first attribute chosen pick the next
  - Step-wise Backward Elimination: Eliminate the worst attribute first and repeat until the attribute set is minimized.
  - Combined step wise selection and elimination of best/worst attribute
  - Decision Tree reduction using algorithms like ID3, C4.5 and CART. All attributes that do not appear in the tree are eliminated.
-

8.4 Attribute Creation

- Create a new attribute set that more effectively captures the relevant information for the analysis.
- Can be domain-specific attribute extraction
- Can also be done by mapping the data to a new coordinate system (eg: Fourier transform, wavelet transform)
- Attribute construction: combining features, data discretization

8.5 Numerosity reduction

8.5.1 Regression methods

- **Linear** regression involves fitting a straight line to the data.
- **Multiple** regression models the output as a linear model of an input feature vector (not a single input variable).
- **Log-linear** models approximate multidimensional probability distributions (eg: the Poisson distribution)

8.5.2 Histograms and Clustering Models

- Histograms: partition data into bins, and store the average of bins (bins may be of equal width or equal frequency)
- Clustering: cluster similar points and store only cluster representations (ie: diameter and centroid), can be hierarchical

9 Data Transformation

A transformation is a one-to-one map from the current value set to a new value set such that each old value can be associated with a unique new value, which is easier and more convenient to analyze.

9.1 Types of transformations

- **Attribute Construction**: Create new attributes from existing set
- **Aggregation**: Summarization, data cube construction
- **Smoothing**: noise reduction, averaging, Gaussian Smoothing
- **Normalization**: Scaling values to within a certain range
- **Discretization**: Grouping data into bins for category wise analysis
- **Concept Hierarchy Generation**: For nominal data, create a tree-like structure that defines the more specific concepts as children of a larger and more general concept.

9.1.1 Normalization

- **Min-max Normalization**: From the range $[min_A, max_A]$ to $[new\_min_A, new\_max_A]$

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \tag{24}$$

- **z-score Normalization**:

$$v' = \frac{v - \mu_A}{\sigma_A} \tag{25}$$

- **Decimal Normalization**:

$$v' = \frac{v}{10^j} \tag{26}$$

  Where $j \in \mathbb{Z}$ is the smallest integer such that $max(|v'|) < 1$

9.1.2 Discretization

- **Binning**: Top down splitting of data, unsupervised
- **Clustering**: Top-down splitting or bottom-up merging, unsupervised
- **Decision Tree Analysis**: Top-down splitting, supervised
- **Correlation Analysis**: Unsupervised, bottom-up merging