Dimensionality Reduction using R

Dr. Tanujit Chakraborty
Assistant Professor of Statistics
Sorbonne University Abu Dhabi

CONTENTS

SL No.	Today's Topics
1	Principal Component Analysis
2	Factor Analysis
3	Market Basket Analysis

PRINCIPAL COMPONENTS ANALYSIS

PRINCIPAL COMPONENTS ANALYSIS

- A dimensionality reduction technique
- Reduces the dimensionality of multivariate data without compromising much on the variation in the original data set.
- Achieved by transforming the original varaible into a new set of variables namely prinipal components (PCAs)
- PCAs are uncorrelated and ordered
- Hence the first few of them account for most of the variation in the orignal variables

PRINCIPAL COMPONENTS ANALYSIS

- Desribes the variation in a set of correlated variables $x = (x_1, x_2, - - -, x_q)$ by a set of uncorrelated variables $y = (y_1, y_2, - - -, y_q)$
- Each principal component is a linear combination of the x variables.
- The new variables are derived in decreasing order of importance.
- Hence y_1 account for maximum possible variation in x among all linear combinations of x
- y_2 account for maximum possible of the remaining variation subject to being uncorrelated to y1. and so on.

PRINCIPAL COMPONENTS ANALYSIS

- A dimensionality reduction technique
- Large number of correlated variables can be reduced to a manageable number of uncorrelated or independent factors.
- The emphasis is on the identification of underlying factors that might explain the dimensions associated with large data sets

$$y_i = a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3 + --- + a_{iq}x_q$$

Where y_i: estimate of i^{th} principal component, a_i: weight or score coefficient, x_i: i^{th} variable and k: number of variables

The coefficients are selected such that

- the first principal component explains largest portion of the total variation
- the second first principal component accounts for the most of the residual variance, etc.

PRINCIPAL COMPONENTS ANALYSIS

- Helps to understand the variability in large data sets with inter correlated variables using a smaller number of uncorrelated factors.
- Explaining variability of a set of n variables using *m* factors where $m < n$
- The emphasis is on the identification of underlying factors that might explain the dimensions associated with large data

Objectives

- Reduces the complexity of a large set of variables by summarizing them in a smaller set of components or factors
- Tries to improve the interpretation of complex data through logical factors

Computation of sample Principal Components

- The first principal component is that linear combination of original variables whose sample variance is greatest amongst all possible such linear combinations
- The second principal component is the linear combination of original variables that account for maximum proportion of the remaining variance subject to being uncorrelated with the first principal component and so on
- The first principal component is

$$y_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + --- + a_{1q}x_q$$

The variance of y_1 be increased by increasing $a_1 = (a_{11}, - - -, a_{1q})$, a restriction must be placed on these coefficients.

The sensible restriction or constraint is to ensure that sum of squares of the coefficients should be equal to one

$$a_1'a_1 = 1$$

Computation of sample Principal Components

- To choose the elements of a_1 which maximizes the variance of y_1 subject to the constraint of $a_1'a_1 = 1$
- Since y_1 is a linear combination of x, the sample variance of y_1 is given by

$$Var(y_1) = a_1'Sa_1$$

where S is the sample covariance matrix of x

The coefficients a_1 of first principal component y_1 is computed by solving

Maximize

$$z = a_1'Sa_1$$

Subject to

$$a_1'a_1 = 1$$

The solution to the above problem (using Lagrange multiplier method) is the Eigen vector of S corresponding to the largest Eigen value of S denoted by λ_1

Computation of sample Principal Components

In general, the coefficients a_i of first principal component y_i is computed by solving

Maximize

$$z = a_i'Sa_i$$

Subject to

$$a_i'a_i = 1$$

The solution to the above problem (using Lagrange multiplier method) is the Eigen vector of S corresponding to the i^{th} largest Eigen value of S denoted by λ_i

Since $a_i'a_i = 1$, the variance of i^{th} principal component y_1 will be λ_i

PRINCIPAL COMPONENTS ANALYSIS

Steps

- Prepare correlation matrix
- Extract a set of principal components using correlation matrix
- Determine the number of principal components
- Interpret results

PRINCIPAL COMPONENTS ANALYSIS

Example: Suppose a researcher wants to determine the underlying benefits consumers seek from the purchase of a toothpaste. A sample of 30 respondents was interviewed. The respondents were asked to indicate their degree of agreement with the following statements using a 7 point scale (1: strongly disagree, 7: strongly agree)

1. It is important to buy a toothpaste that prevents cavities
2. I like a toothpaste that gives shiny teeth
3. A toothpaste should strengthen your gums
4. I prefer toothpaste that freshens breath
5. Prevention of tooth decay is not an important benefit offered by a toothpaste
6. The most important consideration in buying a toothpaste is attractive teeth

Dataset: ***PCA_Factor_Analysis_Example.csv***

PRINCIPAL COMPONENTS ANALYSIS

Step 1: Normalize the data

z transform:

Transformed data = (Data – Mean) / SD

Reading ghe file to R

>mydata = mydata[,2:7]

Transforming the variables

>myzdata = scale(mydata)

Step 2: Check for Correlation

- Variables must be correlated for data reduction

```
>cor(myzdata)
```

Correlation Matrix

		x1	x2	x3	x4	x5	x6
Correlation	x1	1.000	-.053	.873	-.086	-.858	.004
	x2	-.053	1.000	-.155	.572	.020	.640
	x3	.873	-.155	1.000	-.248	-.778	-.018
	x4	-.086	.572	-.248	1.000	-.007	.640
	x5	-.858	.020	-.778	-.007	1.000	-.136
	x6	.004	.640	-.018	.640	-.136	1.000

High correlation between x_1, x_3 & x_5

Good correlation between x_2, x_4 & x_6

Step 4: Method used: Principle Component Analysis

>mymodel = princomp(myzdata)

>summary(mymodel)

Used to identify minimum number of components accounting for maximum variance in the data ; Eigen Values: Amount of variance attributed to a component.

Total Variance = 6 (Sum of all Eigen values)

Prop. variance for PC1= Eigen value of PC1 / Total Variance (2.731/6 = 0.455)

Component	SD	Variance	Proportion of Variance	Cumulative Proportion of Variance
PC 1	1.653	2.732	0.455	0.455
PC 2	1.489	2.217	0.369	0.825
PC 3	0.665	0.442	0.074	0.899
PC 4	0.584	0.341	0.057	0.955
PC 5	0.427	0.182	0.030	0.986
PC 6	0.292	0.085	0.014	1.000
Total		6.000		

PRINCIPAL COMPONENTS ANALYSIS

Step 4: Determine the number of Components

1.Based on Eigen Values: Only components with Eigen value > 1.0 or Eigen value > 0.7 are selected.

2. Based on cumulative % variance: Factors extracted should account for at least 65 % of variance

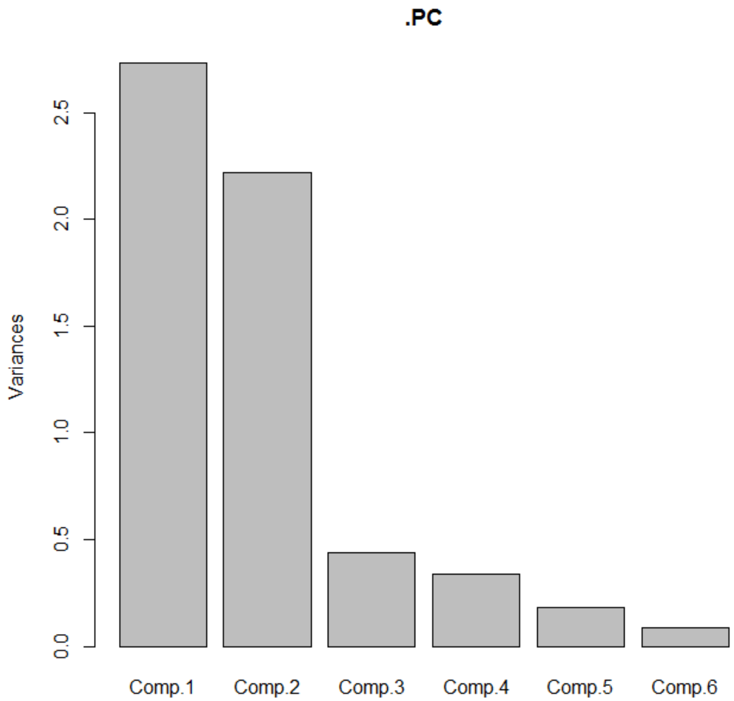
Component	SD	Variance	Proportion of Variance	Cumulative Proportion of Variance
PC 1	1.653	2.732	0.455	0.455
PC 2	1.489	2.217	0.369	0.825
PC 3	0.665	0.442	0.074	0.899
PC 4	0.584	0.341	0.057	0.955
PC 5	0.427	0.182	0.030	0.986
PC 6	0.292	0.085	0.014	1.000
Total		6.000		

Number of factors selected : 2

Step 4: Determine the number of Factors

```
>plot(mymodel)
```

3. Based on Scree plot: Plot of the Eigen values against the number of factors in order of extraction. The number of components is identified based on slope change of scree plot



Number of factors selected : 2

Step 5: Calculate Component Scores– Eigen Vectors

>loadings(mymodel)

$$y_i = a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3 + - - - + a_{ik}x_k$$

	Component	
	y_1	y_2
x_1	0.562	-0.170
x_2	-0.182	-0.534
x_3	0.566	-0.088
x_4	-0.207	-0.530
x_5	-0.526	0.236
x_6	-0.107	-0.585

Step 5: Interpret Components – Eigen Vectors

	Component	
	y_1	y_2
x_1	0.562	-0.170
x_2	-0.182	-0.534
x_3	0.566	-0.088
x_4	-0.207	-0.530
x_5	-0.526	0.236
x_6	-0.107	-0.585

Component 1 is correlated with x_1, x_3 & x_5

Component 2 is correlated with x_2, x_4 & x_6

Step 5: Interpret Components

	Component	
	y_1	y_2
Prevention of Cavities	0.562	-0.170
x_2	-0.182	-0.534
Strong Gum	0.566	-0.088
x_4	-0.207	-0.530
Non Prevention of Tooth Decay	-0.526	0.236
x_6	-0.107	-0.585

Interpretation

Component 1 (y_1) represents the health related benefits

PRINCIPAL COMPONENTS ANALYSIS

Step 5: Interpret Components

	Component	
	y_1	y_2
Prevention of Cavities	0.562	-0.170
Shiny Teeth	-0.182	-0.534
Strong Gum	0.566	-0.088
Fresh Breath	-0.207	-0.530
Non Prevention of Tooth Decay	-0.526	0.236
Attractive Teeth	-0.107	-0.585

Interpretation

Component 2 (y_2) represents the social related benefits

PRINCIPAL COMPONENTS ANALYSIS

Step 6: Reduced Data Set

>pc = mymodel$scores

>cbind(pc[,1], pc[,2])

Respondent	PC1	PC2	Respondent	PC1	PC2
1	1.953	-0.071	16	1.412	0.1352
2	-1.6763	0.9852	17	1.261	0.6098
3	2.4298	0.6577	18	2.5041	-0.2372
4	-0.0908	-1.6975	19	-1.2981	1.3974
5	-1.5154	2.7238	20	-1.2777	-1.7423
6	1.6696	0.0148	21	-1.449	1.7912
7	1.0622	1.1536	22	0.9783	-0.2455
8	2.0882	-0.5402	23	-1.4107	0.8217
9	-1.29	1.3543	24	-0.9281	-2.6799
10	-2.7958	-1.6321	25	1.4305	-0.0294
11	2.0398	0.3893	26	-1.0791	-2.2053
12	-1.6682	0.9421	27	1.4698	0.106
13	2.4379	0.6146	28	-1.5875	-1.2162
14	-0.4251	-1.9974	29	-0.8027	-3.2699
15	-1.6509	1.8801	30	-1.7904	1.987

EXPLORATORY FACTOR ANALYSIS

- Many times it may not be possible to measure some of the concepts directly
- Such cases the concepts are examined indirectly by collecting information on variables which can be directly measurable and assumed to be indicators of the concepts of interest.

Example

It is difficult to conclude a student is interested in science or arts directly.
The students scores on science subjects or arts subjects can be an indicator for the students interest in science or arts.

- Concepts which cannot be measured directly are called latent variables or factors
- The variables which can be directly measured and related to latent variables are called manifest variables

FACTOR ANALYSIS

The method of analysis to uncover the relationship between latent variables and manifest variables is factor analysis

The method is based on multiple regression, except in factor analysis manifest variables is regressed on unobservable latent variables

Types of Factor Analysis: Exploratory and Confirmatory

Exploratory Factor Analysis

Used to investigate the relationship between factors and manifest variables without making any assumption about which manifest variables is related to which factors

Confirmatory Factor Analysis

Used to test whether a specific factor model postulated a priori on the relationship between factors and manifest variables is correct or not

Factor analysis model

A regression model linking the manifest variables to a set of unobserved (or unobservable) latent variables

Assumes that the observed relationships between the manifest variables are the result of relationship between manifest variables and latent variables

The relationship between the manifest variables is measured using covariance matrix or correlation matrix.

Factor analysis model

Let a set of observed or manifest variables x = (x_1, x_2, - - -, x_q) be linked to k unobserved latent variables or common factors f_1, f_2, - - -, f_k, where k < q by the regression model given by

$$x_1 = \lambda_{11} f_1 + \lambda_{12} f_2 + - - - + \lambda_{1k} f_k + \mu_1$$

$$x_2 = \lambda_{21} f_1 + \lambda_{22} f_2 + - - - + \lambda_{2k} f_k + \mu_2$$

- -

- - - - - - - - -- - - - - - - - - - - - - - - -

$$x_q = \lambda_{q1} f_1 + \lambda_{q2} f_2 + - - - + \lambda_{qk} f_k + \mu_q$$

Where

λ_j are regression coefficients of the x variables on the common factors known as factor loadings

Shows how each observed variable x_i, depends on the common factors

Factor analysis model

The regression model is

$$x = \Lambda f + \mu$$

Assumptions

The random disturbance terms μ_1, μ_2, - - -, μ_q are uncorrelated with each other and with the factors f_1, f_2, - - -, f_k.

Hence correlation between the observed variables arise from their relationship with the common factors.

The factors f_1, f_2, - - -, f_k also uncorrelated and occur in the standardized form with mean zero and standard deviation one

Principal Component Method of factor analysis model

Very similar to principal component analysis but not operating directly on S or R but on the reduced covariance matrix S*

$$S^* = S - \psi$$

Steps

- Prepare correlation matrix
- Extract a set of factors using correlation matrix
- Determine the number of factors
- Rotate factors to increase interpretability
- Interpret results

Example: Suppose a researcher wants to determine the underlying benefits consumers seek from the purchase of a toothpaste. A sample of 30 respondents was interviewed. The respondents were asked to indicate their degree of agreement with the following statements using a 7 point scale (1: strongly disagree, 7: strongly agree)

1. It is important to buy a toothpaste that prevents cavities
2. I like a toothpaste that gives shiny teeth
3. A toothpaste should strengthen your gums
4. I prefer toothpaste that freshens breath
5. Prevention of tooth decay is not an important benefit offered by a toothpaste
6. The most important consideration in buying a toothpaste is attractive teeth

Dataset: ***PCA_Factor_Analysis_Example.csv***

Step 1: Normalize the data

z transform:

Transformed data = (Data – Mean) / SD

Reading ghe file to R

```
>mydata = mydata[,2:7]
```

Transforming the variables

```
>myzdata = scale(mydata)
```

Step 2: Check for Correlation

- Variables must be correlated for data reduction

> cor(myzdata)

Correlation Matrix

		x1	x2	x3	x4	x5	x6
Correlation	x1	1.000	-.053	.873	-.086	-.858	.004
	x2	-.053	1.000	-.155	.572	.020	.640
	x3	.873	-.155	1.000	-.248	-.778	-.018
	x4	-.086	.572	-.248	1.000	-.007	.640
	x5	-.858	.020	-.778	-.007	1.000	-.136
	x6	.004	.640	-.018	.640	-.136	1.000

High correlation between x1, x3 & x5

Good correlation between x2, x4 & x6

FACTOR ANALYSIS

Step 3: Check for Sampling (factor) adequacy

```
>library(psych)
> KMO(myzdata)
```

Statistics	Value	Criteria
Kaiser, Meyer, Olkin (KMO)	0.66	> 0.5

Step 4: Identifying the number of factors

Compute eigen values

Choose the factors with eigen values > 1

```
>s = cov(myzdata)
>s_eigen = eigen(s)
>variance = s_eigen$values
```

Factor	Variance	% Variance	Cum % Variance
F1	2.731188	45.52	45.52
F2	2.218119	36.97	82.49
F3	0.441598	7.36	89.85
F4	0.341258	5.69	95.54
F5	0.182628	3.04	98.58
F6	0.085209	1.42	100.00
Total	6		

FACTOR ANALYSIS

Step 4: Determine the number of Factors

1. Based on Eigen Values: Only factors with Eigen value > 1.0 are selected
2. Based on cumulative % variance: Factors extracted should account for at least 65 % of variance

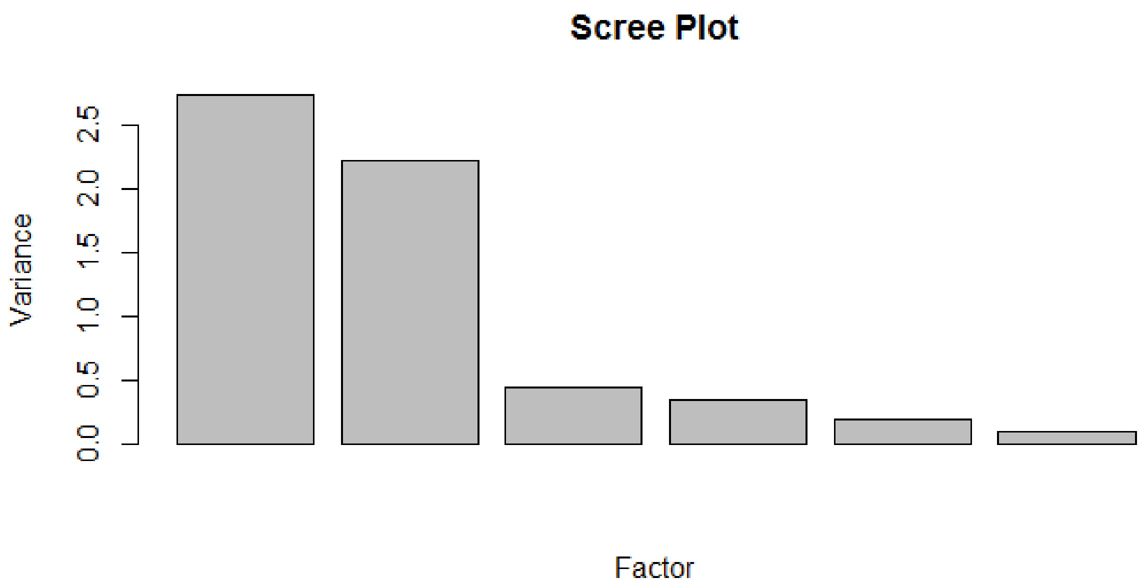
Factor	Variance	% Variance	Cum % Variance
F1	2.731188	45.52	45.52
F2	2.218119	36.97	82.49
F3	0.441598	7.36	89.85
F4	0.341258	5.69	95.54
F5	0.182628	3.04	98.58
F6	0.085209	1.42	100.00
Total	6		

Number of factors selected : 2

Step 4: Determine the number of Factors

```
> barplot(variance, xlab = "Factor", ylab = 'Variance', main = "Scree Plot“)
```

3. Based on Scree plot: Plot of the eigen values against the number of factors in order of extraction. The number of factors is identified based on slope change of scree plot



Number of factors selected : 2

Step 5: Calculate Factor Scores

```
> mymodel = factanal(myzdata,
  2)
```

	Component	
	1	2
x1	0.968	0.000
x2	0.000	0.749
x3	0.898	-0.140
x4	0.000	0.784
x5	-0.887	0.236
x6	0.000	0.830

Interpretation is difficult when the variables are evenly loaded on many factors
Solution: Rotation

Step 5: Calculate Factor Scores: Rotation

A process by which a solution is made more interpretable without changing its underlying mathematical properties.

Types of rotations

1. Orthogonal rotation
2. Oblique rotation

Orthogonal rotation

Restricts the rotated factors to being uncorrelated

Oblique rotation

allows the rotated factors to be correlated

Step 5: Calculate Factor Scores: Rotation

A process by which a solution is made more interpretable without changing its underlying mathematical properties.

Commonly used rotation : Orthogonal rotation

Commonly used orthogonal rotation : varimax rotation

Try to achieve factors with a few large loadings and as many near – zero loadings as possible

Step 5: Calculate Factor Scores: Rotation

```
> myrotatedmodel = factanal(myzdata, 2, rotation = "varimax", scores = "regression")
> myrotatedmodel
```

	Component	
	1	2
x1	0.968	0.000
x2	0.000	0.749
x3	0.898	-0.140
x4	0.000	0.784
x5	-0.887	0.236
x6	0.000	0.830

Step 5: Interpret Components – Eigen Vectors

	Component	
	1	2
x1	0.968	0.000
x2	0.000	0.749
x3	0.898	-0.140
x4	0.000	0.784
x5	-0.887	0.236
x6	0.000	0.830

Component 1 is correlated with x1, x3 & x5

Component 2 is correlated with x2, x4 & x6

Step 5: Interpret Components

	Component	
	1	2
Prevention of Cavities	0.968	0.000
x2	0.000	0.749
Strong Gum	0.898	-0.140
x4	0.000	0.784
Non Prevention of Tooth Decay	-0.887	0.236
x6	0.000	0.830

Interpretation

Component 1 represents the health related benefits

Step 5: Interpret Components

	Component	
	1	2
Prevention of Cavities	0.968	0.000
Shiny Teeth	0.000	0.749
Strong Gum	0.898	-0.140
Fresh Breath	0.000	0.784
Non Prevention of Tooth Decay	-0.887	0.236
Attractive Teeth	0.000	0.830

Interpretation

Component 2 represents the social related benefits

Step 6: Reduced Data Set

```
> output = myrotatedmodel$scores
> output
```

Respondent	Factor1	Factor2	Respondent	Factor1	Factor2
1	1.3046	-0.2413	16	0.8934	-0.342
2	-1.2952	-0.2556	17	0.5714	-0.5502
3	1.1629	-0.7569	18	1.501	-0.24
4	0.1747	1.0108	19	-0.9845	-0.6938
5	-1.428	-1.3608	20	-0.4187	1.259
6	0.9864	-0.2511	21	-1.3132	-0.8042
7	0.4605	-0.9084	22	0.5706	-0.0143
8	1.1867	-0.0515	23	-0.9855	-0.1067
9	-0.7678	-0.6358	24	0.0209	1.7277
10	-1.1191	1.3473	25	0.8821	-0.1881
11	1.0738	-0.6495	26	-0.3011	1.5195
12	-1.0785	-0.1976	27	0.4675	-0.2914
13	1.3796	-0.6989	28	-0.5606	0.7776
14	0.0978	1.2286	29	0.1111	2.1146
15	-1.394	-0.7847	30	-1.1988	-0.9623

MARKET BASKET ANALYSIS

MARKET BASKET ANALYSIS

A modeling technique based upon the logic that if a customer buy a certain group of items, he is more (or less) likely to buy another group of items

Example:

Those who buy cigarettes are more likely to buy match box also.

MARKET BASKET ANALYSIS

Association Rule Mining:

Developing rules that predict the occurrence of of an item based on the occurrence of other items in the transaction

Example

Id	Items
1	Milk, Bread
2	Bread, Biscuits, Toys, Eggs
3	Milk, Biscuits, Toys, Fruits
4	Bread, Milk, Toys, Biscuits
5	Milk, Bread, Biscuits, Fruits

{Milk, Bread} → {Biscuits} with probability = 2 / 3

MARKET BASKET ANALYSIS

Itemset:

A collection of one or more items

k – itemset

An itemset consisting of k items

Id	Items
1	Milk, Bread
2	Bread, Biscuits, Toys, Eggs
3	Milk, Biscuits, Toys, Fruits
4	Bread, Milk, Toys, Biscuits
5	Milk, Bread, Biscuits, Fruits

MARKET BASKET ANALYSIS

Support count:

Frequency of occurrence of an itemset

Example

{Milk, Bread, Biscuits} = 2

Id	Items
1	Milk, Bread
2	Bread, Biscuits, Toys, Eggs
3	Milk, Biscuits, Toys, Fruits
4	Bread, Milk, Toys, Biscuits
5	Milk, Bread, Biscuits, Fruits

Support :

Proportion or fraction of transaction that contain an itemset

Example

{Milk, Bread, Biscuits} = 2 / 5

Id	Items
1	Milk, Bread
2	Bread, Biscuits, Toys, Eggs
3	Milk, Biscuits, Toys, Fruits
4	Bread, Milk, Toys, Biscuits
5	Milk, Bread, Biscuits, Fruits

Frequent Itemset

An itemset whose support is greater than or equal to minimum support

MARKET BASKET ANALYSIS

Confidence

Conditional probability that an item will appear in transactions that contain another items

Example

Confidence that Toys will appear in transaction containing Milk & Biscuits

= {Milk, Biscuits, Toys} / {Milk, Biscuits} = 2 / 3 = 0.67

Id	Items
1	Milk, Bread
2	Bread, Biscuits, Toys, Eggs
3	Milk, Biscuits, Toys, Fruits
4	Bread, Milk, Toys, Biscuits
5	Milk, Bread, Biscuits, Fruits

MARKET BASKET ANALYSIS

Association Rule Mining

1. Frequent Itemset Generation

   Fix minimum support value

   Generate all itemsets whose support $\geq$ minimum support

2. Rule Generation

   Fix minimum confidence value

   Generate high confidence rules from each frequent itemset

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

a. Fix minimum support count

b. Generate all itemsets of length = 1

c. Calculate the support for each itemset

d. Eliminate all itemsets with support count < minimum support count

e. Repeat steps c & d for itemsets of length = 2, 3, ---

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Id	Items
1	A,C,D
2	B,C,E
3	A,B,C,E
4	B,E
5	A,E
6	A,C,E

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 1:

Generate itemsets of length = 1 & calculate support

Item	Support count
A	4
B	3
C	4
D	1
E	5

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 2:

eliminate itemsets with support count < minimum support count (2)

Item	Support count
A	4
B	3
C	4
D	1
E	5

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 2:

eliminate itemsets with support count < minimum support count (2)

Item	Support count
A	4
B	3
C	4
E	5

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 3:

generate itemsets of length = 2

Item	Support count
A, B	1
A, C	3
A,E	3
B, C	2
B, E	3
C,E	3

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 4:

eliminate itemsets with support count < minimum support count (2)

Item	Support count
A, B	1
A, C	3
A,E	3
B, C	2
B, E	3
C,E	3

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 4:

eliminate itemsets with support count < minimum support count (2)

Item	Support count
A, C	3
A,E	3
B, C	2
B, E	3
C,E	3

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 5:

generate itemsets of length = 3

Item	Support count
A, C, E	2
B, C, E	2

Step 6:

generate itemsets of length = 4

Itemset	Support Count
A, B, C, E	1

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Result:

Item	Support count	Support
A, C, E	2	0.33
B, C, E	2	0.33
A , C	3	0.50
A , E	3	0.50
B,C	2	0.33
B,E	3	0.50
C,E	3	0.50

MARKET BASKET ANALYSIS

Association Rule Mining: Apriori Algorithm

Example:

Minimum Support = 0.50

Minimum Confidence = 0.5

Item	Support count	Support
A, C, E	2	0.33
B, C, E	2	0.33
A , C	3	0.50
A , E	3	0.50
B,C	2	0.33
B,E	3	0.50
C,E	3	0.50

Association Rule Mining: Apriori Algorithm

Example:

Minimum Support = 0.50

Minimum Confidence = 0.5

Item	Support	Confidence
A → C	0.50	0.75
A → E	0.50	0.75
B → E	0.50	1.00
C → E	0.50	0.75
C → A	0.50	0.75
E → A	0.50	0.60
E → B	0.50	0.60
E → C	0.50	0.60

Association Rule Mining: Other Measures

Lift

Lift (A $\longrightarrow$ C) = Confidence (A $\longrightarrow$ C) / Support (C)

Example

Item	Confidence	Support	Lift
A $\longrightarrow$ C	0.75	C = 0.67	1.12
A $\longrightarrow$ E	0.75	E = 0.83	0.93

Criteria : Lift $\geq$ 1

Lift (A , C) = 1.12 > Lift (A , E) indicates that A has a greater impact on the frequency of C than it has on the frequency of E

MARKET BASKET ANALYSIS

R Code

Reading the file and variables

```
>target = mydata$items
>ident = mydata$Id
```

Making transactions

```
>library(arules)
>transactions = as(split(target, ident),”transactions”)
```

Generating rules

```
>library(arules)
>myrules = apriori(transactions, parameter = list(support = 0.5, confidence = 0.25, minlen =  2))
```

Displaying rules

```
>myrules
>inspect(myrules)
```

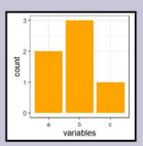
CHEAT SHEET

What type of **DATA VISUALIZATION** to choose?

What do you want to show?

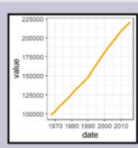
COMPARISON

COMPARISON CHARTS SHOW THE DIFFERENCES BETWEEN VALUES SO YOU CAN QUICKLY COMPARE CATEGORIES AS WELL AS SEE HOW VALUES CHANGE OVER TIME.



BAR PLOT

COMPARING CATEGORIES WITHIN THE SAME MEASURE OR THE SAME MEASURES.

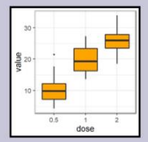


LINE PLOT

COMPARING TRENDS OVER TIME.

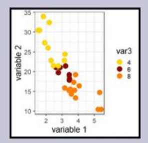
RELATIONSHIP

RELATIONSHIP CHARTS ARE USED TO EXPLORE RELATIONSHIPS BETWEEN VALUES. THEY ALLOW YOU TO FIND CORRELATIONS, OUTLIERS AND CLUSTERS OF DATA.



BOXPLOT

DISPLAYING OUTLIERS AND DATA CLUSTERS.

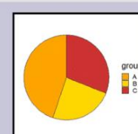


SCATTER PLOT

DISPLAYING THE RELATIONSHIP BETWEEN TWO OR THREE MEASURES FOR A DIMENISON.

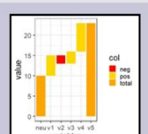
COMPOSITION

COMPOSITION CHARTS ARE USED TO ANALYZE HOW EACH COMPONENT VALUE AFFECTS TO TOTAL.



PIE CHART

DISPLAYING A STATIC COMPOSITION OF VALUES.

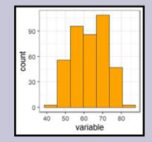


WATERFALL CHART

DISPLAYING THE STATIC COMPOSITION OF A VALUE WITH ACCUMULATION OR SUBTRACTION FROM THE TOTAL.

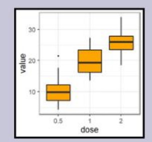
DISTRIBUTION

DISTRIBUTION CHARTS ARE USED TO EXPLORE HOW VALUES ARE GROUPED IN YOUR DATA.



HISTOGRAM

DISPLAYING THE DISTRIBUTION OF DATA INTERVALS.

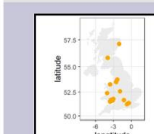


BOXPLOT

DISPLAYING RANGES AND DISTIRIBUTION OF NUMERIC DATA.

GEOGRAPHICAL DATA

GEORGAPHIC CHARTS PRESENT DATA BY GEOGRAPHIC LOCATION ON A MAP AS POINTS OR AREAS.



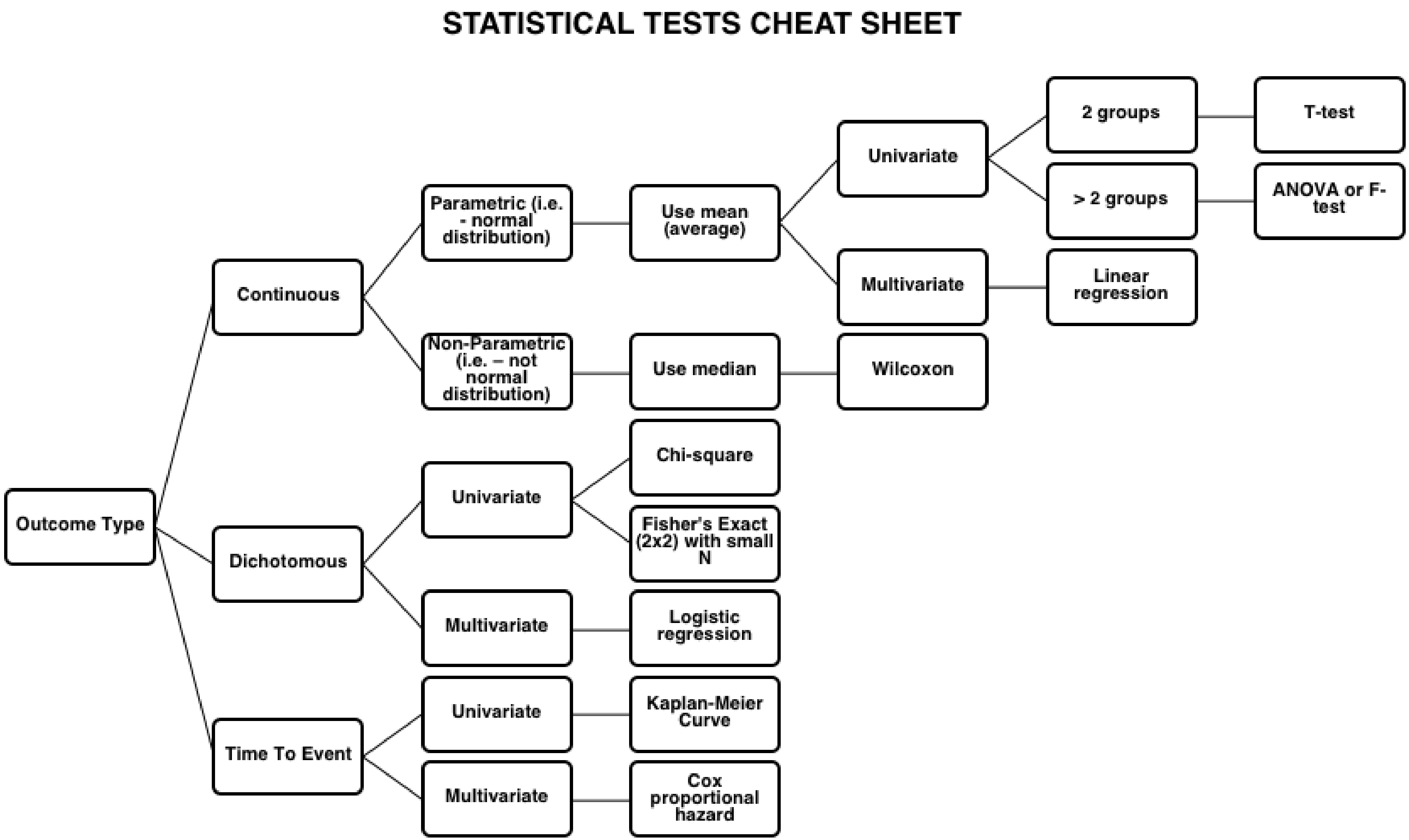
MAP

DISPLAYING DATA REPRESENTED GEOGRAPHICALLY BY A POINT OR AREA.

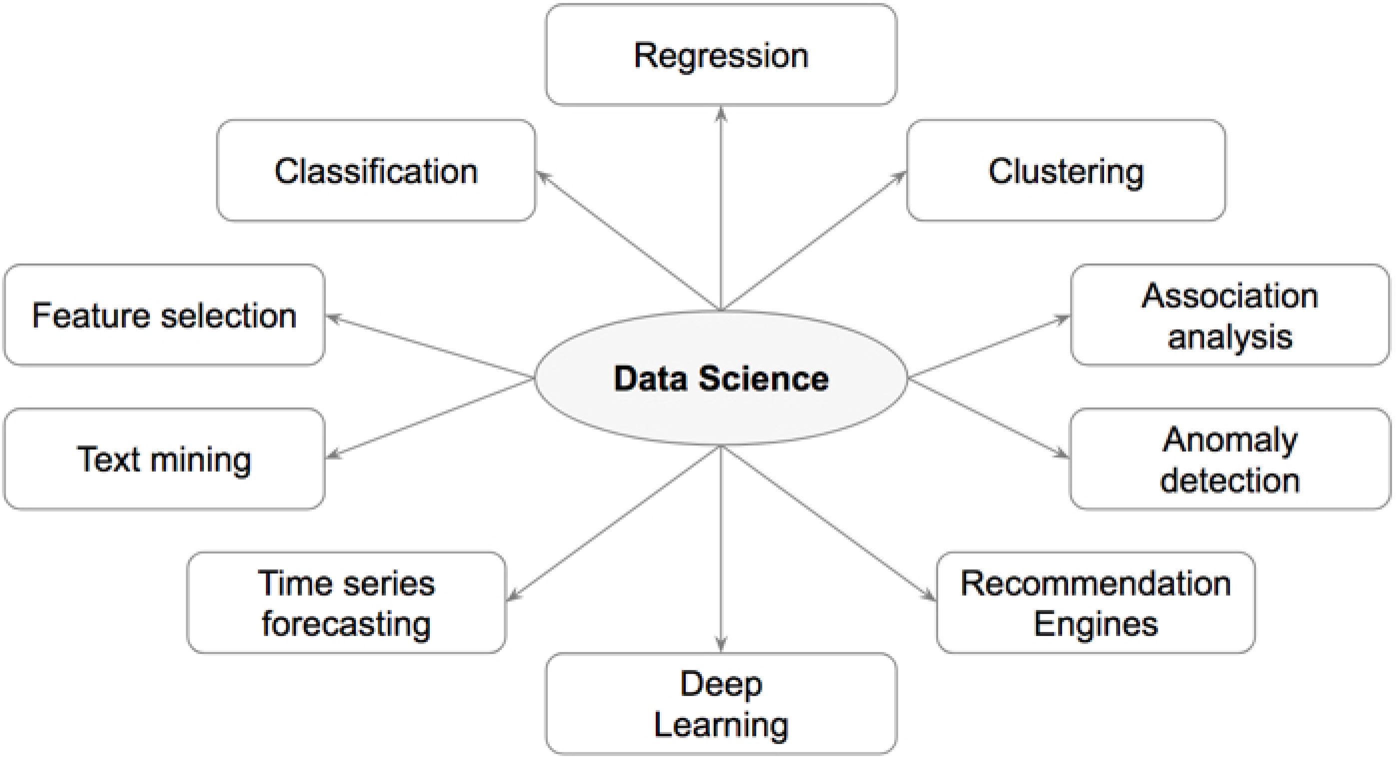
REMEMBER! THESE CHARTS ARE JUST AN EXAMPLE. ALWAYS USE A CHART THAT REPRESENTS YOUR DATA MOST TRANSPARENT AND WITHOUT MISUNDERSTANDING.

PAULINA JASZCZUK | PAWEŁ FIJAŁKOWSKI | ADAM FREJ | MICHAŁ PIASECKI

CHEAT SHEET

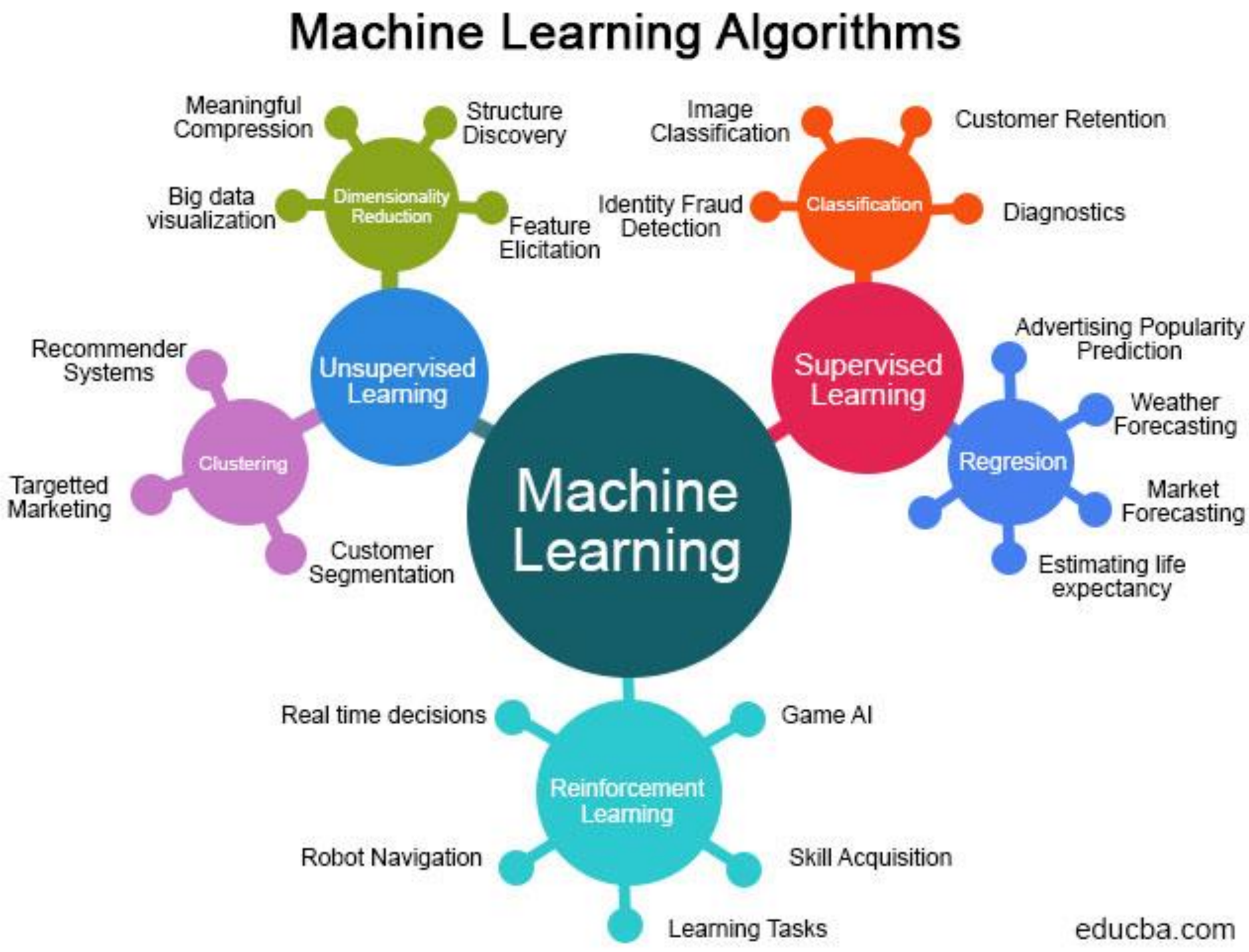

STATISTICAL TESTS CHEAT SHEET
Outcome Type
Continuous
Parametric (i.e. - normal distribution)
Use mean (average)
Univariate
2 groups
T-test
> 2 groups
ANOVA or F-test
Multivariate
Linear regression
Non-Parametric (i.e. – not normal distribution)
Use median
Wilcoxon
Dichotomous
Univariate
Chi-square
Fisher's Exact (2x2) with small N
Multivariate
Logistic regression
Time To Event
Univariate
Kaplan-Meier Curve
Multivariate
Cox proportional hazard

CHEAT SHEET



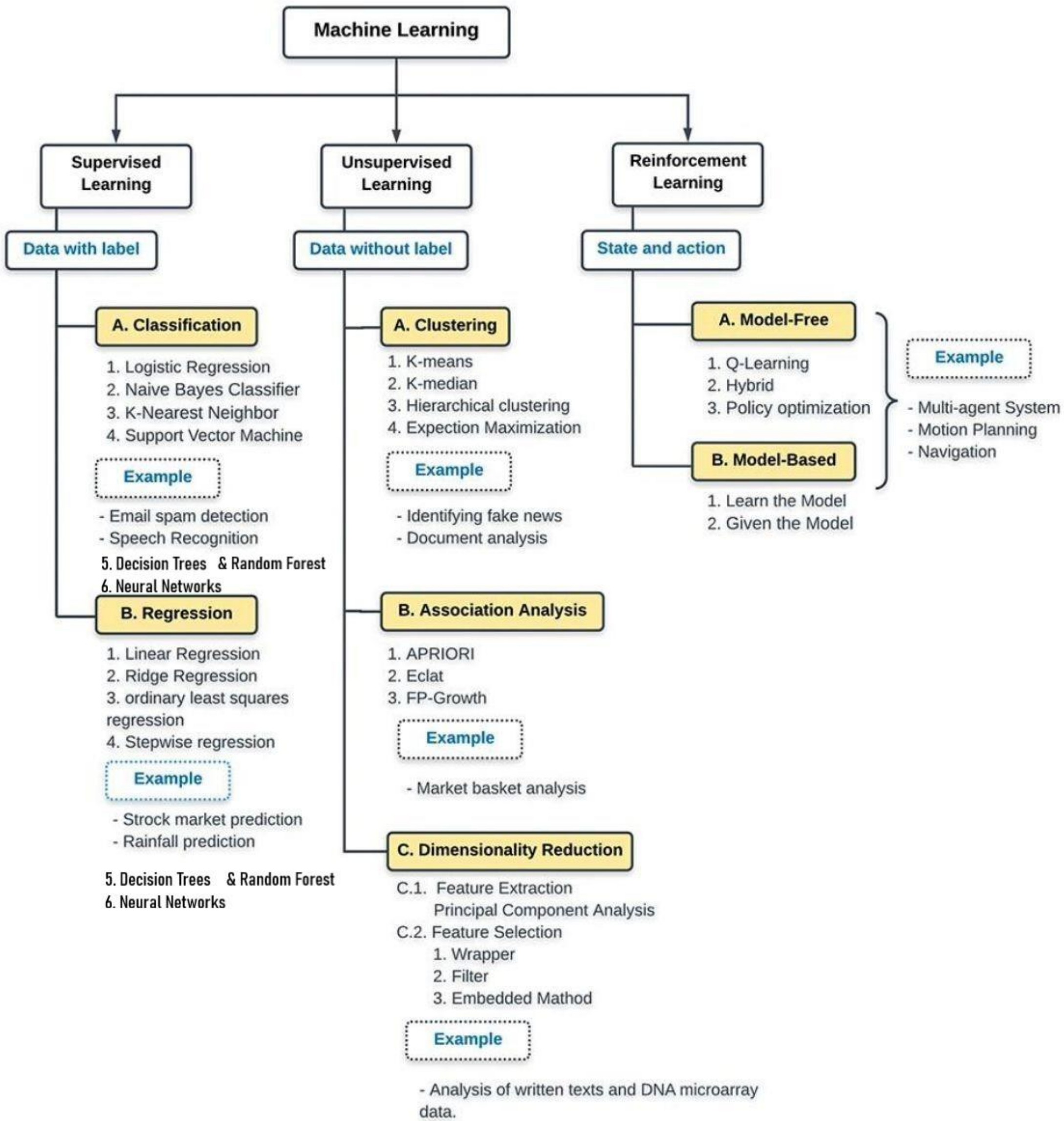
Regression
Classification
Clustering
Feature selection
Data Science
Association analysis
Text mining
Anomaly detection
Time series forecasting
Recommendation Engines
Deep Learning

CHEAT SHEET



Machine Learning Algorithms
Meaningful Compression
Structure Discovery
Big data visualization
Dimensionality Reduction
Feature Elicitation
Image Classification
Customer Retention
Identity Fraud Detection
Classification
Diagnostics
Recommender Systems
Unsupervised Learning
Supervised Learning
Advertising Popularity Prediction
Weather Forecasting
Clustering
Targetted Marketing
Machine Learning
Regresion
Market Forecasting
Customer Segmentation
Estimating life expectancy
Real time decisions
Game AI
Reinforcement Learning
Robot Navigation
Skill Acquisition
Learning Tasks
educba.com

CHEAT SHEET

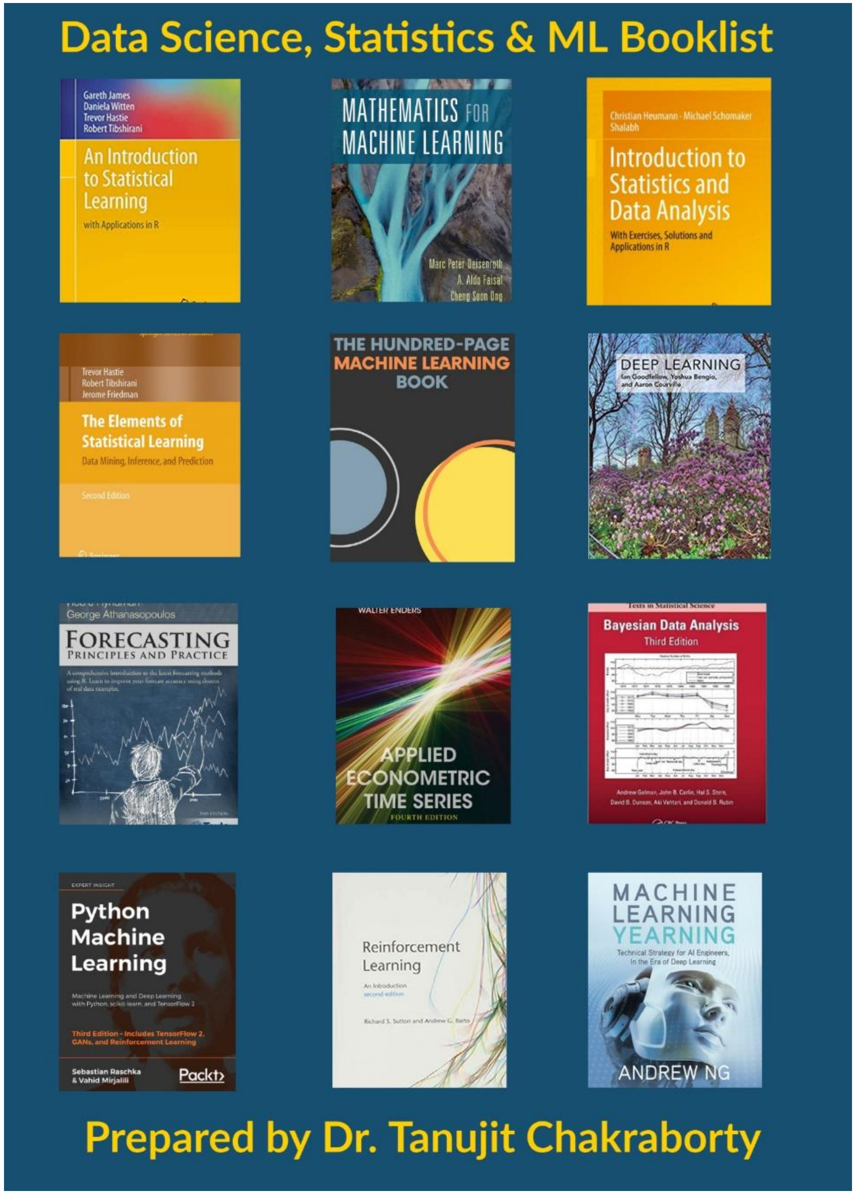


Machine Learning
Supervised Learning
Data with label
A. Classification
1. Logistic Regression
2. Naive Bayes Classifier
3. K-Nearest Neighbor
4. Support Vector Machine
Example
- Email spam detection
- Speech Recognition
5. Decision Trees & Random Forest
6. Neural Networks
B. Regression
1. Linear Regression
2. Ridge Regression
3. ordinary least squares regression
4. Stepwise regression
Example
- Strock market prediction
- Rainfall prediction
5. Decision Trees & Random Forest
6. Neural Networks
Unsupervised Learning
Data without label
A. Clustering
1. K-means
2. K-median
3. Hierarchical clustering
4. Expection Maximization
Example
- Identifying fake news
- Document analysis
B. Association Analysis
1. APRIORI
2. Eclat
3. FP-Growth
Example
- Market basket analysis
C. Dimensionality Reduction
C.1. Feature Extraction
Principal Component Analysis
C.2. Feature Selection
1. Wrapper
2. Filter
3. Embedded Mathod
Example
- Analysis of written texts and DNA microarray data.
Reinforcement Learning
State and action
A. Model-Free
1. Q-Learning
2. Hybrid
3. Policy optimization
B. Model-Based
1. Learn the Model
2. Given the Model
Example
- Multi-agent System
- Motion Planning
- Navigation

CHEAT SHEET

Dependent Variable Type (Ys)	Independent Variable Type (Xs)	Modelling Technique
Numerical	Numerical	1. Linear Regression or Best Subset Regression 2. Non-linear Regression or Regression Splines 3. Regression Trees, Neural Nets, etc.
Numerical	Categorical + Numerical	1. Linear Regression with Dummy Variables 2. Polynomial Regression with Dummy Variables 3. Regression Trees, Neural Nets, etc.
Categorical	Numerical	1. Logistics Regression 2. Classification Trees 3. Support Vector Machines, Neural Nets, etc.
Categorical	Categorical + Numerical	1. Logistic Regression with Dummy Variables 2. Classification Trees 3. Advanced Neural Nets, etc.
Numerical (Time dependent)	Numerical Exogenous Variables	1. ARIMA, ETS, Naïve Model 2. Autoregressive Neural Network 3. RNN, LSTM, etc.

TEXTBOOKS & REFERENCES



END OF MULTIVARIATE DATA ANALYTICS COURSE

> “If you’re brave enough to say goodbye, life will reward you with a new hello.”
>
> – *Paulo Coelho*