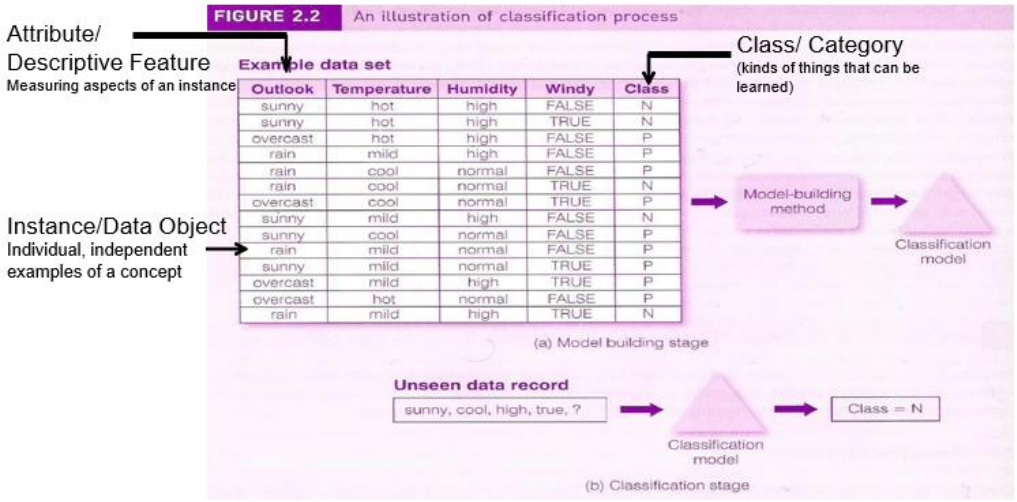
Data Mining

Mohammed Almulla

College of I.T

Data, Preprocessing, and Exploration

In this section, I will talk about data and its types, operations, origins, and quality, also preprocessing data techniques, and finally explore and visualize data. First of all, a data object or instance is an individual independent recording of a real-life object or event. It is characterized by its recorded values on a fixed set of features or attributes. And a feature or attribute is a specific property of the data object. A measurement is assigning a valid value to an attribute according to an appropriate measurement scale. A collection is collecting measurement results or recorded values. Figure 20 shows illustration of classification process with an example of data or object.



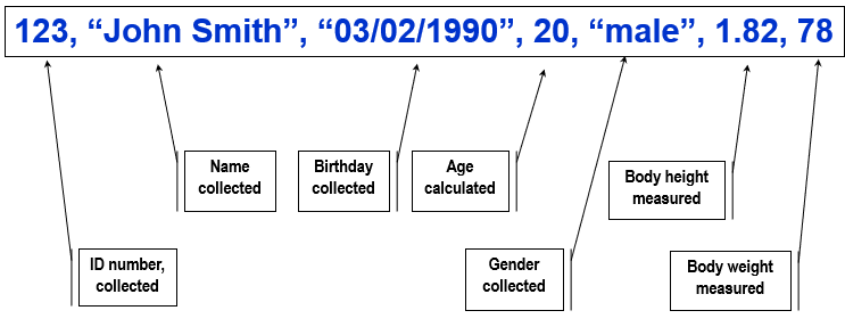


Figure 20: illustration of classification process and an example of data or object

Then, the measurement and measurement errors for data object are precision that is the closeness of measurements to one another, and it represented by the standard deviation, Bias is a systematic variation of measurements from the intended quantity measurement, and it was only known when external reference available, and Accuracy that is the closeness of the measure to the true value. After

that, we must collect errors that are incorrect data recording at the point of entry such as writing the name of Mohammed as “Muhadamed” or Hongbo Du as “Hongpo Do”.

Moreover, the attribute domain types and operations are classified into two categories where the domain is a set of permitted values of the attribute. The first category is qualitative types, and it consists of nominal that is a set of names without order such as gender is male (M) or female (F), and the other is ordinal that is a set of names with the concept of order such as grade can be evaluated like A, B, C, D, or F. The second category is quantitative types, and it consists of an interval that is a set of numeric values both order and difference exist such as a temperature or calendar year, and the other is the ratio that is a set of numeric values order, difference, and ratio are exist such as length in meter or feet. Figure 21 shows a table for attribute domain types.

	Type	Meaning	Operations	Example (+)	Not Applicable	
qualitative	Nominal	Distinct names	= and !=	Country name: UK = UK UK != China	China > UK ?	categorical
	Ordinal	Names with order	=, !=, >, <,>=,<=	Description of day: cold < warm<hot	cold – warm ?	
quantitative	Interval	Numeric values measured in fixed equals units	Ordinal operations, +, -	Year of birth : 1978 < 1980 Difference=2yrs	1978 * 2 means ½ older Year 0?	Numeric
	Ratio	Internal with absolute zero point	Interval operations, *, /, ratio	distance Temperature in Kelvin		

Figure 21: Attribute domain types

Data sets could various forms. It could be a table of records such as a relational table, join of relational tables, numerical spreadsheet or data matrix, and Boolean strings or document-term matrix. Also, it

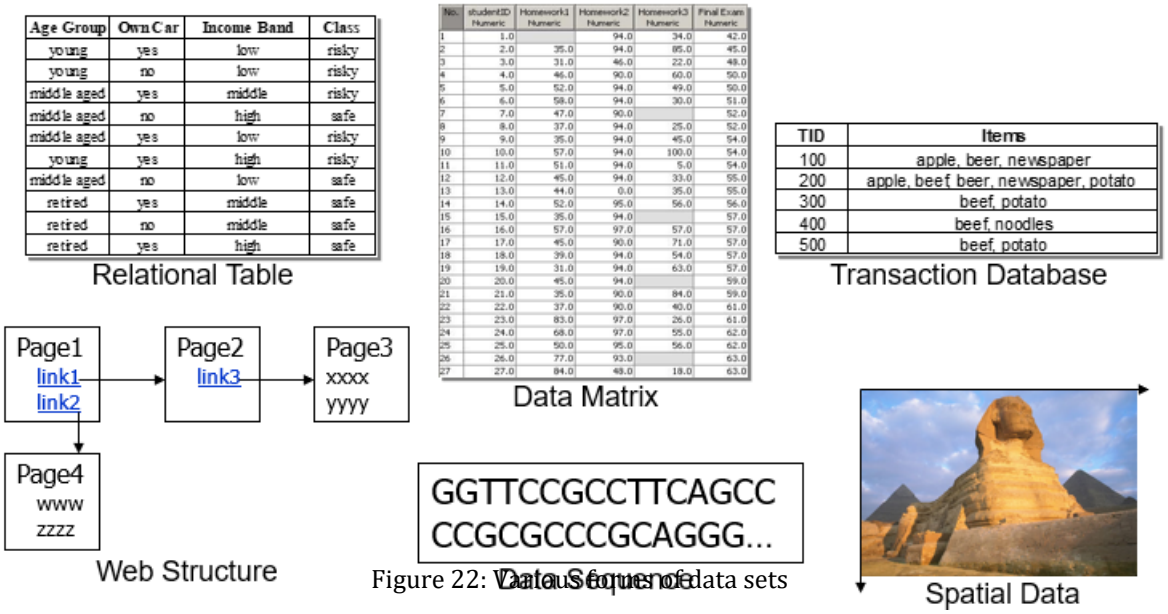


Figure 22: Various forms of data sets

could be ordered data such as time series and temporal sequence, data sequence, and spatial data. It could be graph-based data and non-record-based data. Figure 22 shows various forms of data sets.

Data sets have common properties. Firstly, the type of the data sets indicated by the file structure e.g., ARFF for Weka, DAT for See5, CSV. Second, the size of data sets is measured in terms of the total number of records or the total number of bytes e.g., small (MB), medium (GB), and large (TB). Thirdly, the dimensionality of the data sets is the number of attributes. Fourthly, sparsity that values are skewed to some extreme or sub-ranges and not normally or uniformly distributed. Finally, the resolution that is the right level of data details and is related to the intended purpose. Figure 23 shows the properties of data sets.

Before knowing how to preprocess the data we must know from where the data comes. The data sources could be local data source available, local operational systems from different departments and called internal source, Third-party external data source, and Enterprise or Organisational data warehouse that is an organizational database for decision making. It is a central data repository separate from operational systems that enforcing organization-wide data consistency and integration. It providing data details as well as data summarization and providing data values as well as metadata. It is equipped with data analysis and reporting tools as a data source for data mining.

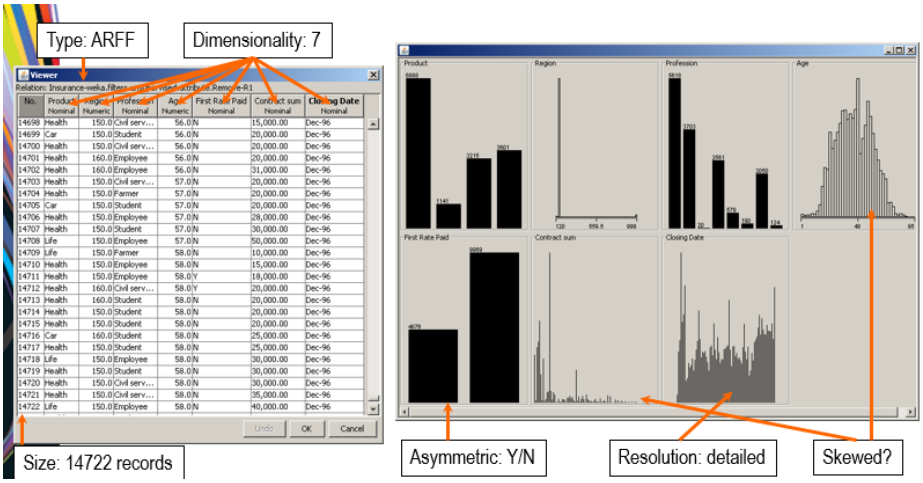


Figure 23: Properties of data sets

If new data appears outside the cluster what will happen? And how can we handle this? This is called data sparsity and it is a term used for how much data we have for a particular dimension or entity of the model. In scientific the term sparse is called on a matrix most of its elements are zero. Figure 24 shows data sparsity.

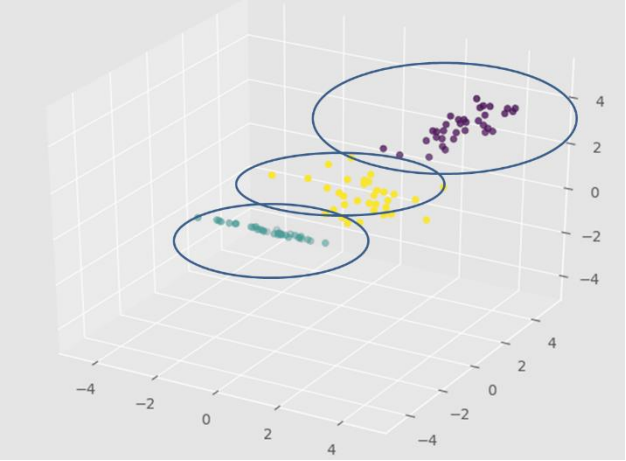


Figure 24: Data Sparsity

In figure 25, shows some basic statistics laws for the mean, median, mode, standard deviation, variance, and skewness. The skewness is a measure of symmetry or more precisely the lack of symmetry. Figure 26 shows skewness

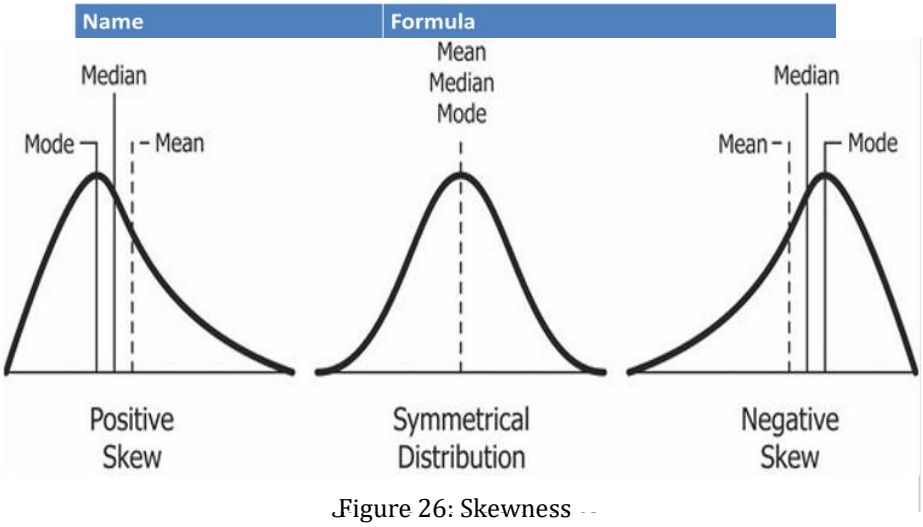


Figure 26: Skewness

By using basic statistics, we will measure errors in the models. So, to measure errors we must calculate three things. Firstly, Bias is a measurement error that depicts the variation of the recorded value from the real value. In other words, it is the sample mean subtracted from the true value with the absolute (abs) of the new value.

$$|Sample\ mean - True\ Value|$$

Secondly, the other measurement error is precision that is how precise is the measurement. And it is the standard deviation of repeated measurement, and less value is better. Finally, Accuracy is data recorded with high precision and low bias. Figure 27 shows accuracy, precision, and bias measurements with a circle on the best measurement.

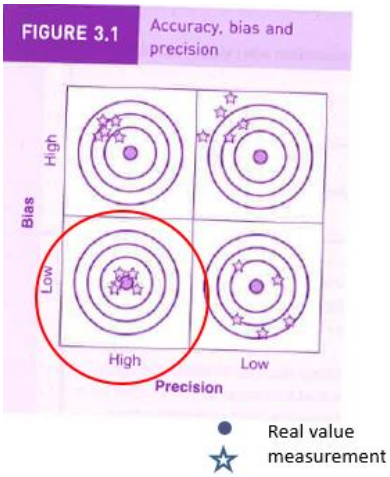


Figure 27: Accuracy, precision, and bias

In the real world, the data can be inaccurate, incomplete, and inconsistent. To make sure this does not happen to our data we must make sure of data quality is a measure of the condition of data based on factors. The factors that data quality can be judged by are accuracy that is data recorded with sufficient precision and little bias, correctness that is data recorded without errors and spurious objects, completeness that is whether if any parts of data records is missing, the consistency that is compliance with established rules and constraints, and redundancy that is unnecessary duplicates. Data quality is important because of garbage in garbage out (GIGO) law that means poor quality input could produce faulty output.

Data Pre-processing

At this point, we will start data pre-processing. The purpose for it is for speedy, cost-effective, and high-quality outcomes of data mining. The tasks for pre-processing data are as follows:

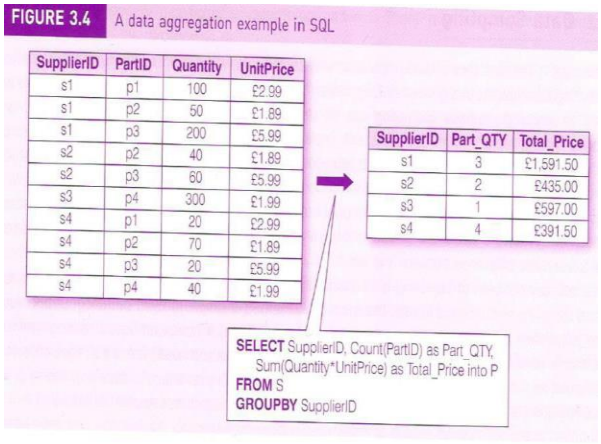
- Data aggregation
- Data sampling
- Dimension reduction
- Feature selection
- Feature creation

- Discretisation/binarization
- Variable transformation
- Dealing with missing values

I will cover each one of them. And after you understand them correctly then you can pre-process data. I advise to practice these steps on a raw data for better understanding.

Data Aggregation

Data aggregation is to summarise low level data details to higher level data abstraction. So to reduce the time of mining, to rescale data values, and to discover more stable patterns. By generalisation using a given concept hierarchy, by applying aggregate functions (e.g., count, sum, average), and dropping some attributes. Figure 28 shows the concept of data aggregation.



TID	Date	Item	Store	Price	Clubcard#	.....
.....	.....	.....	.....	.....	.....	.....
32144	06/06/2006	milk	Buckingham	1.99	1111	.....
11122	04/04/2006	watch	Buckingham	25.99	1011	.....
11122	04/04/2006	battery	Buckingham	3.99	1011	.....
11123	04/04/2006	beer	Buckingham	9.99	1022	.....
22244	04/04/2006	beer	MK	6.99	1022	.....
22244	04/04/2006	nappies	MK	10.89	1022	.....
23311	05/04/2006	beer	MK	6.99	1011	.....
.....	.....	.....	.....	.....	.....	.....

Date	Store	AveragePrice	.....
.....	.....	.....	.....
06/06/2006	Buckingham	1.99	.....
04/04/2006	Buckingham	13.32	.....
04/04/2006	MK	8.94	.....
05/04/2006	MK	6.99	.....
.....	.....	.....	.....

Number of Items	TotalPrice	Clubcard#	.....
.....	.....	.....	.....
1	1.99	1111	.....
3	36.97	1011	.....
2	27.87	1022	.....
.....	.....	.....	.....

FIGURE 3.5 Attribute-oriented induction for data aggregation: (a) input hierarchies, (b) input table and (c) output table

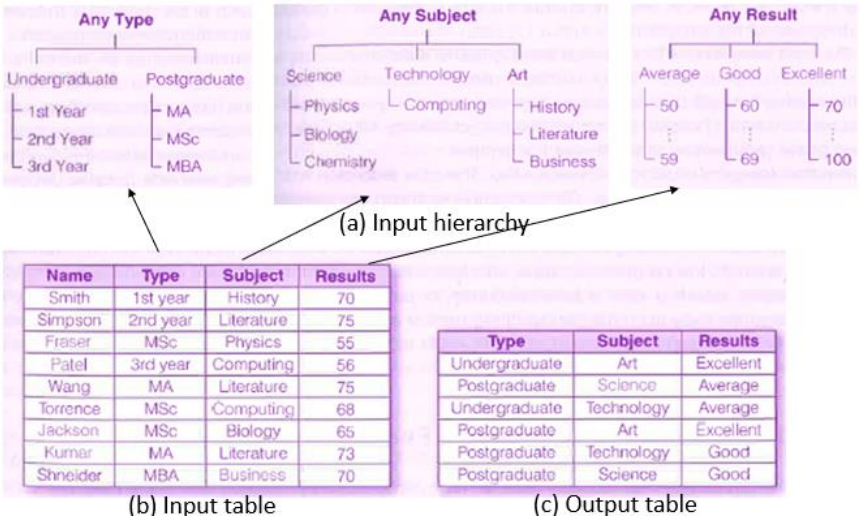


Figure 28: Data Aggregation Concept

Data Sampling

It is selecting a subset of the given data set. We use it to make it possible to use sophisticated mining algorithms within a time limit. And caution the sample must be representative of the original data set. We can do sampling by using random sampling, stratified sampling, and progressive sampling, With or without replacement. The random sampling is effective if all data objects can be listed, and the bad news is that the population is biased. Systematic sampling is to take samples on a fixed intervals and the size of sample and data set must be known, and even systematic sampling is biased. Stratified sampling is to group or cluster then sample from each group, and it avoid missing a group of data objects, and the issue

for stratified sampling is the criteria of grouping, size of data must be known, and different size groups. Progressive sampling start with small sample, then based on evaluation criteria increase the sample size gradually in stages. Figure 29 shows data sampling.

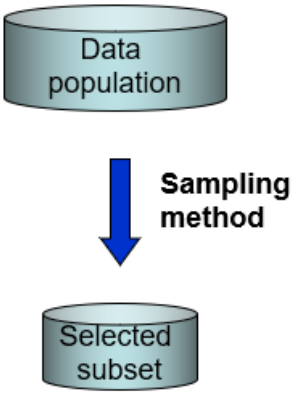


Figure 29: Data Sampling

Before talking about feature selection, I will talk about techniques to handle small data sets. And to handle them we could use simple models or combinations of learning algorithm, e.g., Bagging and boosting. Also, we must remove outliers because small data set is easily affected by noise or outliers. Even we can use techniques to extend the data sets like Synthetic samples or SMOTE that over sample minority class by duplicating or synthesizing instances belonging to the minority class. SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line. Also, there are another technique to extend the data set called input smearing that estimate new data from the data set by make small random changes to the original instances. Figure 30 shows SMOTE technique.

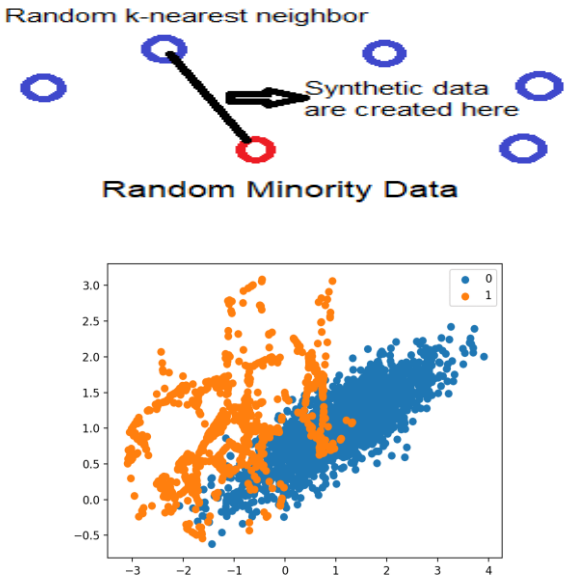


Figure 30: SMOTE technique

Also, another technique that is good in handling small data sets is bagging or also known as bootstrap aggregating. It combines the predicted classifications (prediction) from multiple models, or from the same type of model for different learning data. It is also used to address the inherent instability of results when applying complex models to relatively small data sets. Figure 31 shows bagging technique.

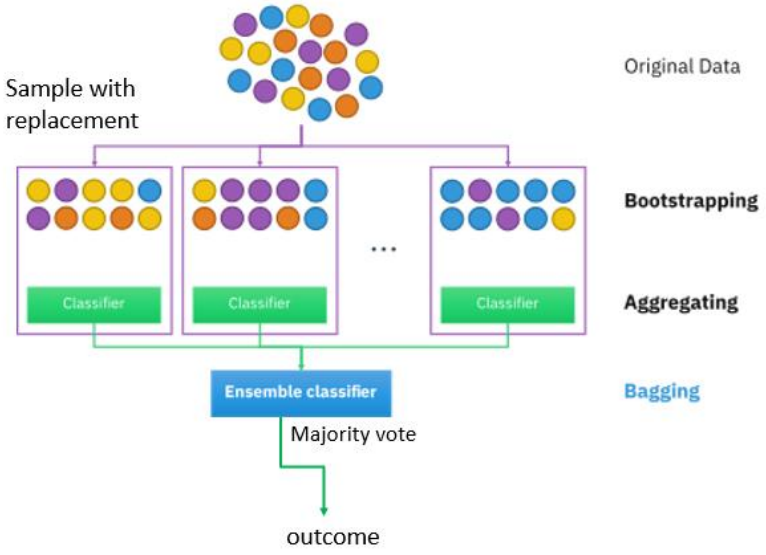


Figure 31: Bagging technique

Furthermore, to handle also noisy data we will use for it interquartile range. And a quartile is a type of quantile which divides the number of data points into four quarters of equal size and be sure the data is ordered.

$$Interquartile\ Range\ (IQR) = Q3 - Q1$$

Figure 32 shows interquartile range example.

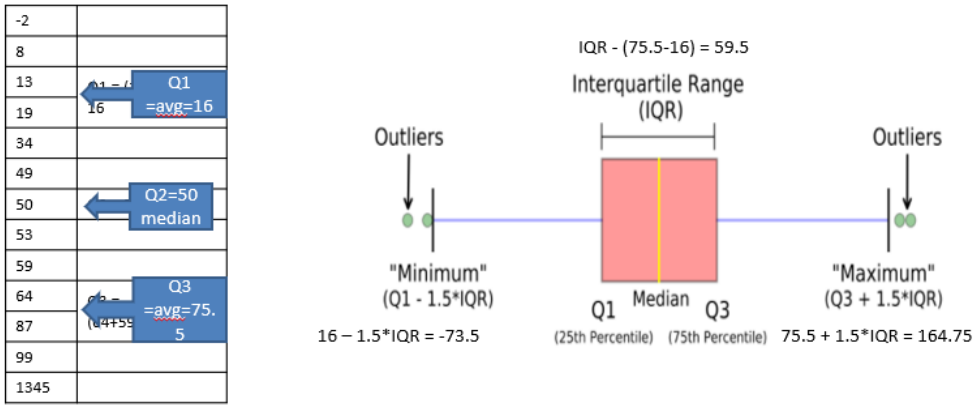


Figure 32: Interquartile range example

Feature Selection

It is reducing dimensionality by selecting a subset of attributes. The purpose for feature selection is to remove or reduce redundant features, and to remove irrelevant features with no useful information for the mining task. We will do feature selection manually with common sense and domain knowledge, by

remove missing values, by remove Low Variance attributes because if the variance is low or close to zero, then a feature is approximately constant and will not improve the performance of the model. Also, to do feature selection manually by reduce Highly Correlated features such as year of birth and age, by remove Low correlated feature with target attribute by using Pearson Correlation that is a measure of the strength of the linear relationship between two interval or numeric variables, and Chi-square test that is a measure of relationship. Also, on how to do feature selection is by letting the mining solution to select suitable features, and by filter and wrapper approaches. Figure 33 shows feature selection diagram.

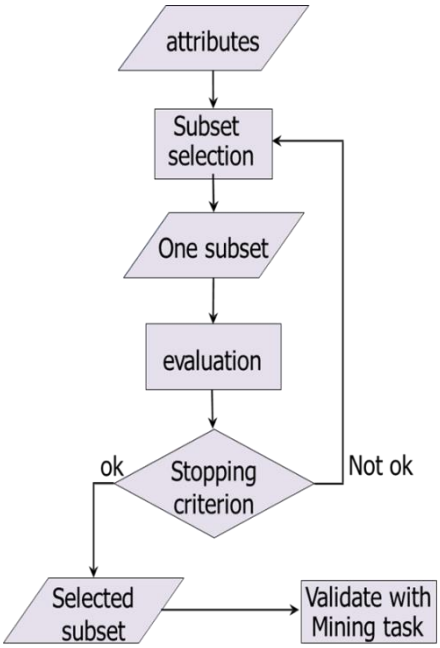


Figure 33: Feature Selection

Let us talk more about covariance and correlation. Covariance measures the linear relationship between two variables or attributes. While correlation can be positive if one variable increases the other increases too, negative if one variable increases the other decreases, and zero no correlation. To calculate correlation between two numeric attributes we will use Pearson coefficient.

$$Covariance(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$Correlation = \frac{Covariance(x, y)}{\sigma x \times \sigma y}$$

To calculate correlation between two nominal attributes we will use Chi-squared.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

And, to calculate correlation between mixed attributes we will use Point Biserial or Logistic Regression.

Dimensionality Reduction

It reduces redundancy implied among attributes. The curse of dimensions that as dimensionality increases data become more diverse, and any patterns are getting less significant and more peculiar, and the processing time may increase substantially. And we need to reduce the dimensionality to reduce redundancy and effects of the curse. We will reduce the dimensions by linear algebra techniques such as Principal component analysis (PCA) that is a statistical procedure that orthogonally transforms the original n coordinates of a data set into a new set of n coordinates, independent component analysis (ICA), random forests or ensemble trees, missing values ratio, low variance filter, high correlation filter, and Single value decomposition (SVD). Also, we can reduce the dimensionality by using feature selection that described above.

Feature Creation

It is to create a new set of features from the original features. The purpose of feature creation in the new feature space, meaningful and relevant patterns can be extracted more easily, and the number of features may be reduced. We will do feature creation by using feature extraction methods to extract new features from the existing ones, e.g., extracting colour, texture and shape from image of pixel values, by mapping data to a new space, e.g., wavelet transformation of pixel values of images to a frequency domain, and by constructing new features from the existing ones using domain knowledge, e.g., using transaction dates to construct a new feature customer tenure that indicates the loyalty of the customer to the company. Figure 34 shows an example on feature creation.

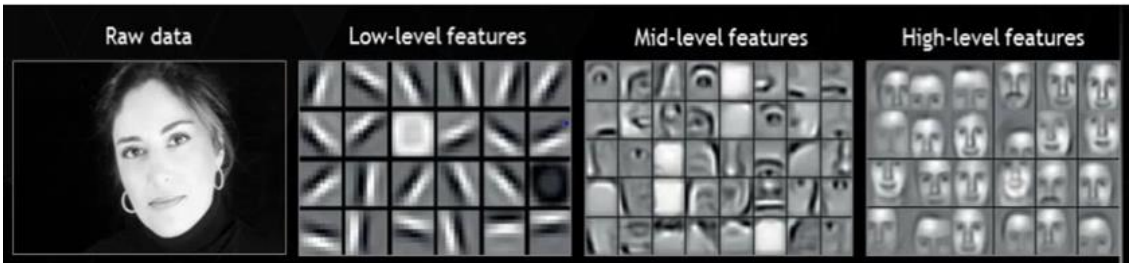


Figure 34: Feature Creation

Data Discretization

It is to convert continuous attribute values to discrete categorical values. The purpose for data discretization is requirement for some data mining solutions, and better data mining results (not

always). To do data discretization is by Deciding how many categories to have and where split points should be, and mapping values to categories. Figure 35 shows data discretization.

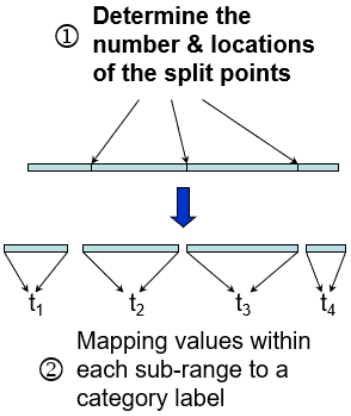


Figure 35: Data Discretization

The methods that are used for discretization are classified based on supervised and unsupervised learning. The unsupervised that is without concern to the outcome of a specific attribute, normally used for clustering and association rule mining e.g., equal width, equal depth, clustering. The supervised that is with respect to the outcome of the class attribute, normally used for classification could be entropy based or error based. Figure 36a an figure 36 b shows discretization example with the two methods (Supervised & Unsupervised).

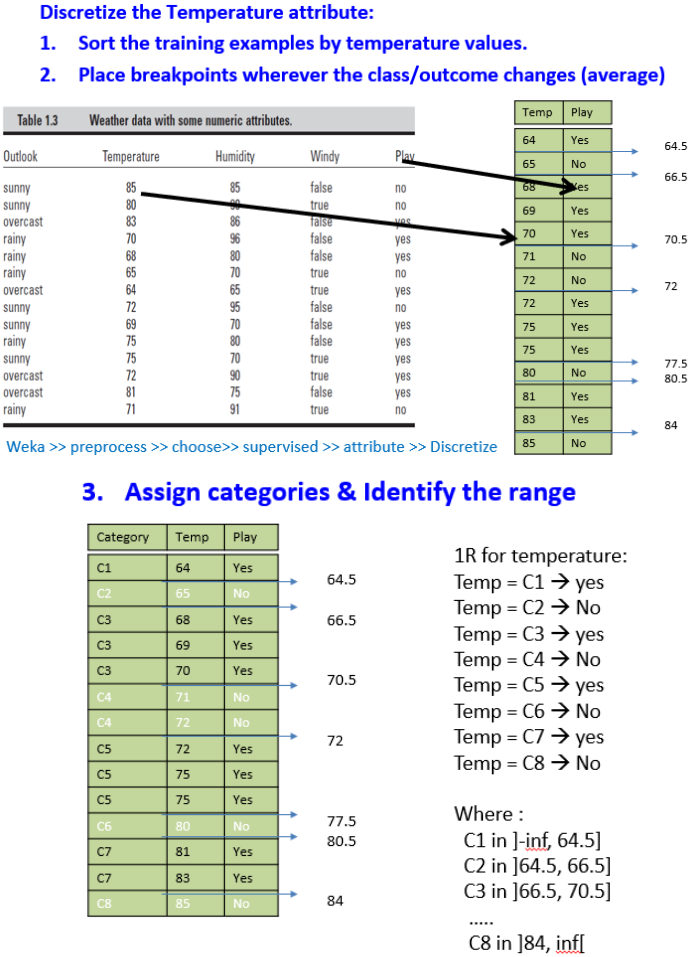


Figure 36a: Data Discretization Example using Supervised

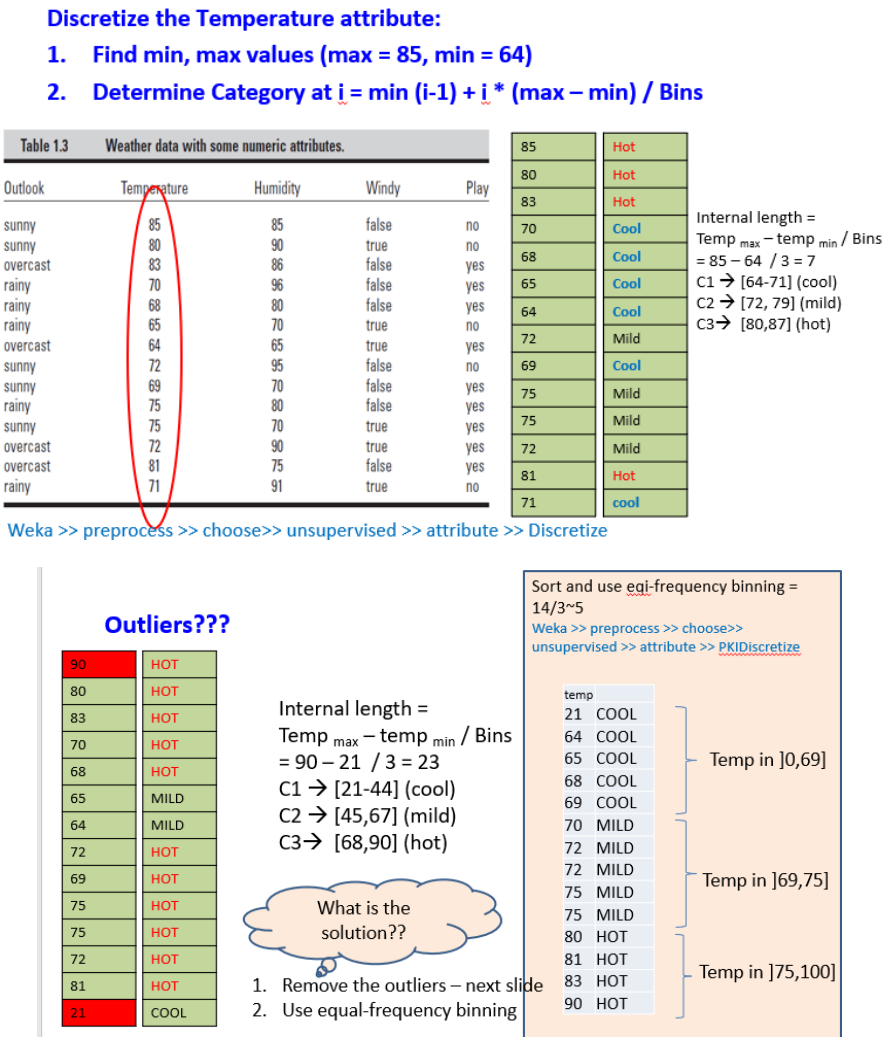


Figure 36b: Data Discretization Example using Unsupervised

Variable Transformation

It is transforming all values of an attribute to other values. The purpose of variable transformation is removing the effect of the outlier values, make the result data visualisation more interpretable, and make the values more comparable. We will do variable transformation by apply log function, standardization, and normalization in [0, 1] range.

$$z = \frac{x - \mu}{\sigma}$$

Standardization Formula

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Normalization Formula

Figure 37 shows variable transformation.

Employee	salary	Log(Salary)	Standardization	Normalization
1	10	1	-0.45	0.05
2	14	1.146128	-0.4	0.07
3	9	0.954243	-0.47	0.04
4	2	0.30103	-0.57	0.01
5	12	1.079181	-0.43	0.06
6	14	1.146128	-0.4	0.07
7	200	2.30103	2.38	1
8	22	1.342423	-0.28	0.11
9	18	1.255273	-0.34	0.09
10	14	1.146128	-0.4	0.07
11	12	1.079181	-0.43	0.06
12	11.5	1.060698	-0.43	0.05
13	7	0.845098	-0.5	0.03
14	9	0.954243	-0.47	0.04
15	1	0	-0.59	0
16	133	2.123852	1.38	0.66
17	200	2.30103	2.38	1

Figure 37: Variable Transformation

Nominal to Numeric

Also known as data binarization. It is changing the type of non-numeric attributes to a numeric type. The purpose of that data mining technique might be limited to numeric attributes because of metric usage (distance). To transfer nominal to numeric:

1. Apply domain knowledge, e.g., grades: A=90, B=80, …
2. Binary vector representation (one hot encoding)
    a. Each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.
    b. E.g., Grades: A, B, C, D, and F (5 attributes will replace the grade)

A = [1 0 0 0 0]
B = [0 1 0 0 0]
C = [0 0 1 0 0]
D = [0 0 0 1 0]
F = [0 0 0 0 1]

Handling Missing Values

It is to treat attributes with null values. And the reasons of missing values are malfunctioning equipment, changes in experiments design, collection of different datasets, measurement not possible, etc. The purpose for handling them is improve data quality, and better mining results. We will handle missing values by:

1. Elimination (may not always be possible).
2. Using sensible default, e.g., Spending Amount is set to 0.

3. By data imputation
    a. Average, median, or mode of the whole data population
    b. Average, median or mode of the nearest neighbors
4. Postponing the handling and making the mining methods adaptive to missing values

Now you learn how to preprocess a data set. So, the next step is data exploration to explore our data set and visualize it. The first section for data exploration is data visualization.

Data Exploration

Exploring data before mining and knowing it is essential for successful data mining. The purpose for knowing data is better understanding of the characteristics of data, better decision over data pre-processing tasks, and being able to discover some hidden patterns. The categories of data exploration techniques are statistics, data visualization (Using graphs and plots to visualize data, I explain it below), and online analytic processing (OLAP) that is an approach to answer multi-dimensional analytical queries swiftly in computing, and it is part of business intelligence (BI). Also, data exploration uses exploratory data analysis (EDA) that is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics.

Data Visualisation

We will use it because human eyes are good at spotting patterns, particularly visual patterns. Figure 38 shows examples on data visualization.

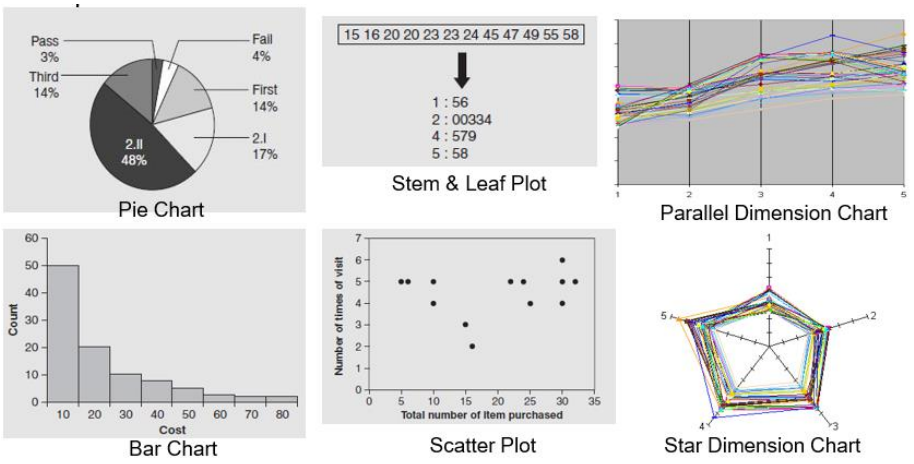


Figure 38: Data Visualisation Examples

References & Useful Resources:

- CSBP 320: Data Mining, UAE University
- Textbook: Hongbo Du.*Data Mining Techniques and Applications: an introduction*, 1st Edition