Data Preprocessing

Data preprocessing

The aim is to construct a full $n \times P$ data matrix with units on the rows and variables on the columns.
Obstacles : real-life data may be inconsistent, noisy and incomplete.

- Inconsistency : e.g. duplications, out-of-range values ;
- Noise : e.g. low quality sensors or sources ;
- Incompleteness : non response, lost analysis units .

Data preprocessing activities :

1. summarization
2. cleansing (imputation, noise reduction, detection of aberrant units)
3. transformation (integration, normalization, discretization)
4. reduction (aggregation, variable selection, dimension reduction)

Data preprocessing

Nature of variables (or features or attributes)

- Qualitative
  - Ordinal (e.g. income level : low, medium, high)
  - Nominal (binary, categorical e.g. eye color)
- Quantitative
  - Discrete (e.g. counts)
  - Continue (e.g. income)

Other kinds of (important) data types : documents, web pages, source codes, multimedia streams, ...

[Ex.] Classify the following variables from a household's survey :

1. Family name
2. Marital status
3. Server workload (e.g. in flops)
4. Number of cars
5. Date

1. Data summarization

a. Univariate distributions (summary measures)

Central tendency measures

- Mean : $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Median : splits the data roughly in two equal parts
- Mode : the most frequent value
- Quantiles : Q_1; Q_2(=median); Q_3

Dispersion measures

- Variance : $s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- Standard deviation $s_X = \sqrt{s_X^2}$
- Interquartile range : $IQR = Q_3 - Q_1$
- Range : $X_{\max} - X_{\min}$

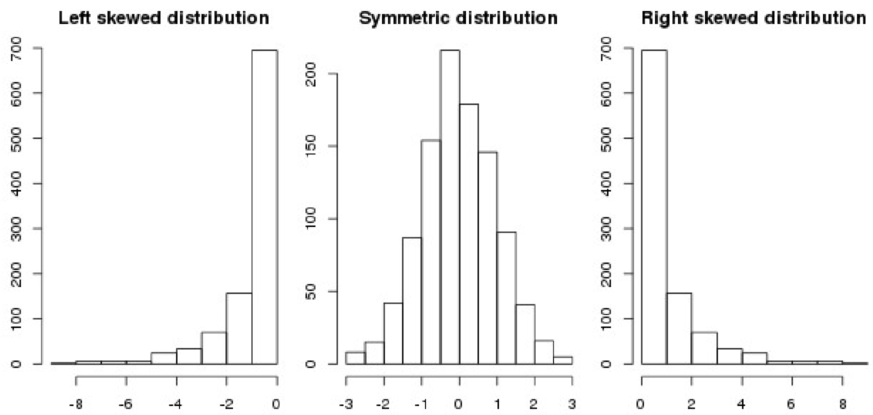
Shape measures

- Skewness (symmetry) coefficient. $SC = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \hat{X})^3}{s^3}$
- Kurtosis coefficient : how flat a distribution is. $KC = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \hat{X})^4}{s^4} - 3$

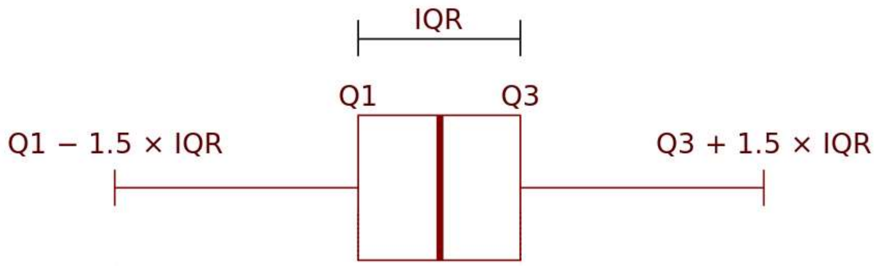
1. Data summarization

a. Univariate distributions (Graphical displays)

Histogram



Boxplot



1. Data summarization

a. Bivariate distributions

- (Cross) covariance between the variables X and Y

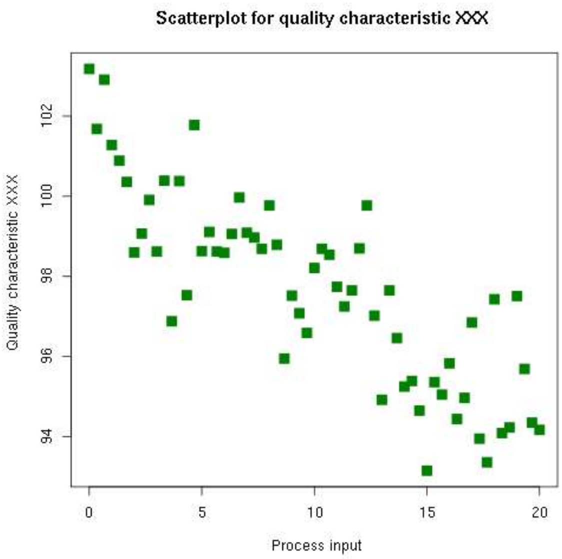
$$s_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

- Correlation coefficient measures only linear relationships

$$r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y}$$

We have that $|r_{X,Y}| \leq 1$ (prove it !)

Scatter plot



2. Data cleansing

a. Missing values

- A missing value occurs when an attribute is not recorded for a unit (they are usually coded, i.e. 999, `NA`)
- Many reasons : nonresponse, error, mistake.
- It may concern all the attributes of the unit or some of them.
- Missing values can reveal important information about the data ==> they can false the whole DM analysis.

Golden rule : all the efforts must be done during the data acquisition.

Copy Paste Quit

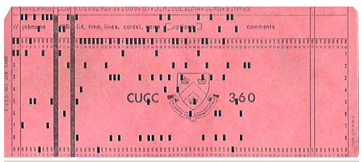
	row.names	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	13	4.8	3	1.4	0.1	setosa
2	10	4.9	3.1	1.5	0.1	setosa
3	107	4.9	2.5	4.5	1.7	
4	147	6.3	2.5	5	1.9	virginica
5	75	6.4	2.9	4.3	1.3	versicolor
6	2	4.9	3	1.4	0.2	setosa
7	138	6.4	3.1	5.5	1.8	virginica
8	34	5.5	NA	1.4	0.2	setosa
9	96	5.7	3	4.2	1.2	versicolor
10	113	6.8	3	5.5	NA	virginica

2. Data cleansing

a. Missing values (elimination & imputation)

However, one sometimes eliminates from the analysis the units with missing values or imputes them, i.e. fills them up with artificial values.

- If the missing values are completely at random and a very low fraction of n :
  - hot deck imputation (very bad idea !)
  - list-wise elimination
- More complex imputation schemes include :
  - mean / median value imputation for a quantitative attribute
  - mode imputation for a qualitative attribute
  - regression imputation



Copy Paste Quit

	row.names	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	13	4.8	3	1.4	0.1	setosa
2	10	4.9	3.1	1.5	0.1	setosa
3	107	4.9	2.5	4.5	1.7	
4	147	6.3	2.5	5	1.9	virginica
5	75	6.4	2.9	4.3	1.3	versicolor
6	2	4.9	3	1.4	0.2	setosa
7	138	6.4	3.1	5.5	1.8	virginica
8	34	5.5	NA	1.4	0.2	setosa
9	96	5.7	3	4.2	1.2	versicolor
10	113	6.8	3	5.5	NA	virginica

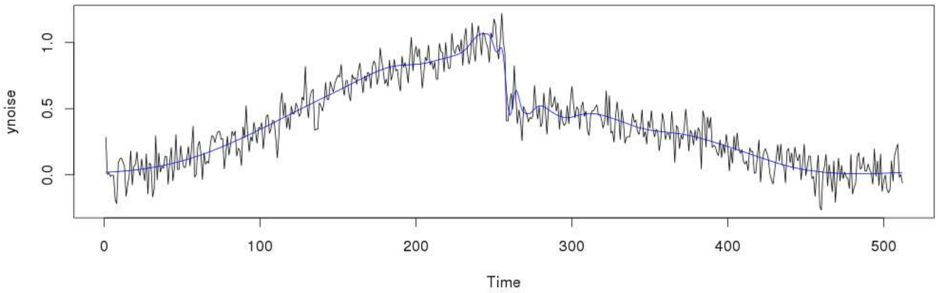
Also, you may rely on robust algorithms that will work even with `NA` .

2. Data cleansing

b. Noisy data

- Noise can be seen as unstructured (randomly) unwanted data
- Can be due to low quality technology in acquisition or transmission (i.e. cheap microphones or cables).
- Noise may difficult data mining (or fake it)

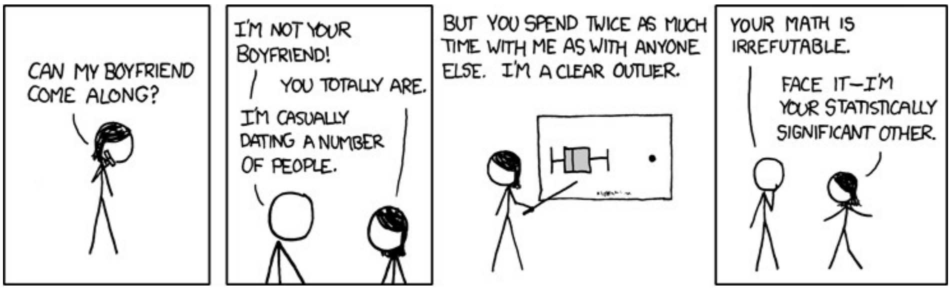
Noise can be reduced using smoothing filters or thresholding.



2. Data cleansing

c. Outliers & Influential units

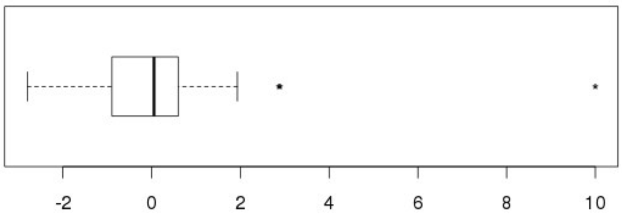
- An outlier is an unit that have a different probability structure from the pack. It may be due to measurement error or heavy-tailed distributions (i.e. high kurtosis) .
- An unit is influential if its deletion noticeably alters the result of the analysis.

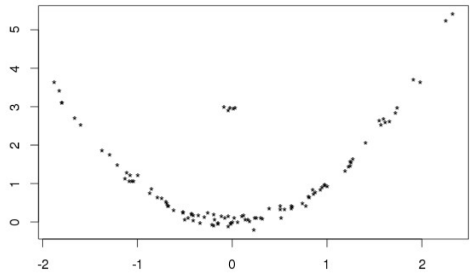


2. Data cleansing

c. Outliers & Influential units

Detection of outliers using graphical tools





A frequently used rule-of-the thumb is to assume that observations laying outside $Q_2 \pm 1.5 \times \text{IQR}$ are outliers (be very careful with this kind of rule).

3. Data transformation

Why would someone choose to transform the data ?

- Some techniques may need to transform the data in order to make it dimensionless, e.g. correlation coefficient, or to rend some hypothesis more reasonable (e.g. log for stabilize variance)
- Sometimes we are interest on categories instead of numerical scales (e.g. low, mid, high income instead of actual nominal income)
- Some techniques can not handle categorical values with more than two categories (i.e. binary variables)
- The target variables were not recorded but you have a proxy
- De-noising (check section 2.c)
- Aggregation : in order to change resolution of data

3. Data transformation

Data normalization

Min-Max normalization

- If x is the original attribute we compute a new attribute x^* by computing

$$x^* = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$$

- Linear transformation
- Maps data from the original range $[x_{\text{min}}, x_{\text{max}}]$ to $[0, 1]$

$z-$score normalization

- If x is the original attribute we compute a new attribute z_x by computing

$$z_x = \frac{x - \bar{x}}{s_x}$$

- Associated to the normalization of a normal random variable
- We use it implicitly when we compute correlations.
- The $z-$score normalizes an attribute to have zero mean and unitary standard deviation (Prove it !)

4. Data reduction

Reduction can be performed by

- selecting instances (i.e. rows of the data matrix)
- selecting features/attributes (i.e. columns of the data matrix)
- combining instances : e.g. data aggregation
- combining features : e.g. PCA (coming soon !)

Copy Paste Quit

	row.names	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	13	4.8	3	1.4	0.1	setosa
2	10	4.9	3.1	1.5	0.1	setosa
3	107	4.9	2.5	4.5	1.7	
4	147	6.3	2.5	5	1.9	virginica
5	75	6.4	2.9	4.3	1.3	versicolor
6	2	4.9	3	1.4	0.2	setosa
7	138	6.4	3.1	5.5	1.8	virginica
8	34	5.5	NA	1.4	0.2	setosa
9	96	5.7	3	4.2	1.2	versicolor
10	113	6.8	3	5.5	NA	virginica

The reduction may be wanted to reduce the computational time of some algorithms.