Introduction to Data Science

Last Update: 7 July 2021

Chapter 3
Descriptive Analysis

Papangkorn Inkeaw, PhD

Department of Computer Science, Faculty of Science
Chiang Mai University

Outline

Descriptive Analysis

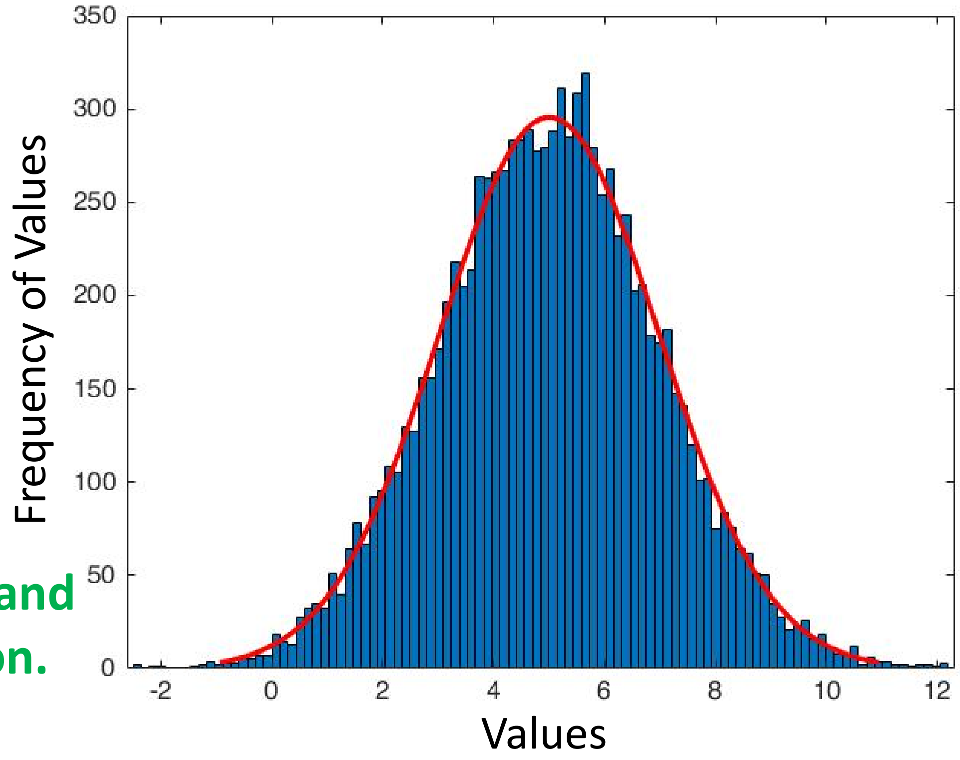
1. Descriptive Statistics with Pivot Tables
   - Mean, Median and Mode
   - Variance and Standard Deviation
   - Skewness and Kurtosis
   - Covariance Matrix
2. Cluster Analysis
   - Distances
   - K-means Clustering
   - Hierarchical Clustering
   - Density-based Spatial Clustering
3. Association Analysis
   - Itemset Mining
   - Association Rules

Descriptive Statistics with Pivot Tables

Mean, Median and Mode

Descriptive Statistics with Pivot Tables

	X_1	X_2	…	X_{10}
$\mathbf{x}_1$				
…				
$\mathbf{x}_n$				



We can slice a feature/variable and describe it as a data distribution.

A distribution in statistics is a function that shows:

- the possible values for a variable (x-axis)
- how often they occur (y-axis).

Mean, Median and Mode

Descriptive Statistics with Pivot Tables

Mean

- A measure of a central or typical value for a probability distribution.
- The sum of all measurements divided by the number of observations in the data set.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Example:

Job performance: 7, 10, 11, 15, 10, 10, 12, 14, 16, 12

Mean of job performance:

$$\bar{x} = \frac{7+10+11+15+10+10+12+14+16+12}{10} = \frac{117}{10} = 11.7$$

Mean, Median and Mode

Descriptive Statistics with Pivot Tables

Median

- Reflect the central tendency of the sample in such a way that it is uninfluenced by extreme values or outliners.
- The middle value that separates the higher half from the lower half of the data set.
- To compute the middle value, we need to arrange all the numbers from smallest to greatest.
- Then,

$$\tilde{x} = \begin{cases} x_{\frac{(n+1)}{2}}, & if\ n\ is\ odd, \\ \dfrac{\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)}{2}, & if\ n\ is\ even, \end{cases}$$

Example:

Job performance: 7, 10, 11, 15, 10, 10, 12, 14, 16, 12

Median of job performance:

n = 10. So, n is even

$$\tilde{x} = \frac{x_5 + x_6}{2} = \frac{11 + 12}{2} = 11.5$$

11.5

7	**10**	**10**	**10**	**11**	**12**	**12**	**14**	**15**	**16**
				x_5	x_6				

Mean, Median and Mode

Descriptive Statistics with Pivot Tables

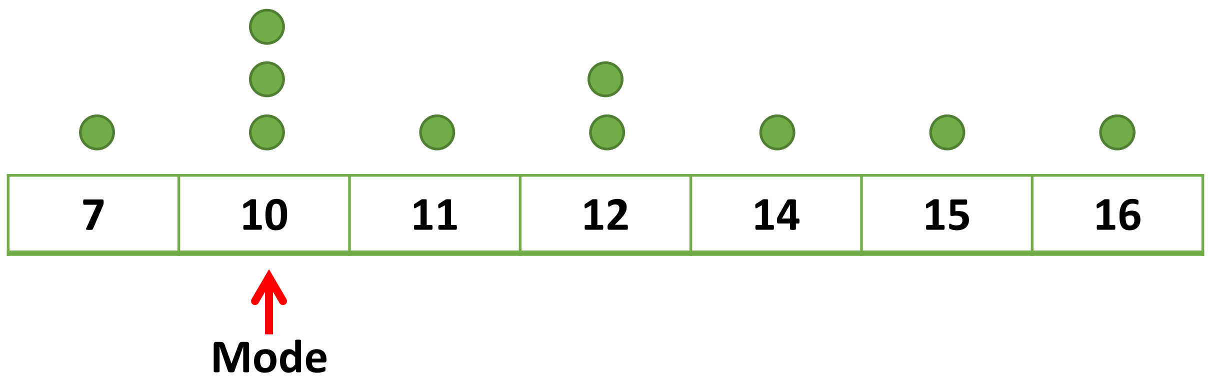
Mode

- The most frequent value in the data set.

Example:

Job performance: 7, 10, 11, 15, 10, 10, 12, 14, 16, 12

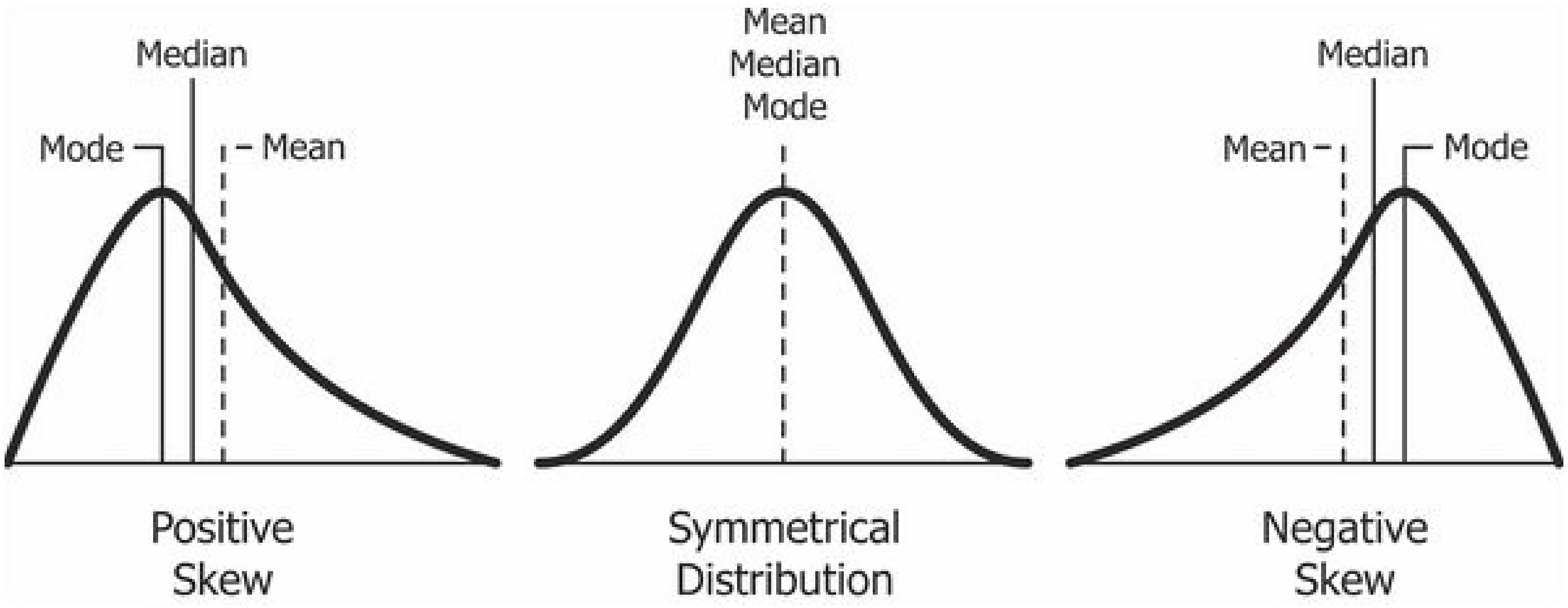
Mode of job performance:



Mean, Median and Mode

Descriptive Statistics with Pivot Tables

Geometric visualization of the mode, median and mean of an arbitrary probability density function



Source: https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa

Mean, Median and Mode

Descriptive Statistics with Pivot Tables

Recall:

Provides	Categorical Attribute		Numerical Attribute	
	Nominal	**Ordinal**	**Interval-scaled**	**Ratio-scaled**
Mode	/	/	/	/
Median		/	/	/
Mean			/	/

Mean, Median and Mode

Descriptive Statistics with Pivot Tables

	IQ X_1	Job performance X_2
$\mathbf{x}_1$	99	7
$\mathbf{x}_2$	105	10
$\mathbf{x}_3$	105	11
$\mathbf{x}_4$	106	15
$\mathbf{x}_5$	108	10
$\mathbf{x}_6$	112	10
$\mathbf{x}_7$	113	12
$\mathbf{x}_8$	115	14
$\mathbf{x}_9$	118	16
$\mathbf{x}_{10}$	134	12
Mean		11.7
Median		11.5
Mode		10

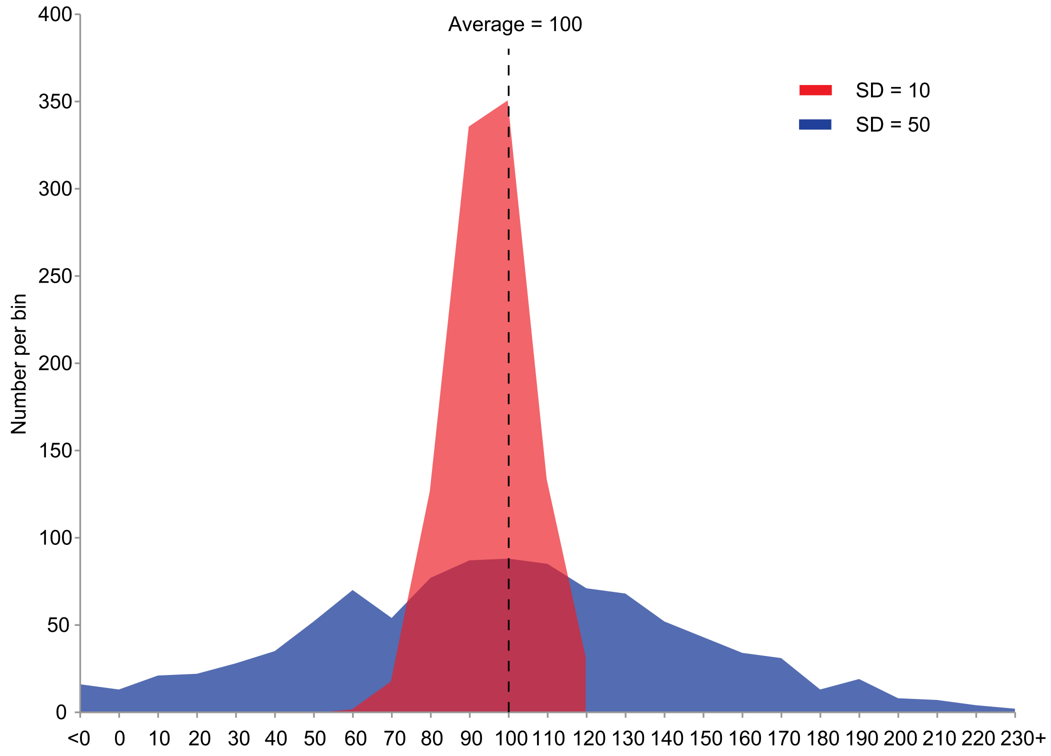
Quiz:
Find the mean, median and mode of IQ.

Variance and Standard Deviation

Descriptive Statistics with Pivot Tables

Standard Deviation (SD, s)

- A measure that is used to quantify the amount of variation or dispersion of a set of data values.
- A low standard deviation indicates that the data points tend to be close to the mean.
- A high standard deviation indicates that the data points are spread out over a wider range of values.



Source:
https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Comparison_standard_deviations.svg

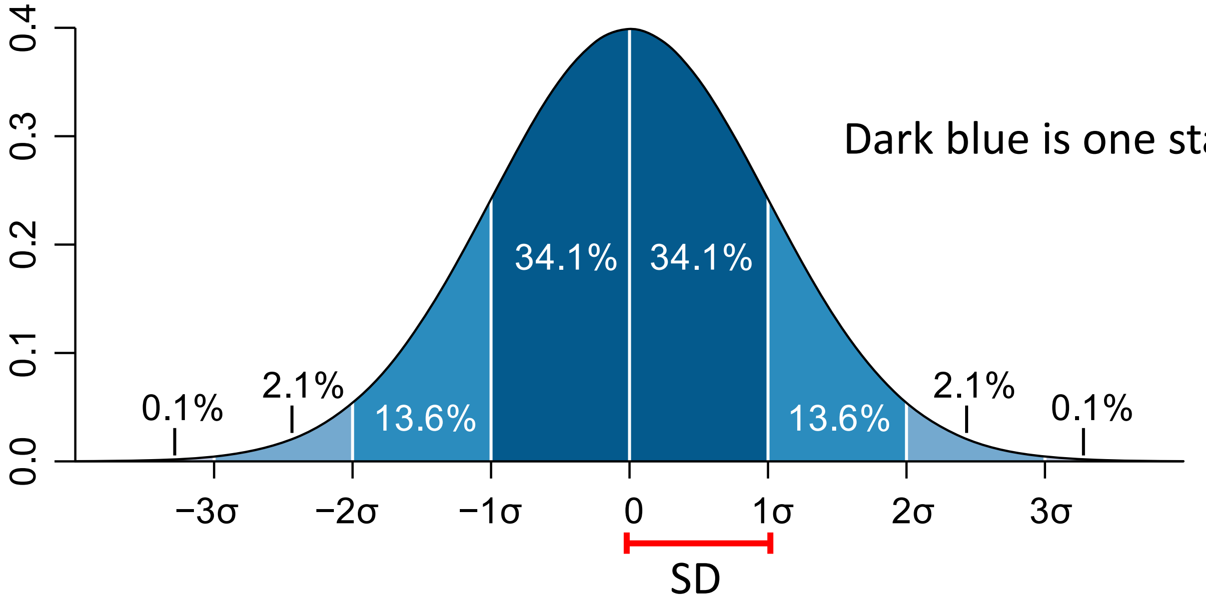
Variance and Standard Deviation

Descriptive Statistics with Pivot Tables

Standard Deviation (SD, s)

The formula for the sample standard deviation is

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$



Dark blue is one standard deviation on either side of the mean.

Source:
https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Standard_deviation_diagram.svg

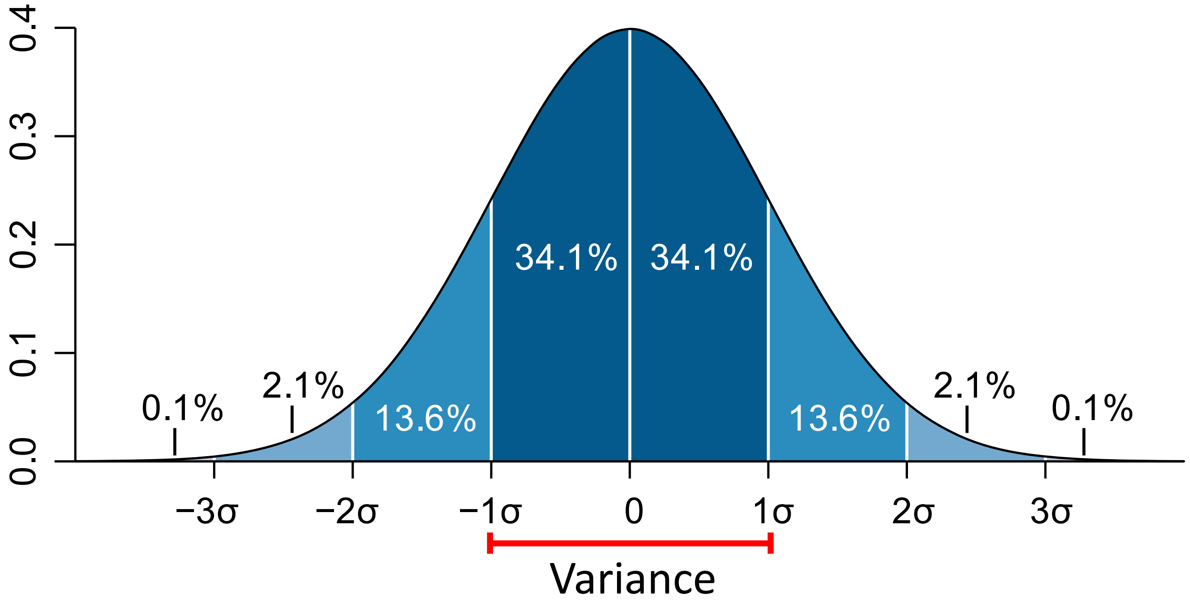
Variance and Standard Deviation

Descriptive Statistics with Pivot Tables

Variance (σ)

- How far a set of numbers are spread out from their average value.
- It is the square of the standard deviation

$$\mathrm{var}(X) = s^2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$



Source:
https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Standard_deviation_diagram.svg

Variance and Standard Deviation

Descriptive Statistics with Pivot Tables

Example

- Job performance; X={7, 10, 11, 15, 10, 10, 12, 14, 16, 12}
- Mean of job performance $\bar{x}$: 11.7
- Standard Deviation; $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 2.71$
- Variance; $\text{var}(X) = SD^2 = 2.71^2 = 7.34$

Job performance x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
7	-4.7	22.09
10	-1.7	2.89
11	-0.7	0.49
15	3.3	10.89
10	-1.7	2.89
10	-1.7	2.89
12	0.3	0.09
14	2.3	5.29
16	4.3	18.49
12	0.3	0.09
$\sum_{i=1}^{n}(x_i - \bar{x})^2$		66.1
$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$		2.71

Variance and Standard Deviation

Descriptive Statistics with Pivot Tables

	IQ X_1	Job performance X_2
$\mathbf{x}_1$	99	7
$\mathbf{x}_2$	105	10
$\mathbf{x}_3$	105	11
$\mathbf{x}_4$	106	15
$\mathbf{x}_5$	108	10
$\mathbf{x}_6$	112	10
$\mathbf{x}_7$	113	12
$\mathbf{x}_8$	115	14
$\mathbf{x}_9$	118	16
$\mathbf{x}_{10}$	134	12
Mean	111.5	11.7
SD		2.71
Variance		7.34

$$\mathrm{var}(X) = s^2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Quiz:

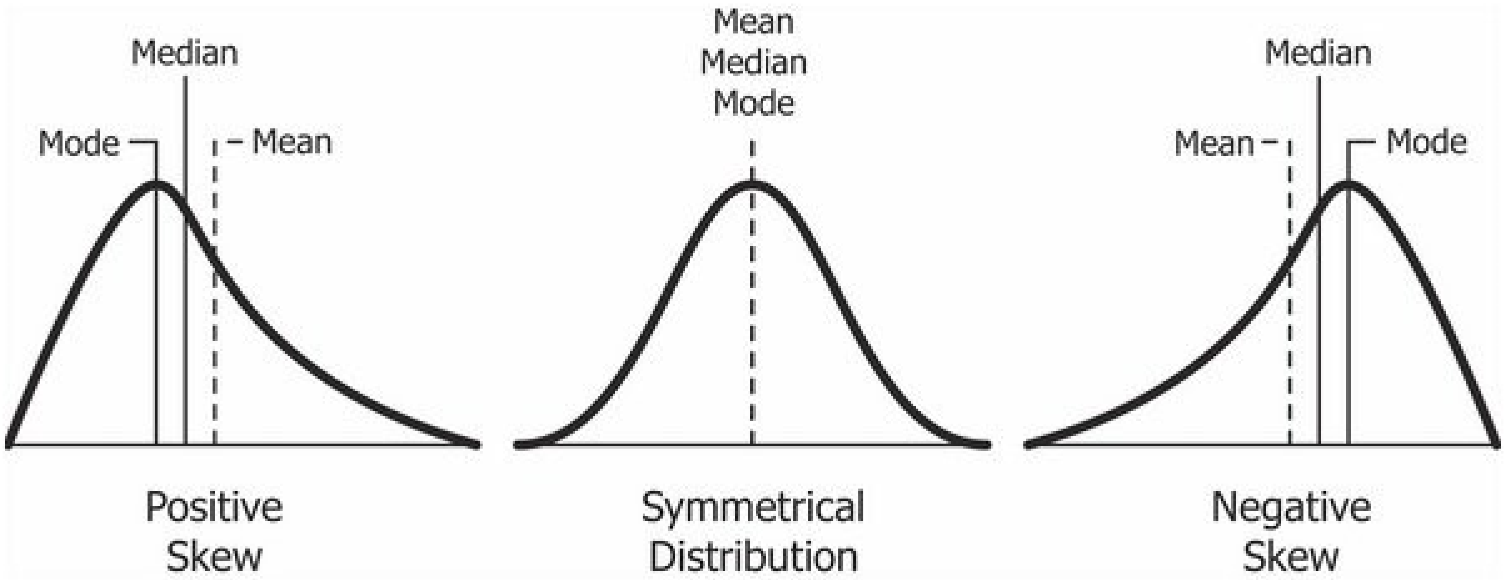
Find the SD and variance of IQ.

Skewness and Kurtosis

Descriptive Statistics with Pivot Tables

Skewness

- Skewness is usually described as a measure of a **dataset's symmetry** – or lack of symmetry.
- The normal distribution has a skewness of 0.



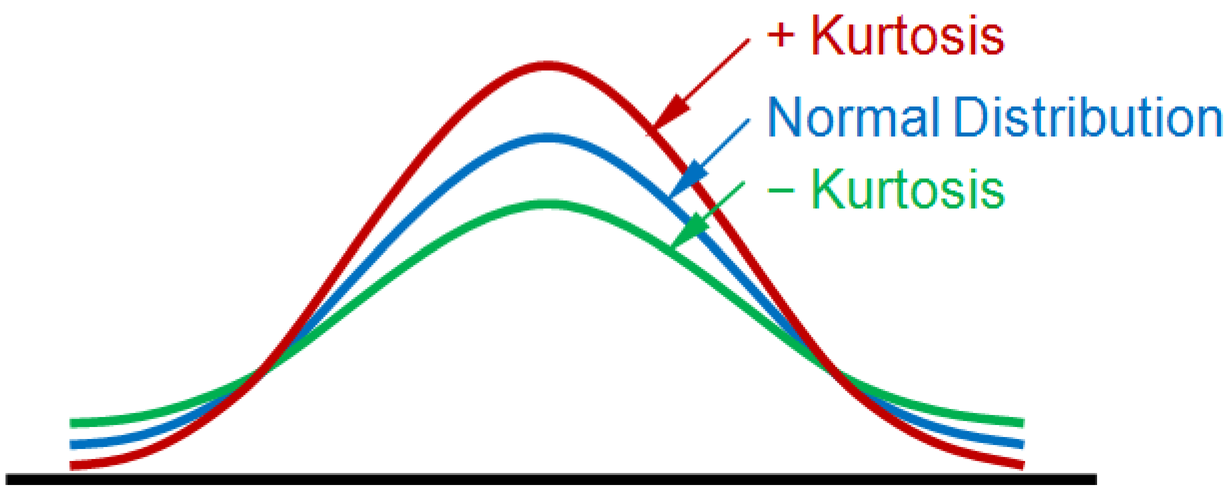
Source: https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa

Skewness and Kurtosis

Descriptive Statistics with Pivot Tables

Kurtosis

- Measures the **tail-heaviness of the distribution**.
- The excess kurtosis for a standard normal distribution is 0.



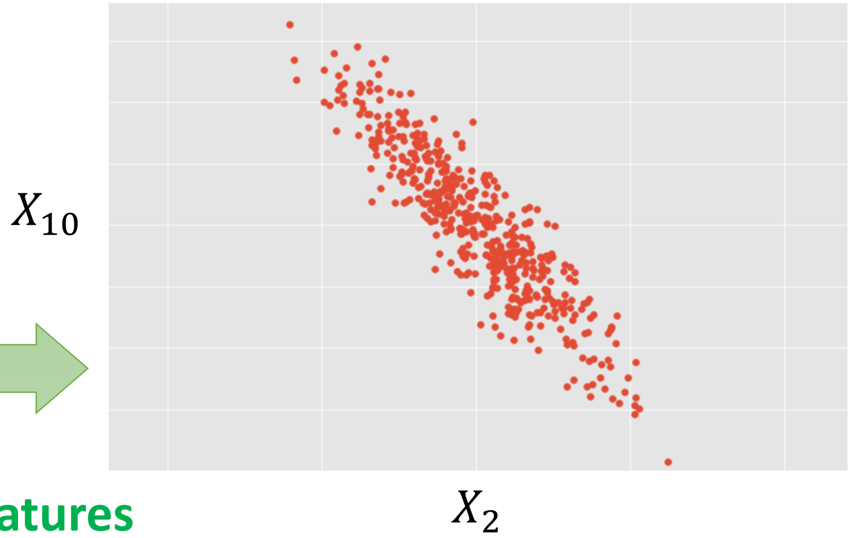
Source: https://www.statext.com/android/kurtosis.html

Covariance Matrix

Descriptive Statistics with Pivot Tables

	X_1	X_2	…	X_{10}
$\mathbf{x}_1$				
…				
$\mathbf{x}_n$				

The joint variability of two random variables can be described by **covariance**



We can slice any variables/features and display them as a scatter plot

Covariance Matrix

Descriptive Statistics with Pivot Tables

Covariance

- How much two random variables vary together.
- The covariance of random variables X and Y, denoted by $\mathrm{cov}(X, Y)$ can be computed by:

$$\mathrm{cov}(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

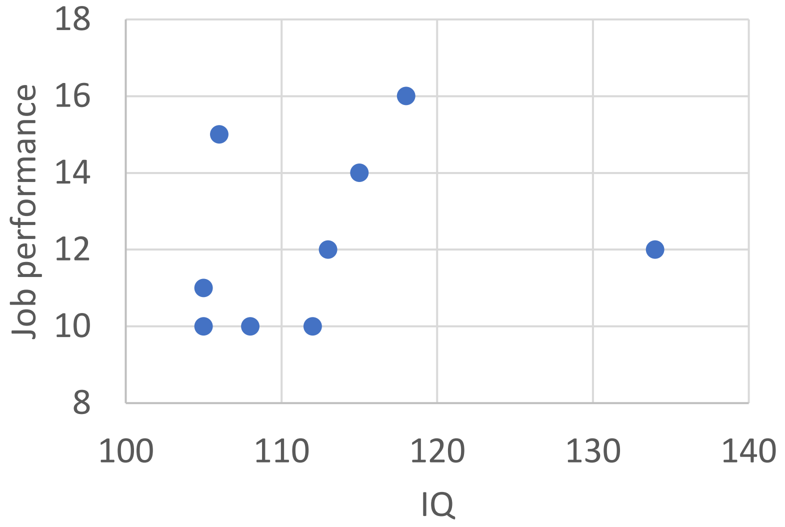
- The value of covariance lies between $-\infty$ and $+\infty$.

Covariance Matrix

Descriptive Statistics with Pivot Tables

Example

	IQ X	Job performance Y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
$\mathbf{x}_1$	99	7	-12.5	-4.7	58.75
$\mathbf{x}_2$	105	10	-6.5	-1.7	11.05
$\mathbf{x}_3$	105	11	-6.5	-0.7	4.55
$\mathbf{x}_4$	106	15	-5.5	3.3	-18.15
$\mathbf{x}_5$	108	10	-3.5	-1.7	5.95
$\mathbf{x}_6$	112	10	0.5	-1.7	-0.85
$\mathbf{x}_7$	113	12	1.5	0.3	0.45
$\mathbf{x}_8$	115	14	3.5	2.3	8.05
$\mathbf{x}_9$	118	16	6.5	4.3	27.95
$\mathbf{x}_{10}$	134	12	22.5	0.3	6.75
Mean	111.5	11.7		SUM	104.5



$$\mathrm{cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

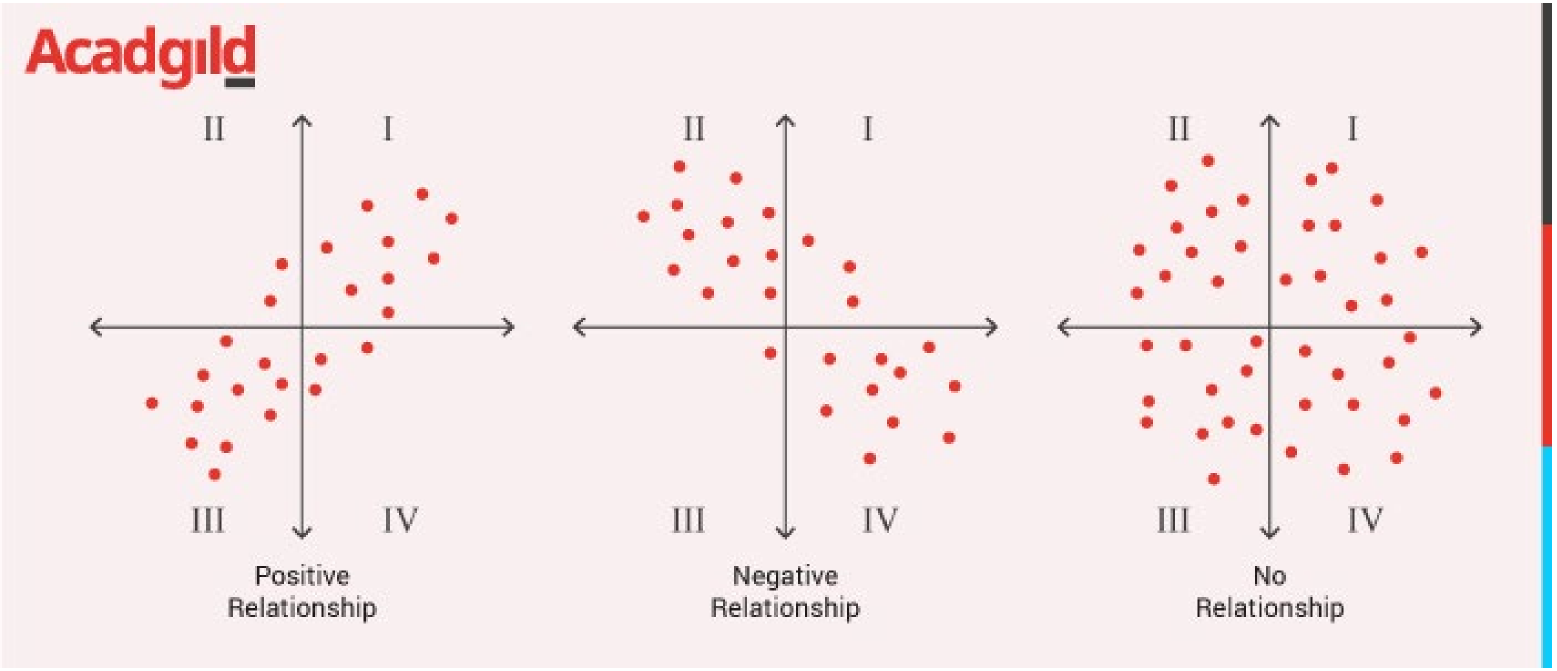
$$\mathrm{cov}(X,Y) = \frac{104.5}{9} = 11.61$$

What dose it mean?

Covariance Matrix

Descriptive Statistics with Pivot Tables

Covariance



Source: https://acadgild.com/blog/covariance-and-correlation

A **positive covariance** means both variables tend to move upward or downward in value at the same time.

A **negative covariance** means the variables will move away from each other.

A **zero covariance** means there is no relationship.

Covariance Matrix

Descriptive Statistics with Pivot Tables

Correlation

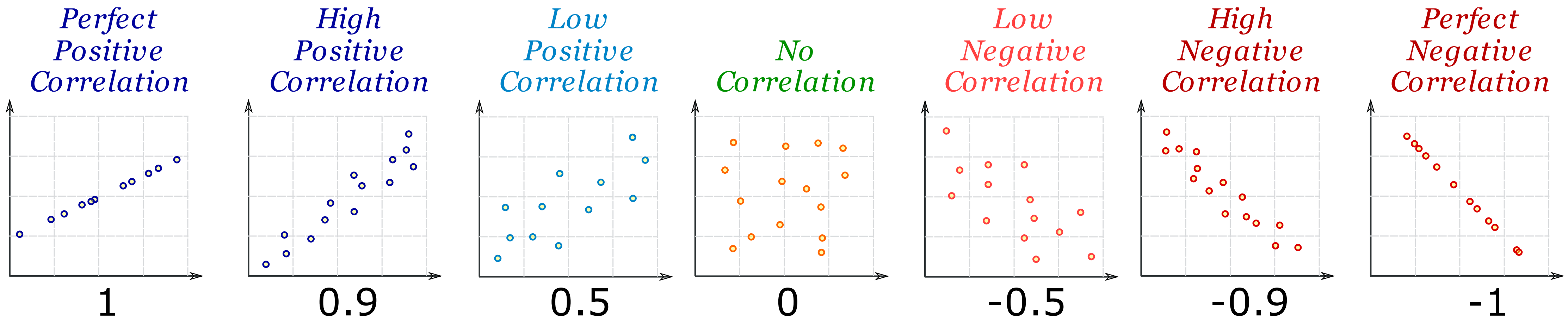
- Unit measure of change between two variables change with respect to each other.
- A normalized form of covariance.

$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{s_X s_Y}$$

- The value of correlation lies between -1 and $+1$.
  - If the correlation coefficient is one, it means that if one variable moves a given amount, the second moves proportionally in the same direction.
  - If correlation coefficient is zero, no relationship exists between the variables.
  - If correlation coefficient is –1, it means that one variable increases, the other variable decreases proportionally.

Covariance Matrix

Descriptive Statistics with Pivot Tables



The value of covariance lies between -1 and $+1$.

- If the correlation coefficient is <u>one</u>, it means that if one variable moves a given amount, the second moves proportionally in the same direction.
- If correlation coefficient is <u>zero</u>, <u>no relationship</u> exists between the variables.
- If correlation coefficient is <u>–1</u>, it means that one variable increases, the other variable decreases proportionally.

Covariance Matrix

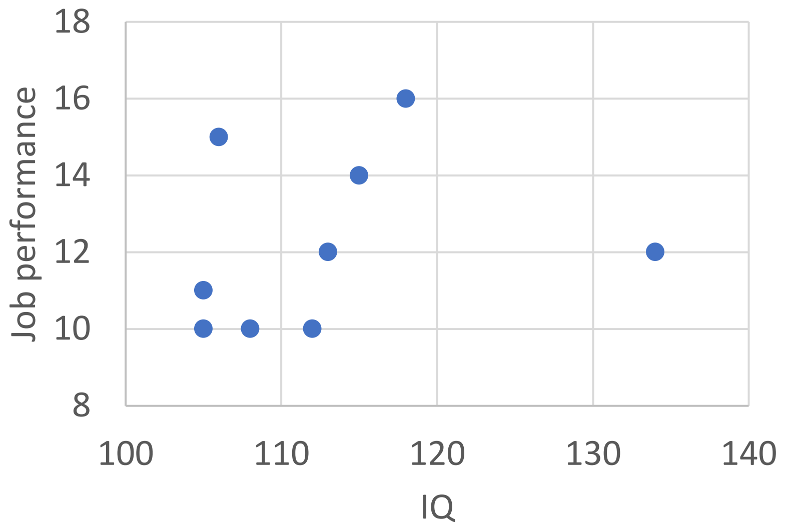
Descriptive Statistics with Pivot Tables

Example

	IQ X	**Job performance** Y
$\mathbf{x}_1$	99	7
$\mathbf{x}_2$	105	10
$\mathbf{x}_3$	105	11
$\mathbf{x}_4$	106	15
$\mathbf{x}_5$	108	10
$\mathbf{x}_6$	112	10
$\mathbf{x}_7$	113	12
$\mathbf{x}_8$	115	14
$\mathbf{x}_9$	118	16
$\mathbf{x}_{10}$	134	12
Mean	111.5	11.7
SD	9.70	2.71

$$\mathrm{cov}(X, Y) = 11.61$$

$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{s_X s_Y} = \frac{11.61}{9.70 \times 2.71} = \frac{11.61}{26.287} = 0.44$$

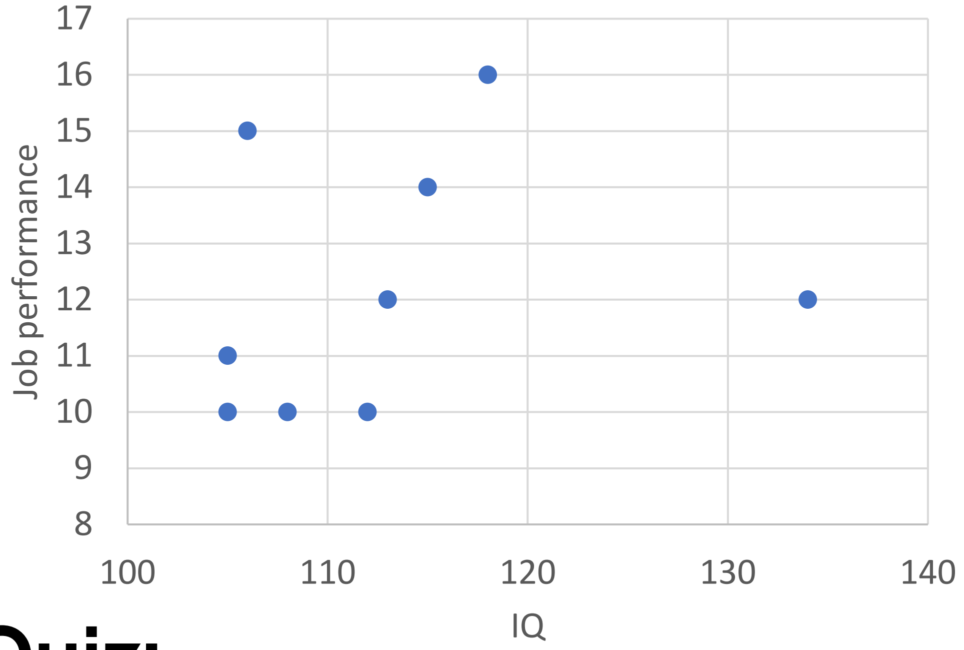


Covariance Matrix

Descriptive Statistics with Pivot Tables

Example

	IQ X	Job performance Y
$\mathbf{x}_1$	99	7
$\mathbf{x}_2$	105	10
$\mathbf{x}_3$	105	11
$\mathbf{x}_4$	106	15
$\mathbf{x}_5$	108	10
$\mathbf{x}_6$	112	10
$\mathbf{x}_7$	113	12
$\mathbf{x}_8$	115	14
$\mathbf{x}_9$	118	16
$\mathbf{x}_{10}$	134	12
Mean	111.5	11.7
SD	9.70	2.71



$$\text{cov}(X, Y) = 11.61$$

$$\text{corr}(X, Y) = 0.44$$

Quiz:
What do the covariance and correlation tell about the relation between IQ and job performance?

Covariance Matrix

Descriptive Statistics with Pivot Tables

Covariance Matrix

- A matrix whose element in the *i, j* position is the covariance between the *i*-th and *j*-th features.

	X_1	X_2	…	X_{10}
$\mathbf{x}_1$				
…				
$\mathbf{x}_n$				

Data Matrix

$$C = \begin{array}{c} \\ X_1 \\ X_2 \\ \\ X_{10} \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & & X_{10} \end{array} \\ \begin{bmatrix} \operatorname{cov}(X_1, X_1) & \operatorname{cov}(X_1, X_2) & \cdots & \operatorname{cov}(X_1, X_{10}) \\ \operatorname{cov}(X_2, X_1) & \operatorname{cov}(X_2, X_2) & \cdots & \operatorname{cov}(X_2, X_{10}) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_{10}, X_1) & \operatorname{cov}(X_{10}, X_2) & \cdots & \operatorname{cov}(X_{10}, X_{10}) \end{bmatrix} \end{array}$$

Covariance Matrix

Practice

Problem

- จงวาดภาพลักษณะการกระจายของข้อมูลตัวแปร sepal length พร้อมระบุ ตำแหน่งของค่า mean, median และ mode ในภาพด้วย

- ถ้าค่าสหสัมพันธ์ (Correlation) ระหว่างตัวแปร petal length และ petal width มีค่าเท่ากับ 0.96 จงบอกลักษณะความสัมพันธ์ร่วมระหว่าง 2 ตัวแปรนี้ (ไม่มีความสัมพันธ์กัน, มีความสัมพันธ์กันในเชิงลบ, มีความสัมพันธ์กันในเชิงบวก)

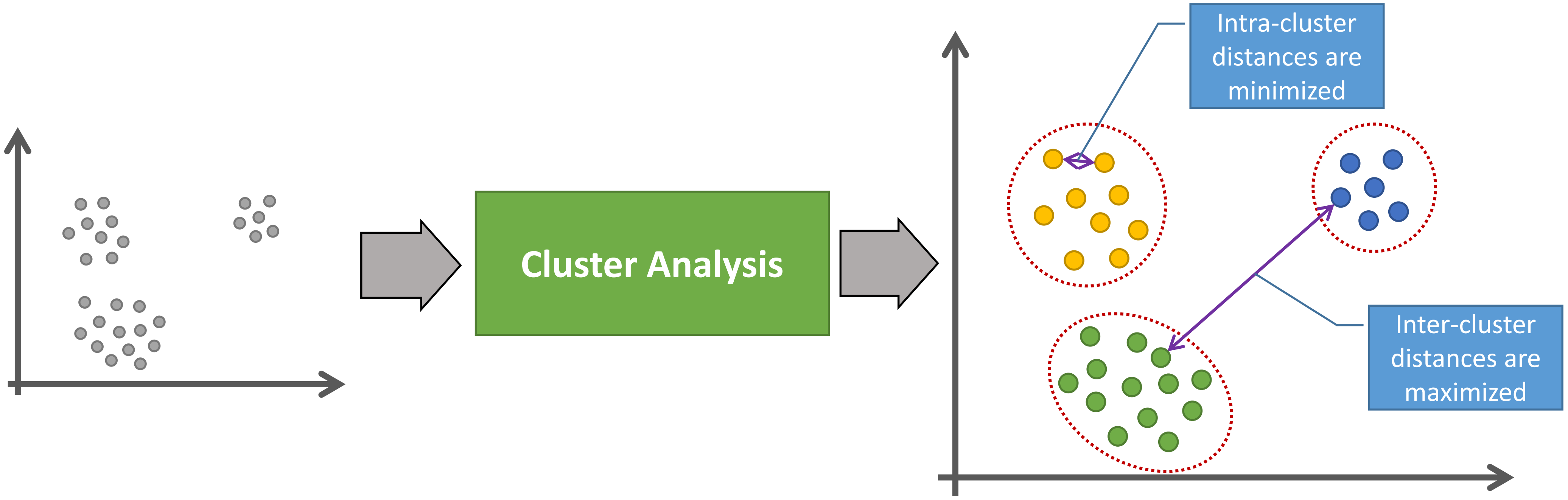
Variable / Value	sepal length	sepal width	petal length	petal width
Mean	5.84	3.05	3.76	1.20
Median	5.80	3.00	4.35	1.30
Mode	5.00	3.00	1.50	0.20
S.D.	0.83	0.43	1.76	0.76
Skewness	0.31	0.33	- 0.27	- 0.10
Kurtosis	- 0.55	0.29	- 1.40	- 1.34

Cluster Analysis

Cluster Analysis

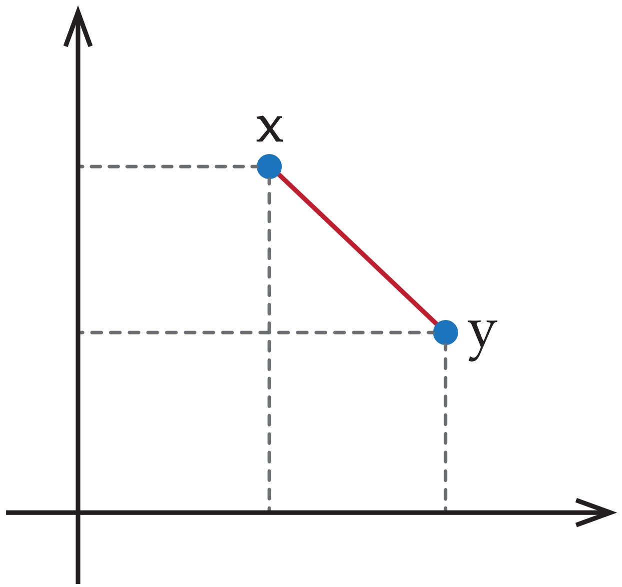
Finding groups of datapoints such that:

- The datapoints in the same group will be like one another.
- The datapoints in a group are different from the datapoints in other groups.
- The group of similar data points is called a **Cluster.**



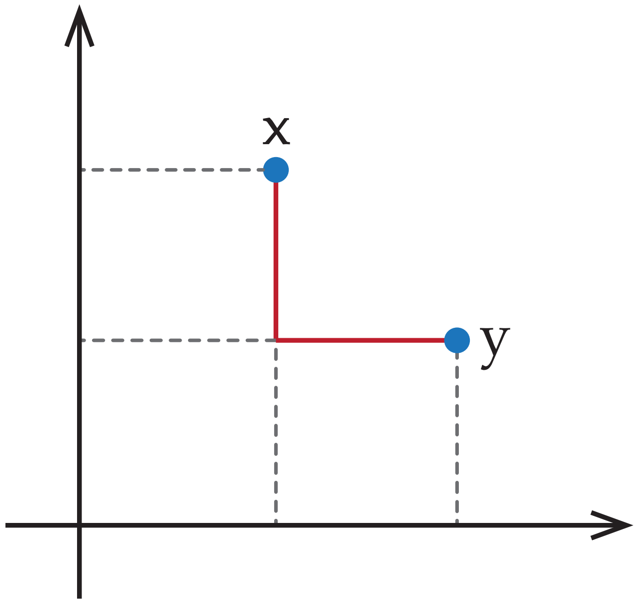
Distances and Similarity

Cluster Analysis



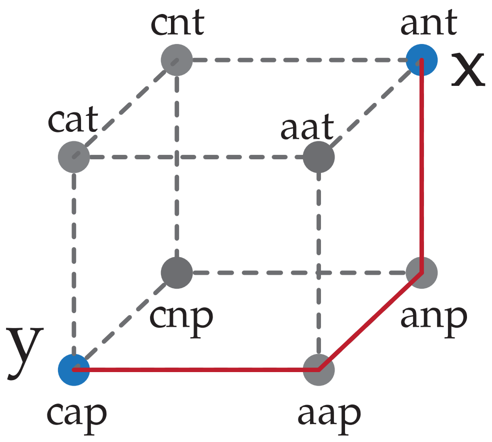
Euclidean distance

$$d_{euc}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$



Manhattan distance

$$d_{manh}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} |x_i - y_i|$$



Hamming distance

$$d_{hamm}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} (x_i \neq y_i)$$

The number of mismatched values

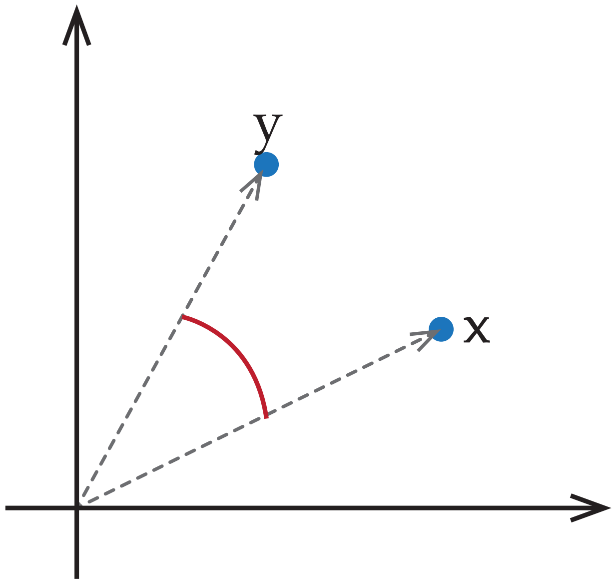
Commonly used to measure distance between two numerical datapoints.

Commonly used for categorical datapoints.

If it is 0, it means that both objects are identical.

Distances and Similarity

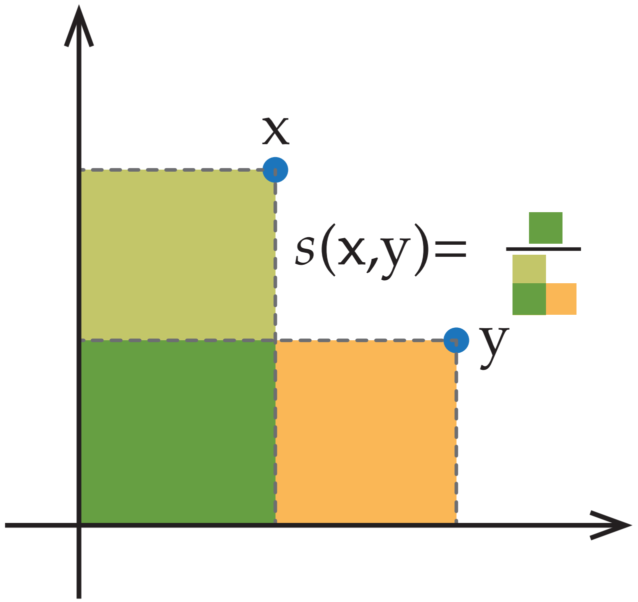
Cluster Analysis



Cosine similarity

$$s_{cos}(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{p} x_i y_i}{\sqrt{\sum_{i=1}^{p} x_i^2}\sqrt{\sum_{i=1}^{y} y_i^2}}$$

Commonly used for numerical datapoints.



Jaccard coefficient

$$s_{jacc}(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{p} \min(x_i, y_i)}{\sum_{i=1}^{p} \max(x_i, y_i)}$$

Commonly used for categorical datapoints.

The range of score varies between 0 and 1.
If score is 1, it means that they are same.

Distances and Similarity

Cluster Analysis

	X_1	X_2	…	X_{10}
$\mathbf{x}_1$				
…				
$\mathbf{x}_n$				

Data Matrix

We can represent all pair datapoints as a **distance matrix**

Element in the *i, j* position is the distance between the *i*-th and *j*-th datapoints.

$$D = \begin{array}{c} \\ \mathbf{x}_1 \\ \mathbf{x}_2 \\ \\ \mathbf{x}_n \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & & \mathbf{x}_n \end{array} \\ \begin{bmatrix} d(\mathbf{x}_1, \mathbf{x}_1) & d(\mathbf{x}_1, \mathbf{x}_2) & \cdots & d(\mathbf{x}_1, \mathbf{x}_n) \\ d(\mathbf{x}_2, \mathbf{x}_1) & d(\mathbf{x}_2, \mathbf{x}_2) & \cdots & d(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ d(\mathbf{x}_n, \mathbf{x}_1) & d(\mathbf{x}_n, \mathbf{x}_2) & \cdots & d(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix} \end{array}$$

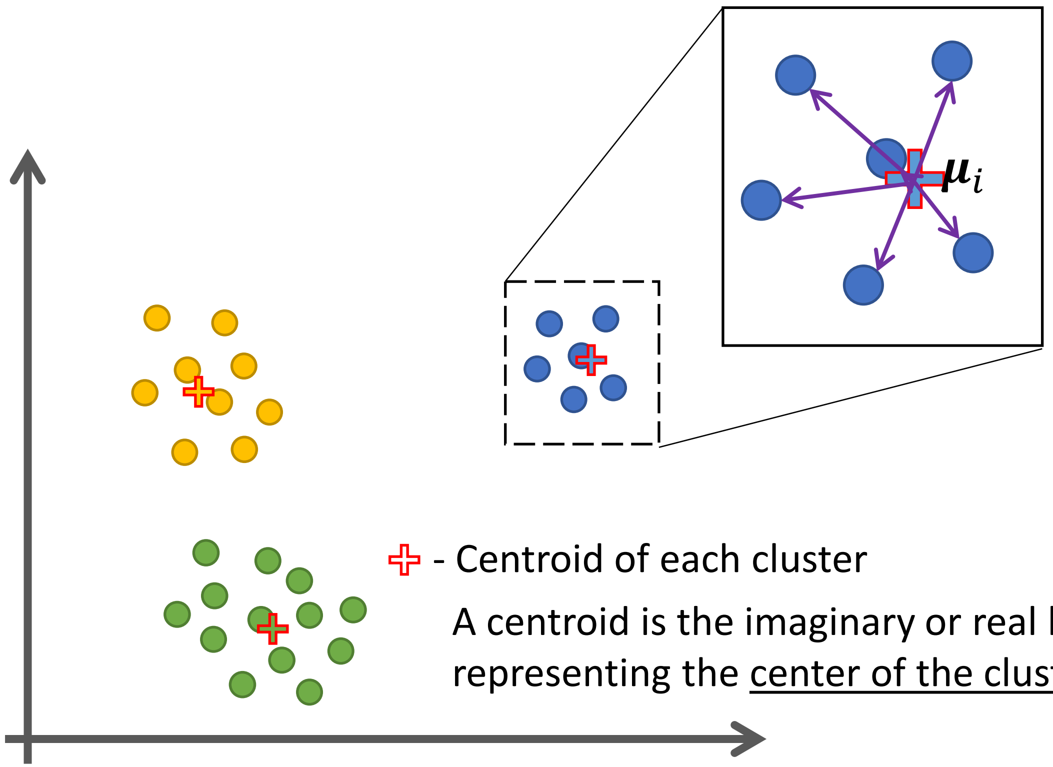
Distance Matrix

K-means Clustering

Cluster Analysis

K-means

Every data point is allocated to each of the clusters through reducing the sum of squared error.



A centroid is the imaginary or real location representing the center of the cluster.

Intra-cluster sum of squared error for a cluster:

$$\sum_{\mathbf{x}_j \in C_i} d(\mathbf{x}_j, \boldsymbol{\mu}_i)^2$$

C_i - set of datapoints in cluster j

Sum of squared error:

$$\sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} d(\mathbf{x}_j, \boldsymbol{\mu}_i)^2$$

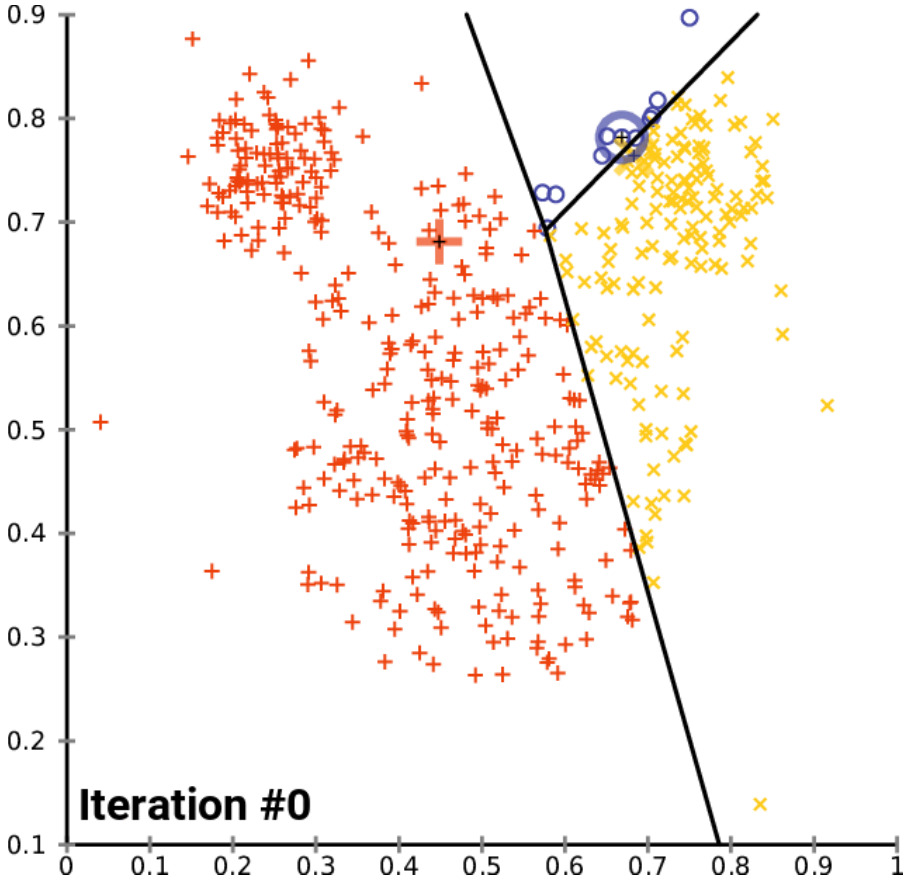
k – number of clusters

K-means Clustering

Cluster Analysis

How the k-means works

STEP 1: Identifies *k* number of centroids
(k is a parameter of the k-means)
STEP 2: Randomly initialize *k* centroids
STEP 3: Allocates every data point to the nearest cluster
STEP 4: Update each centroid (mean)
STEP 5: Go to STEP 3 until centroids have stabilized



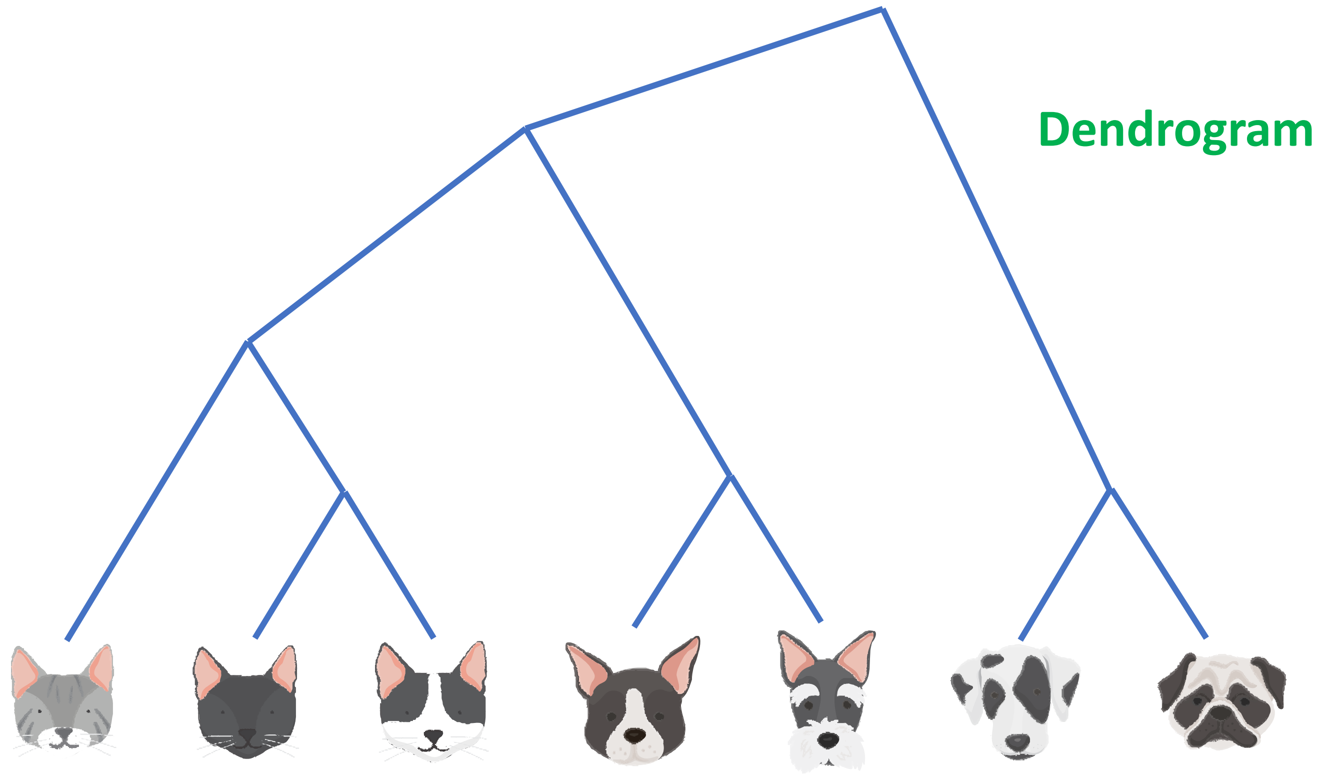
Source:
https://commons.wikimedia.org/wiki/File:K-means_convergence.gif

Hierarchical Clustering

Cluster Analysis

Agglomerative Hierarchical clustering

Iteratively merge the two closest clusters until only a single cluster remains.



Hierarchical Clustering

Cluster Analysis

How the agglomerative hierarchical clustering works

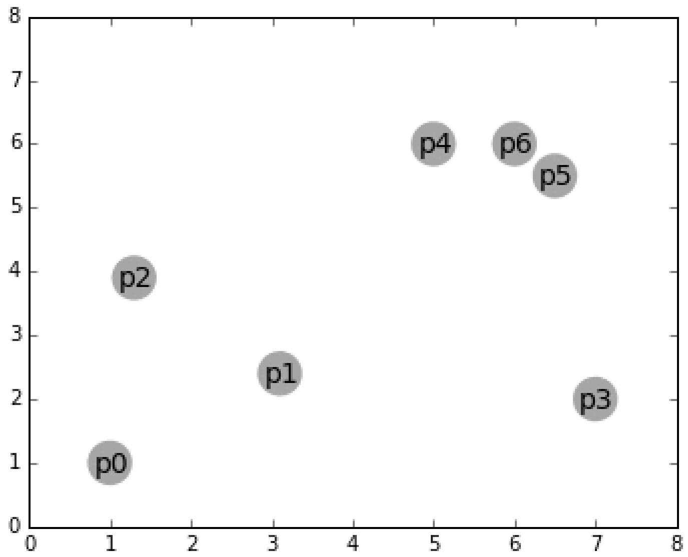
STEP 1: Compute the proximity matrix (distance or similarity matrix)

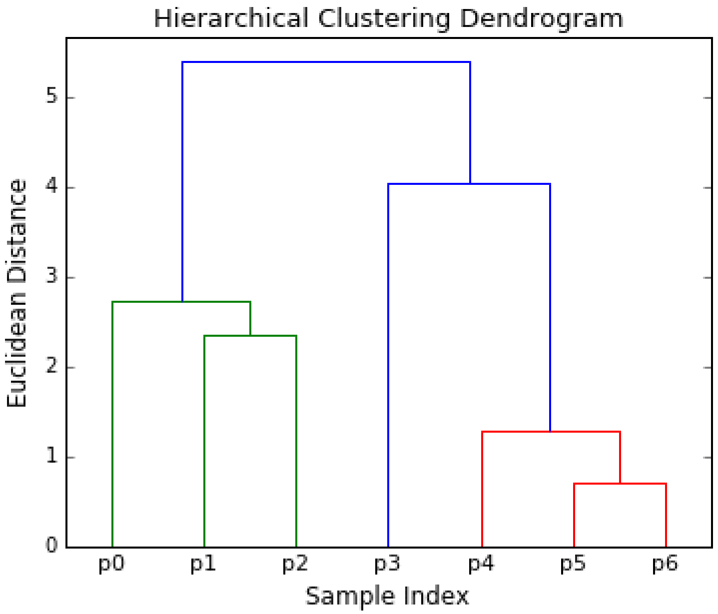
STEP 2: Let each data point be a cluster

STEP 3: Merge the two closest clusters

STEP 4: Update the proximity matrix

STEP 5: Go to STEP 3 until only a single cluster remains





Source:
https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

Hierarchical Clustering

Cluster Analysis

Agglomerative hierarchical clustering

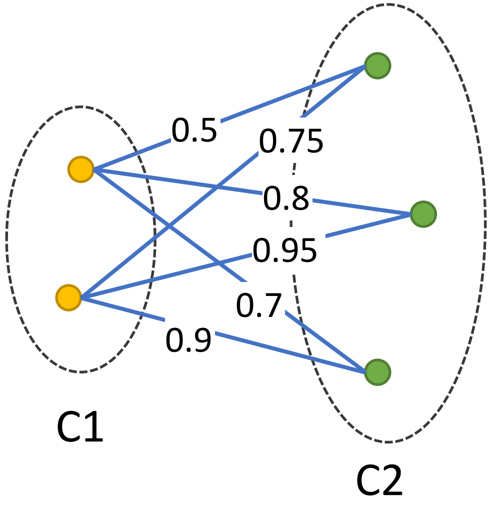
STEP 1: Compute the proximity matrix
STEP 2: Let each data point be a cluster
STEP 3: Merge the two closest clusters
STEP 4: Update the proximity matrix
STEP 5: Go to STEP 3 until only a single cluster remains

As we merge datapoints to form a cluster (set of datapoints)

How can we measure the distance/similarity between two sets?

Linkage Criteria: Distance between sets of observations

1. Minimum of the distance between points x_i and x_j such that x_i belongs to C1 and x_j belongs to C2
2. Maximum of the distance between points x_i and x_j such that x_i belongs to C1 and x_j belongs to C2
3. Average distance of all-pair data points
4. Distance Between Centroids
5. and etc.



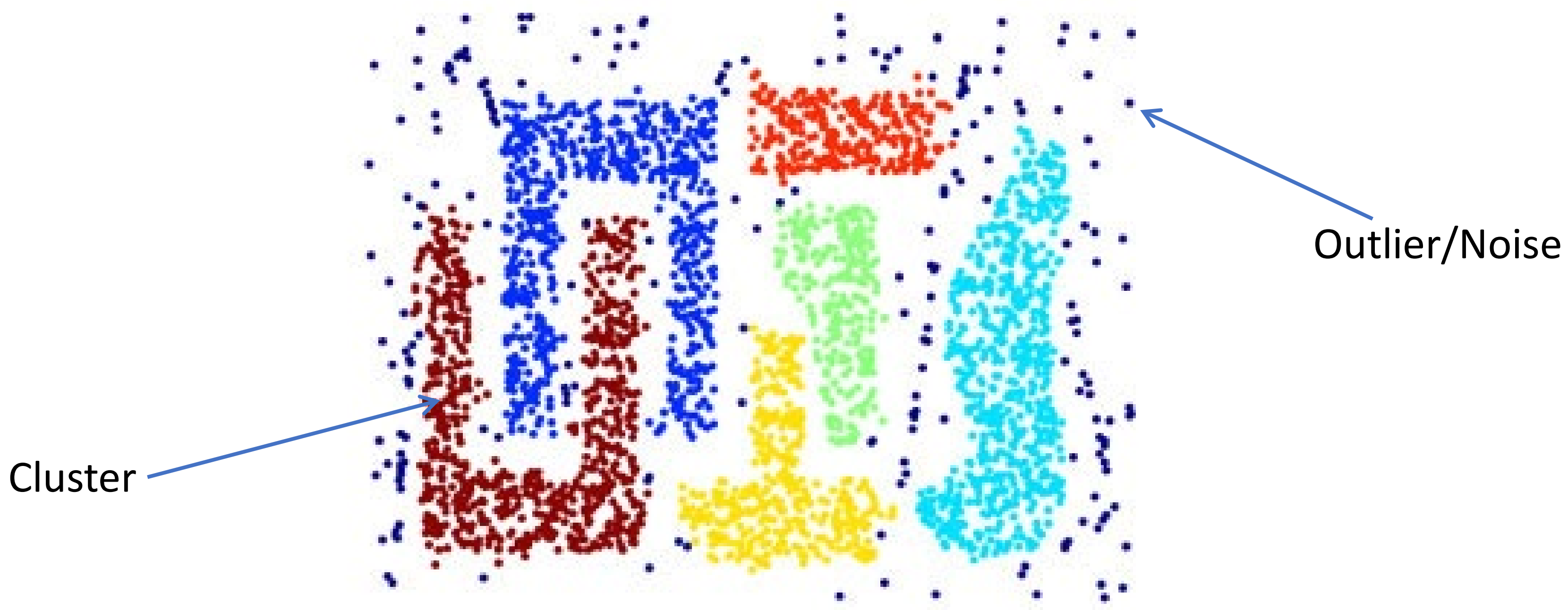
Minimum (single-linkage clustering): 0.5
Maximum (complete-linkage clustering): 0.95
Average linkage clustering: 0.77

Density-based Spatial Clustering

Cluster Analysis

Use the local density of points to determine the clusters.

- Groups together points that are closely packed together (point in high-density regions).
- Marking points that lie alone in low-density regions as outliers.



Density-based Spatial Clustering

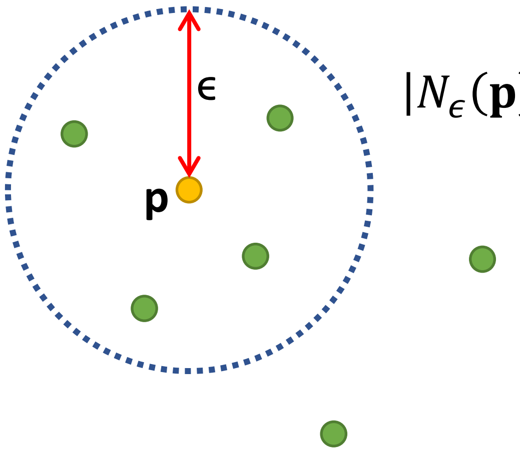
Cluster Analysis

How do we measure density of a region?

- **Density at a point** - Number of points within a circle of Radius *Eps (ϵ)* from point **p**.

  ϵ-neighborhood: $N_\epsilon(\mathbf{p}) = \{\mathbf{q} \in \mathbf{D} | \ d(\mathbf{p}, \mathbf{q}) \leq \ \epsilon\}$

- **Dense Region -** For each point in the cluster, the circle with radius ϵ contains at least minimum number of points (*MinPts*).



$|N_\epsilon(\mathbf{p})| = 4$

Density-based Spatial Clustering

Cluster Analysis

How do we measure density of a region?

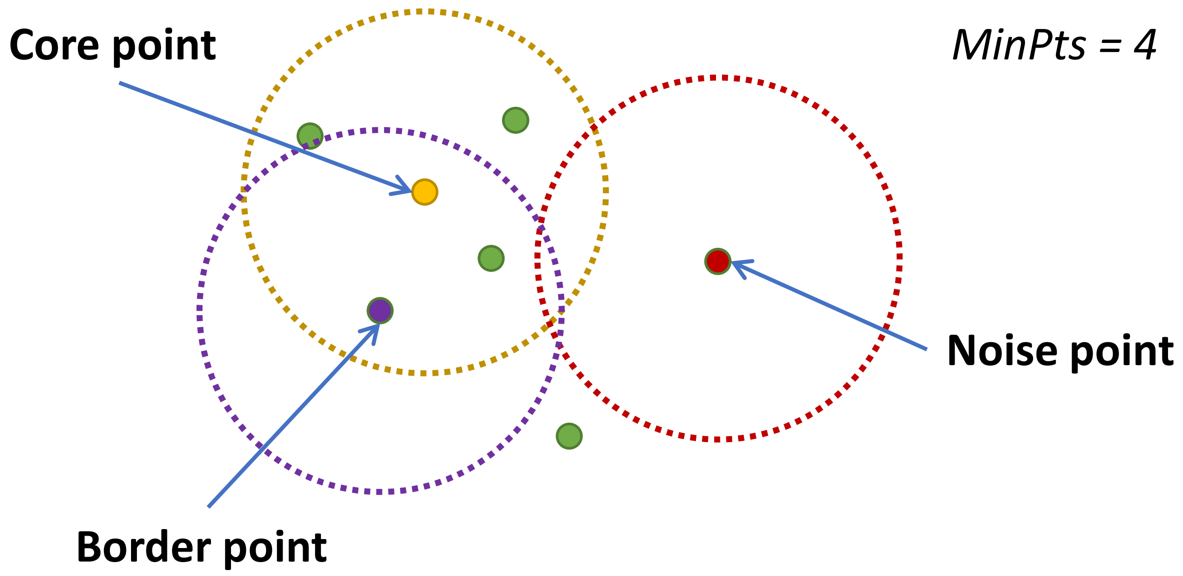
- **Density at a point** - Number of points within a circle of Radius *Eps (ϵ)* from point **p**.

  ϵ-neighborhood: $N_\epsilon(\mathbf{p}) = \{\mathbf{q} \in \mathbf{D} |\ d(\mathbf{p}, \mathbf{q}) \leq \epsilon\}$

- **Dense Region -** For each point in the cluster, the circle with radius ϵ contains at least minimum number of points (*MinPts*).

A point p can be classified as:

- **Core point** – if $|N_\epsilon(\mathbf{p})| \geq MinPts$
- **Border point** – if $|N_\epsilon(\mathbf{p})| < MinPts$ and **p** belong to ϵ-neighborhood of some core point
- **Noise point** – if **p** is neither a core nor a border point



Density-based Spatial Clustering

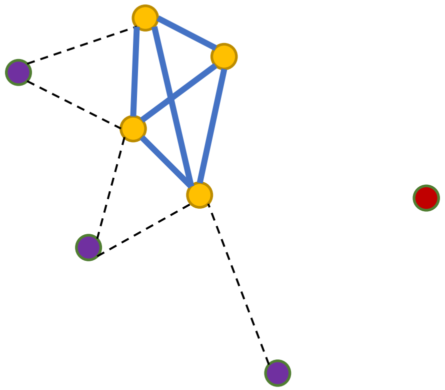
Cluster Analysis

How the DBSCAN works

STEP 1: Find ϵ-neighborhood of every point, and identify the core points

STEP 2: Find the connected components of core points on the neighbor graph, ignoring all non-core points.

STEP 3: Assign each non-core point to a nearby cluster if the cluster is an ϵ - neighbor, otherwise assign it to noise.



MinPts = 4

core points

Connected Components -
There exists an edge between two core points

Practice

Problem

A) k-mean B) Hierarchical Clustering C) DBSCAN

จงจับคู่วิธีการวิเคราะห์กลุ่มข้อมูล (ด้านบน) ที่เกี่ยวข้องกับข้อความต่อไปนี้ (อาจมีมากกว่า 1 ตัวเลือก)

1) การรวมข้อมูลเป็นกลุ่มจากกลุ่มที่มีขนาดเล็กเป็นกลุ่มที่มีขนาดใหญ่ขึ้นจนกระทั่งได้จำนวนกลุ่มข้อมูลตามที่ต้องการ

2) จัดข้อมูลที่อยู่ในบริเวณที่ความหนาแน่นของข้อมูลบริเวณเดียวกันให้อยู่ในกลุ่มข้อมูลเดียวกัน

3) แบ่งกลุ่มข้อมูลที่ทำให้ผลรวมระยะห่างระหว่างข้อมูลในกลุ่มข้อมูลเดียวกันมีค่าน้อยที่สุด

4) ต้องกำหนดจำนวนกลุ่มข้อมูลที่ต้องการเป็นพารามิเตอร์

5) ต้องกำหนดรัศมีของจุดข้อมูล เพื่อคำนวณหาความหนาแน่นของข้อมูล ณ จุดข้อมูล แต่ละจุด

6) จุดข้อมูลบางจุดอาจถูกจัดเข้ากลุ่มข้อมูลใดเลย แต่ถูกระบุเป็น Outliers

Association Analysis

Association Analysis

Uncover associations between items (attributes)

- How likely are two sets of items to **co-occur**.
- How likely are two sets of items to **conditionally occur**.

Frequent Item Sets: (Milk, Bread), (Banana, Apple)

Association Rules: (Bread → Milk)

A prototypical application of association analysis is **Market Basket Analysis**

Association Analysis

	Banana	*Milk*	…	*Bread*
$\mathbf{x}_1$				
…				
$\mathbf{x}_n$				

Market baskets

Frequent Item Sets

Association Analysis

Items

All possible things that can be put into the basket

Example:

Items $I = \{Banana, Milk, Apple, Bread\}$

Items

	$Banana$	$Milk$	$Apple$	$Bread$
$\mathbf{x}_1$	0	1	1	0
$\mathbf{x}_2$	1	1	0	0
$\mathbf{x}_3$	0	1	0	1
…				
$\mathbf{x}_n$	1	0	1	0

Market baskets

Item Set

- A possible combinations of elements in the baskets
- Possible things that can be bought together

For example: 15 possible item sets

$\{Banana\}, \{Milk\}, \{Apple\}, \{Bread\}$

$\{Banana, Milk\}, \{Banana, Apple\}, \{Banana, Bread\}, \{Milk, Apple\}, \{Milk, Bread\}, \{Apple, Bread\}$

$\{Banana, Milk, Apple\}, \{Banana, Milk, Bread\}, \{Banana, Apple, Bread\}, \{Milk, Apple, Bread\}$

$\{Banana, Milk, Apple, Bread\}$

Frequent Item Sets

Association Analysis

Support

- an indication of how frequently the itemset appears in the dataset.
- The proportion of transactions in the dataset $\mathbf{D}$ that contain an item set X, denoted $sup(X, \mathbf{D})$

Example

$$sup(\{Milk\}, \mathbf{D}) = \frac{7}{10} = 0.7$$

$$sup(\{Banana, Apple\}, \mathbf{D}) = \frac{2}{10} = 0.2$$

$$sup(\{Milk, Apple, Bread\}, \mathbf{D}) = \frac{2}{10} = 0.2$$

Items

$\mathbf{D}$	*Banana*	*Milk*	*Apple*	*Bread*
$\mathbf{x}_1$	0	1	1	0
$\mathbf{x}_2$	1	1	0	0
$\mathbf{x}_3$	0	1	0	1
$\mathbf{x}_4$	1	0	1	0
$\mathbf{x}_5$	0	1	1	1
$\mathbf{x}_6$	1	1	0	1
$\mathbf{x}_7$	0	1	1	1
$\mathbf{x}_8$	0	0	1	0
$\mathbf{x}_9$	0	1	0	1
$\mathbf{x}_{10}$	1	0	1	1

Transection

Market baskets

Frequent Item Sets

Association Analysis

An item set X is said to be frequent in D if

$$sup(X, \mathrm{D}) \geq minsup$$

where $\boldsymbol{minsup}$ is a user defined *minimum support threshold*

sup	Item Set
7	$\{Milk\}$
6	$\{Apple\}$ and $\{Bread\}$
5	$\{Milk, Bread\}$
4	$\{Banana\}$
3	$\{Milk, Apple\}$ and $\{Apple, Bread\}$
2	$\{Banana, Milk\}$ and $\{Banana, Apple\}$ and $\{Banana, Bread\}$ and $\{Milk, Apple, Bread\}$
1	$\{Banana, Milk, Bread\}$ and $\{Banana, Apple, Bread\}$

Frequent Item Sets

$minsup$ =0.3

Items

	$Banana$	$Milk$	$Apple$	$Bread$
$\mathbf{x}_1$	0	1	1	0
$\mathbf{x}_2$	1	1	0	0
$\mathbf{x}_3$	0	1	0	1
$\mathbf{x}_4$	1	0	1	0
$\mathbf{x}_5$	0	1	1	1
$\mathbf{x}_6$	1	1	0	1
$\mathbf{x}_7$	0	1	1	1
$\mathbf{x}_8$	0	0	1	0
$\mathbf{x}_9$	0	1	0	1
$\mathbf{x}_{10}$	1	0	1	1

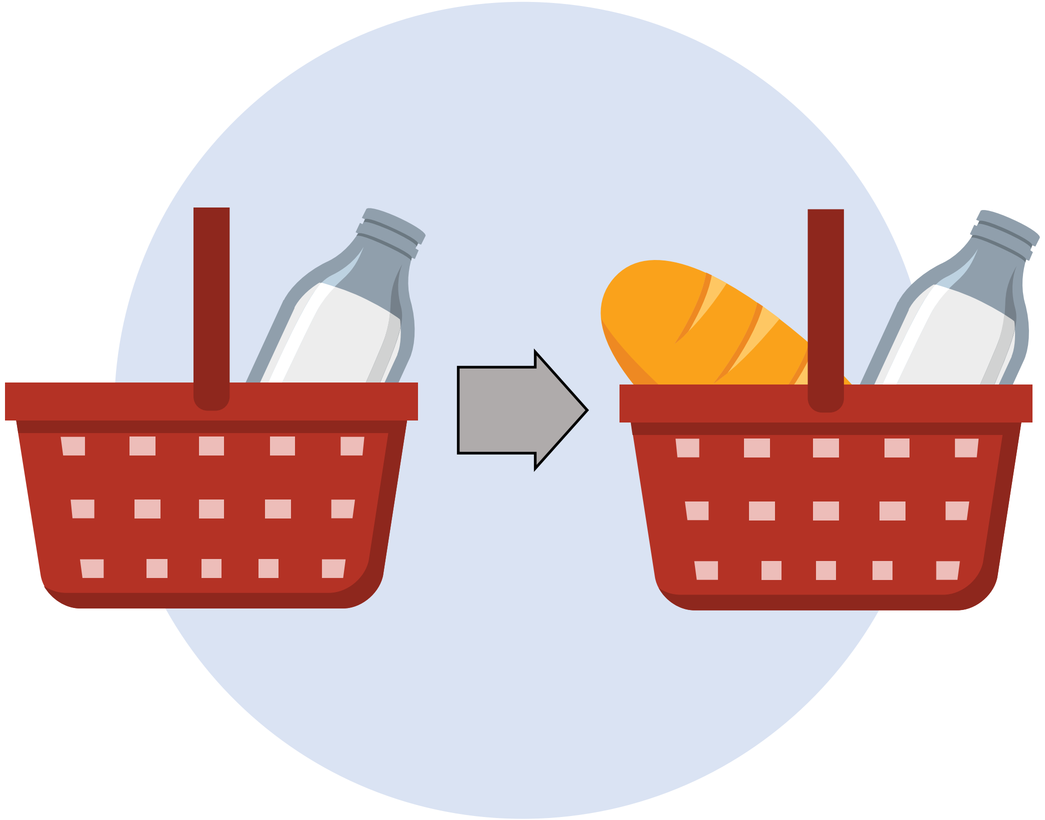
Market baskets

Association Rules

Association Analysis

Association Rule

- An expression $X \rightarrow Y$ where X and Y are item sets and they are disjoint.
- The customer has purchased items in the set X then he is likely to purchase items in the set Y.



Example

$$\{Milk\} \rightarrow \{Bread\}$$

The customer has purchased *milk* then he is likely to purchase *bread*.

Please note that association rules are not commutative, i.e. $\{Milk\} \rightarrow \{Bread\}$ does not equal $\{Bread\} \rightarrow \{Milk\}$.

Association Rules

Association Analysis

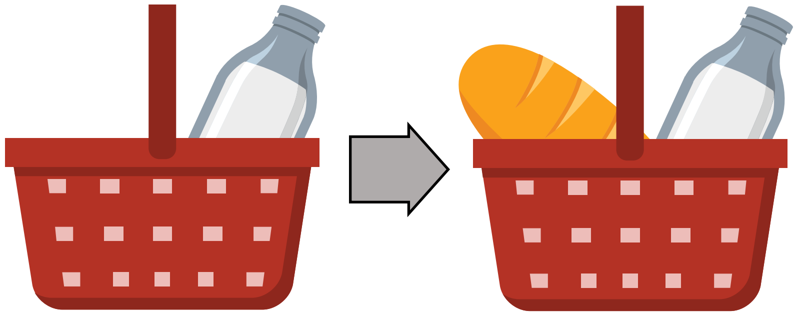
Support of Association Rule

- The number of transaction in which both X and Y co-occur as subsets, where X and Y are item sets

$$sup(X \rightarrow Y) = sup(X \cup Y)$$

Example

$$sup(\{Milk\} \rightarrow \{Bread\}) = sup(\{Milk, Bread\}) = \frac{5}{10} = 0.5$$



Items

	$Banana$	$Milk$	$Apple$	$Bread$
$\mathbf{x}_1$	0	1	1	0
$\mathbf{x}_2$	1	1	0	0
$\mathbf{x}_3$	0	1	0	1
$\mathbf{x}_4$	1	0	1	0
$\mathbf{x}_5$	0	1	1	1
$\mathbf{x}_6$	1	1	0	1
$\mathbf{x}_7$	0	1	1	1
$\mathbf{x}_8$	0	0	1	0
$\mathbf{x}_9$	0	1	0	1
$\mathbf{x}_{10}$	1	0	1	1

Market baskets

Association Rules

Association Analysis

Confident of Association Rule

- Measures how much the consequent (item) is dependent on the antecedent (item)
- The conditional probability that a transaction contains Y given that it contains X

$$conf(X \rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)}$$

Example

$$conf(\{Milk\} \rightarrow \{Bread\}) = \frac{sup(\{Milk, Bread\})}{sup(\{Milk\})}$$

$$= \frac{0.5}{0.7} = 0.71$$

Items

	$Banana$	$Milk$	$Apple$	$Bread$
$\mathbf{x}_1$	0	1	1	0
$\mathbf{x}_2$	1	1	0	0
$\mathbf{x}_3$	0	1	0	1
$\mathbf{x}_4$	1	0	1	0
$\mathbf{x}_5$	0	1	1	1
$\mathbf{x}_6$	1	1	0	1
$\mathbf{x}_7$	0	1	1	1
$\mathbf{x}_8$	0	0	1	0
$\mathbf{x}_9$	0	1	0	1
$\mathbf{x}_{10}$	1	0	1	1

Market baskets

Association Rules

Association Analysis

A rule $X \rightarrow Y$ is said to be frequent if

$$sup(X \rightarrow Y) \geq minsup$$

A rule $X \rightarrow Y$ is said to be strong if

$$conf(X \rightarrow Y) \geq minconf$$

where $\boldsymbol{minsup}$ is a user defined *minimum support threshold*
$\boldsymbol{minconf}$ is a user-specified *minimum confidence threshold*

Example

Given $\boldsymbol{minsup}$ = 0.3 and $\boldsymbol{minconf}$ = 0.5

The rule $\{Milk\} \rightarrow \{Bread\}$ is

- Frequent because $sup(\{Milk, Bread\}) = 0.5 \geq 0.3$
- Strong because $conf(\{Milk\} \rightarrow \{Bread\}) = 0.71 \geq 0.5$

Items

	$Banana$	$Milk$	$Apple$	$Bread$
$\mathbf{x}_1$	0	1	1	0
$\mathbf{x}_2$	1	1	0	0
$\mathbf{x}_3$	0	1	0	1
$\mathbf{x}_4$	1	0	1	0
$\mathbf{x}_5$	0	1	1	1
$\mathbf{x}_6$	1	1	0	1
$\mathbf{x}_7$	0	1	1	1
$\mathbf{x}_8$	0	0	1	0
$\mathbf{x}_9$	0	1	0	1
$\mathbf{x}_{10}$	1	0	1	1

Market baskets

Association Rules

Association Analysis

Lift

- Called improvement or impact
- Measure the difference — measured in ratio — between the confidence of a rule and the expected confidence.
- Lift of a rule $X \rightarrow Y$ is defined as

$$Lift(X \rightarrow Y) = \frac{sup(X \cup Y)}{sup(X) \times sup(Y)}$$

- $Lift(X \rightarrow Y)$ = 1 means that there is no correlation within the itemset.
- $Lift(X \rightarrow Y)$ > 1 means that products in the itemset, **X**, and **Y**, are more likely to be bought together.
- $Lift(X \rightarrow Y)$ < 1 means that products in itemset, **X**, and **Y**, are unlikely to be bought together.

Example

$$Lift(\{Milk\} \rightarrow \{Bread\}) = \frac{sup(\{Milk\} \cup \{Bread\})}{sup(\{Milk\}) \times sup(\{Bread\})} = \frac{0.5}{0.7 \times 0.6} = 1.19$$

Items

	Banana	*Milk*	*Apple*	*Bread*
$\mathbf{x}_1$	0	1	1	0
$\mathbf{x}_2$	1	1	0	0
$\mathbf{x}_3$	0	1	0	1
$\mathbf{x}_4$	1	0	1	0
$\mathbf{x}_5$	0	1	1	1
$\mathbf{x}_6$	1	1	0	1
$\mathbf{x}_7$	0	1	1	1
$\mathbf{x}_8$	0	0	1	0
$\mathbf{x}_9$	0	1	0	1
$\mathbf{x}_{10}$	1	0	1	1

Market baskets

Practice

- $sup(\{\text{tomato}\}, \mathbf{D}) = ?$
- $sup(\{\text{avocado}\}, \mathbf{D}) = ?$
- $sup(\{\text{tomato,avocado}\}, \mathbf{D}) = ?$
- $sup(\{\text{tomato}\} \rightarrow \{\text{avocado}\}) = ?$
- $conf(\{\text{tomato}\} \rightarrow \{\text{avocado}\}) = ?$
- ถ้ากำหนดให้ $\boldsymbol{minsup}$ = 0.25 และ $\boldsymbol{minconf}$ = 0.3 แล้ว กฎความสัมพันธ์ {tomato} → {avocado} จัดว่าเป็นกฎที่เกิดขึ้นบ่อย และน่าเชื่อถือหรือไม่
- $Lift(\{\text{tomato}\} \rightarrow \{\text{avocado}\}) = ?$
- นักศึกษาควรแนะนำลูกค้าให้ซื้อ avocado พร้อมกับ tomato หรือไม่

D	green grapes	avocado	tomato	corn
$\mathbf{x}_1$	0	1	1	0
$\mathbf{x}_2$	1	1	0	0
$\mathbf{x}_3$	0	1	0	1
$\mathbf{x}_4$	1	0	1	0
$\mathbf{x}_5$	0	1	1	1
$\mathbf{x}_6$	1	1	0	1
$\mathbf{x}_7$	0	1	1	1
$\mathbf{x}_8$	0	0	1	0

Reference and Further Study

- Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag.
- Joanes, D. N. and C. A. Gill (1998). "Comparing measures of sample skewness and kurtosis." In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.1, pp. 183–189.
- Raymond H. Meyers, Ronald E. Walpole and, Sharon L. Meyers, and Keying E. Ye (2012). *Probability & Statistics for Engineers & Scientists*. 9th. USA: Prentice Hall.
- Theodoridis, Sergios and Konstantinos Koutroumbas (2008). *Pattern Recognition*. 4th. USA: Academic Press, Inc.
- Westfall, Peter H. (2014). "Kurtosis as Peakedness" In: *The American Statistician* 68.3, pp. 191–195.
- Zaki, Mohammed J. and Wagner Meira Jr (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. USA: Cambridge University Press.