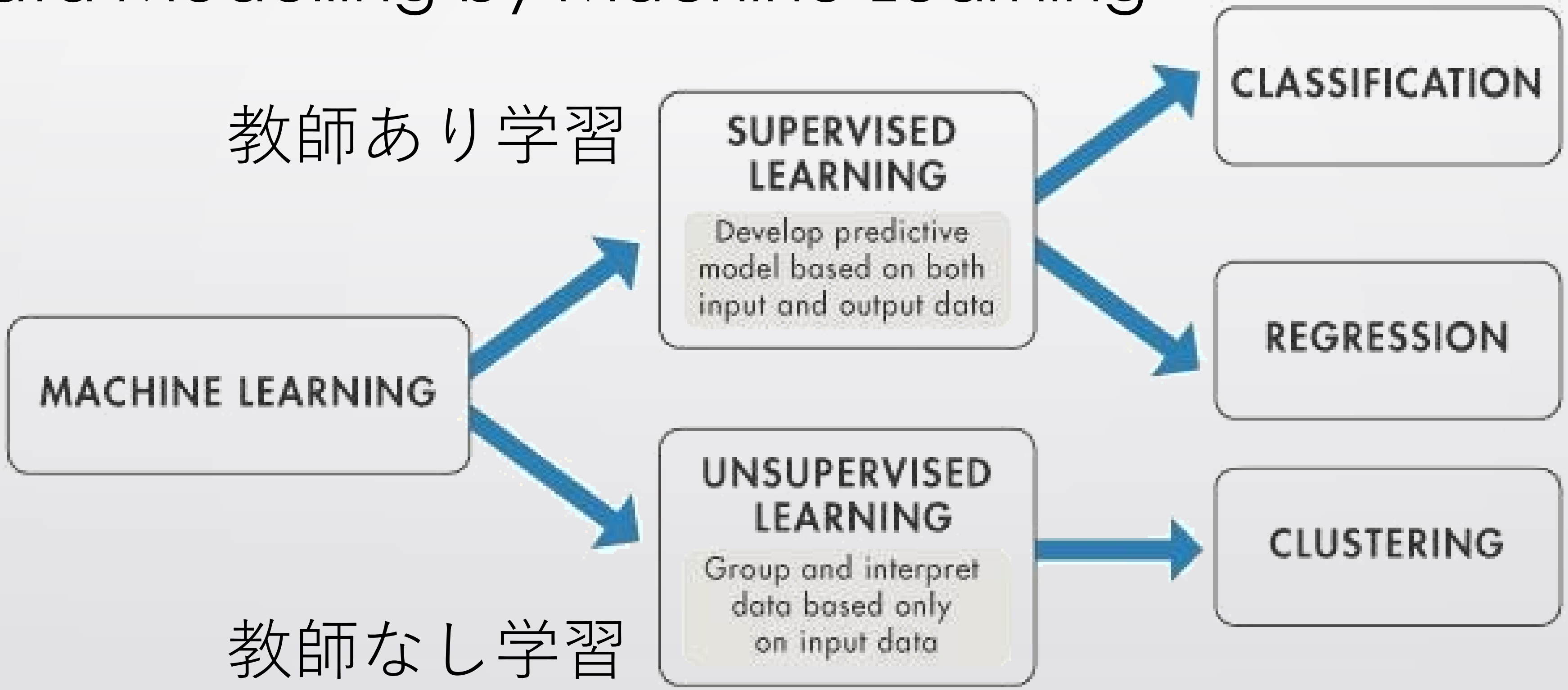
データマイニング
Data Mining

11: クラスタリング① Clustering

土居 裕和 Hirokazu Doi
長岡技術科学大学 Nagaoka University of Technology

機械学習によるモデル化
Data Modelling by Machine Learning

教師あり学習



分類

回帰

クラスタリング

教師なし学習

Supervised Learning versus Unsupervised Learning (Mathworks, n.d.)

クラスタリングの種類 Types of Clustering

非階層的クラスタリング
Non-Hierarchical Clustering

階層的クラスタリング
Hierarchical Clustering

モデル・ベース・クラスタリング
Model-Based Clustering

データの統計的分布についての仮定をおく
Make presumptions about statistical distribution of data

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	~~setosa~~
1	4.9	3.0	1.4	0.2	~~setosa~~
2	4.7	3.2	1.3	0.2	~~setosa~~
3	4.6	3.1	1.5	0.2	~~setosa~~
4	5.0	3.6	1.4	0.2	~~setosa~~
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	~~virginica~~
146	6.3	2.5	5.0	1.9	~~virginica~~
147	6.5	3.0	5.2	2.0	~~virginica~~
148	6.2	3.4	5.4	2.3	~~virginica~~
149	5.9	3.0	5.1	1.8	~~virginica~~

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
...	...	...	...	...
145	6.7	3.0	5.2	2.3
146	6.3	2.5	5.0	1.9
147	6.5	3.0	5.2	2.0
148	6.2	3.4	5.4	2.3
149	5.9	3.0	5.1	1.8

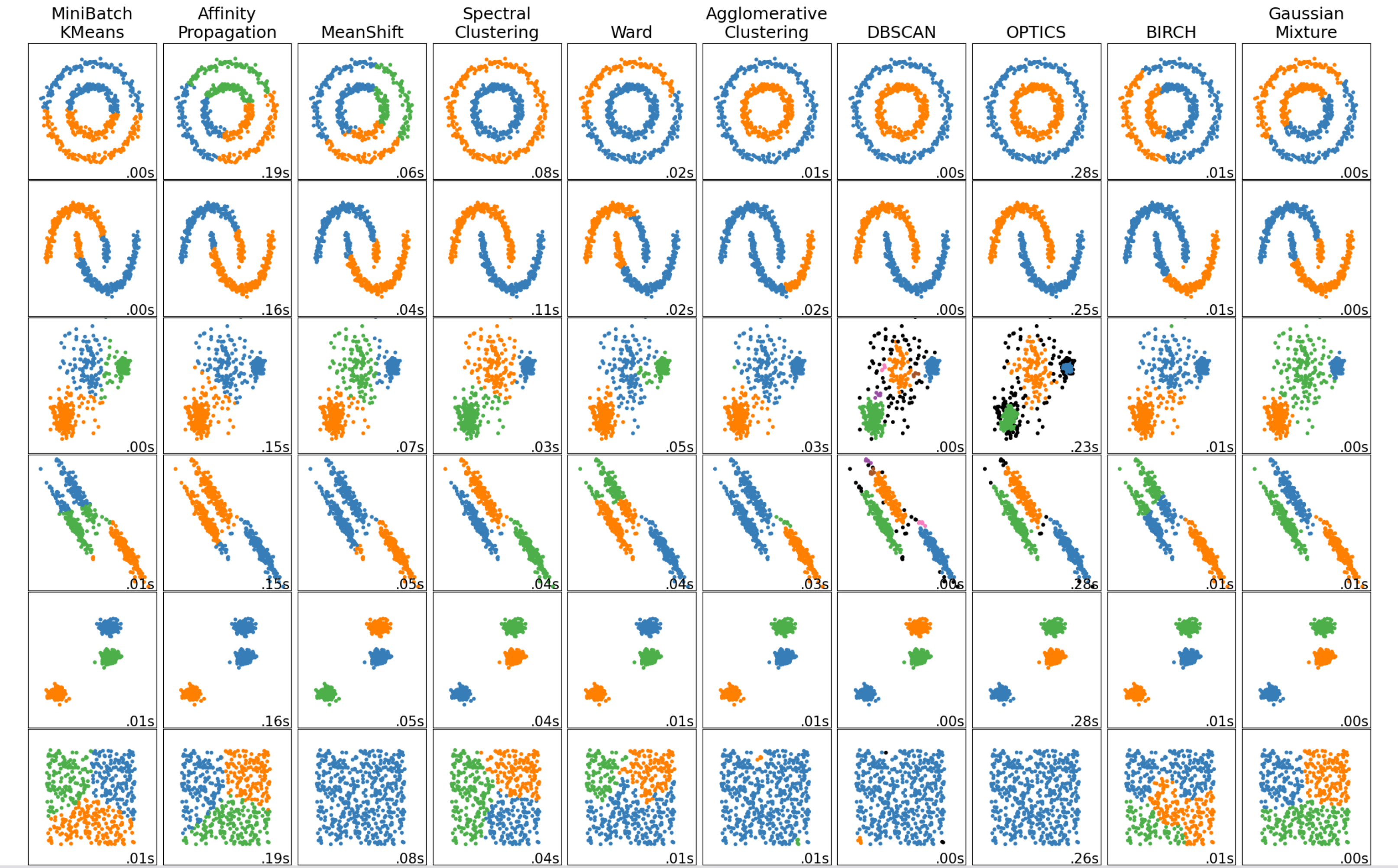
クラスタリング

Final result of clustering depends on

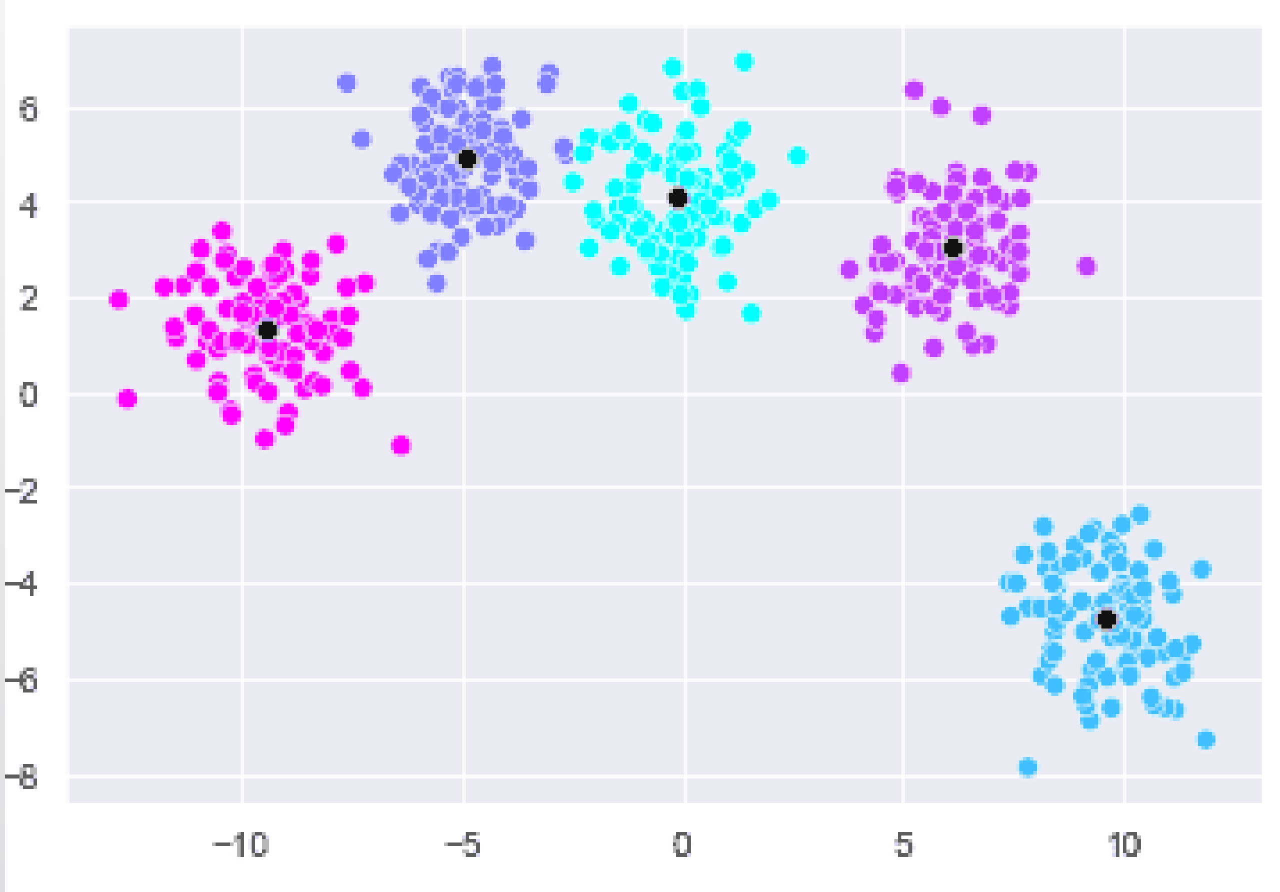
Type of algorithm
アルゴリズムの種類

Parameter Setting
パラメータ設定

https://scikit-learn.org/stable/modules/clustering.html



*K*平均クラスタリング *k*-means clustering



クラスターの数を指定しなくてはいけない

You have to specify the number of clusters, *k*.

*K*平均クラスタリング *k*-means clustering

非階層的クラスタリングの代表的なアルゴリズム

Representative algorithm of non-hierarchical clustering

各クラスターの中心とデータとの距離に基いてクラスタリングを行う

Clustering based on distance between data point and center of each cluster

予めクラスターの数を指定する必要がある

It is necessary to specify the number of clusters beforehand

K平均クラスタリング *k*-means clustering

$D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}, \boldsymbol{x}_{\boldsymbol{i}} \in R^d$

d次元データがN個ある There are N d-dimensional data points

$\boldsymbol{\mu}_k$： k番目のクラスターの代表ベクトル
Representative vector of k–th cluster

$M(\boldsymbol{\mu}_k)$: $\boldsymbol{\mu}_k$のボロノイ領域
Voronoi region of representative vector $\boldsymbol{\mu}_k$

K平均クラスタリング k-means clustering

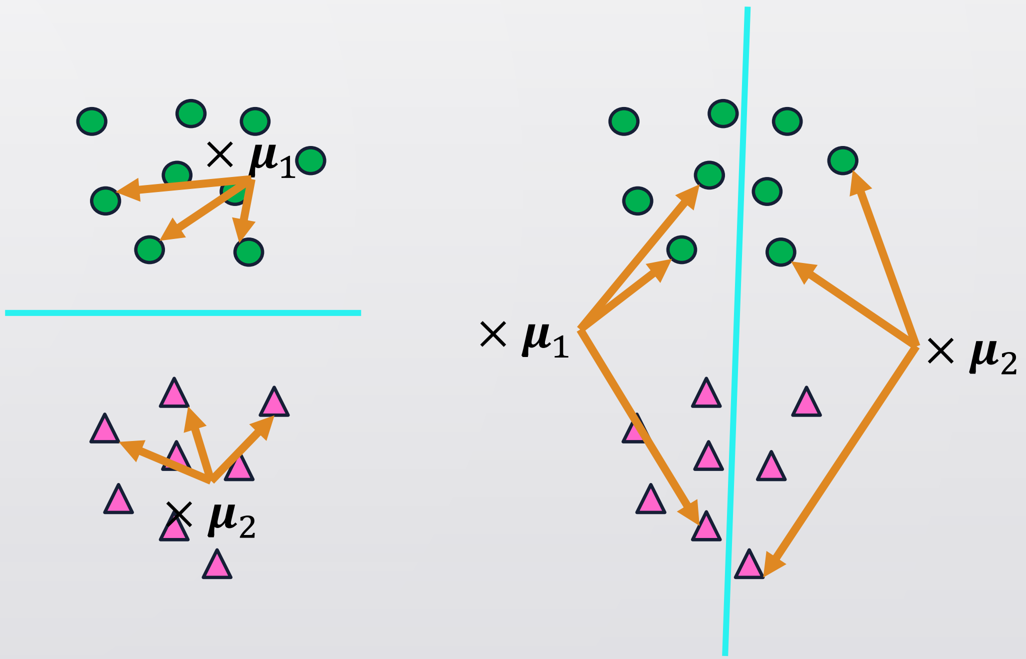
$$q_{i,k} = \begin{cases} 1 & (x_i \in M(\boldsymbol{\mu}_k)\text{の場合}) \\ 0 & \end{cases}$$

In case of $x_i \in M(\boldsymbol{\mu}_k)$

$$J(q_{i,k}, \mu_k) = \sum_{i=1}^{N} \sum_{k=1}^{K} q_{i,k} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2$$

$J(q_{i,k}, \mu_k)$を最小化する$q_{i,k}$と$\boldsymbol{\mu}_k$を求める

Find $q_{i,k}$ and $\boldsymbol{\mu}_k$ that minimize $J(q_{i,k}, \mu_k)$



K平均クラスタリング k-means clustering

$$J(q_{i,k}, \mu_k) = \sum_{i=1}^{N} \sum_{k=1}^{K} q_{i,k} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2$$

$$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = 0 \qquad -2\sum_{i=1}^{N} q_{i,k}(\boldsymbol{x}_i - \boldsymbol{\mu}_k) = 0$$

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} q_{i,k}\boldsymbol{x}_i}{\sum_{i=1}^{N} q_{i,k}}$$

$\boldsymbol{\mu}_k$と$q_{i,k}$を同時に最適化するにはどうすればいいか？

How can we optimize $\boldsymbol{\mu}_k$ and $q_{i,k}$ simultaneously?

K平均クラスタリング k-means clustering

$$J(q_{i,k}, \mu_k) = \sum_{i=1}^{N} \sum_{k=1}^{K} q_{i,k} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2$$

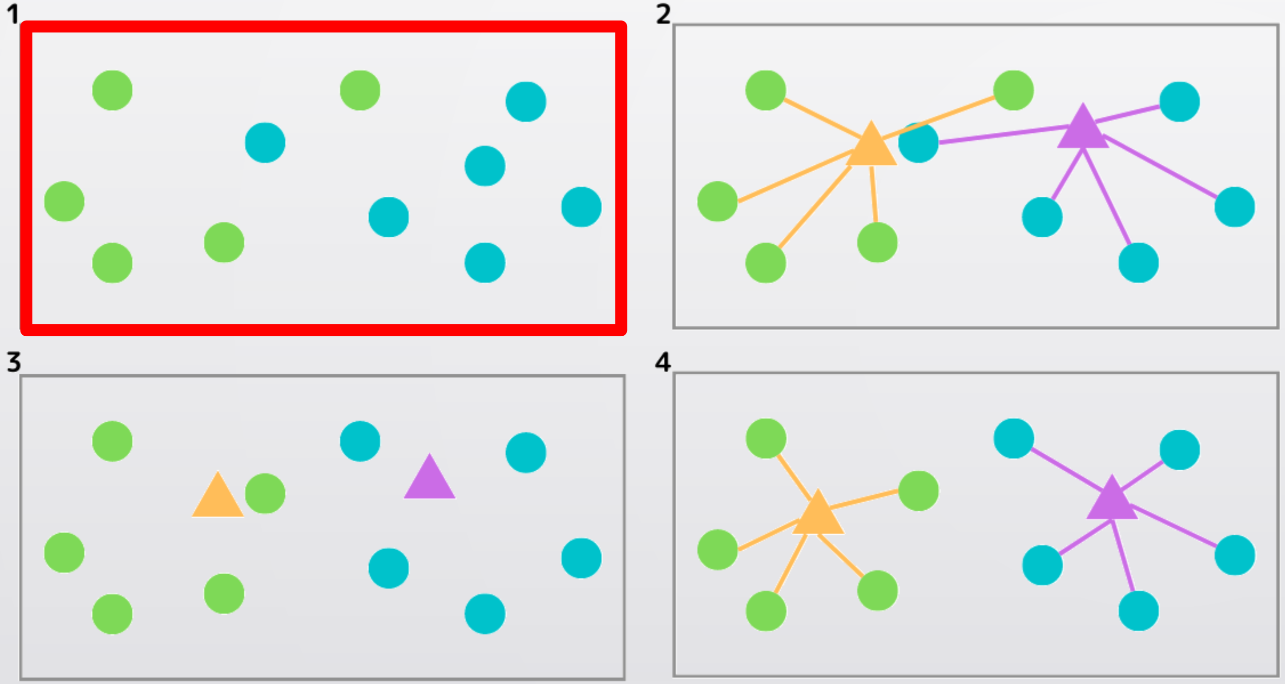
$$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = 0 \quad -2 \sum_{i=1}^{N} q_{i,k} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) = 0$$

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} q_{i,k} \boldsymbol{x}_i}{\sum_{i=1}^{N} q_{i,k}}$$

$\boldsymbol{\mu}_k$と$q_{i,k}$を同時に最適化するにはどうすればいいか？

How can we optimize $\boldsymbol{\mu}_k$ and $q_{i,k}$ simultaneously?

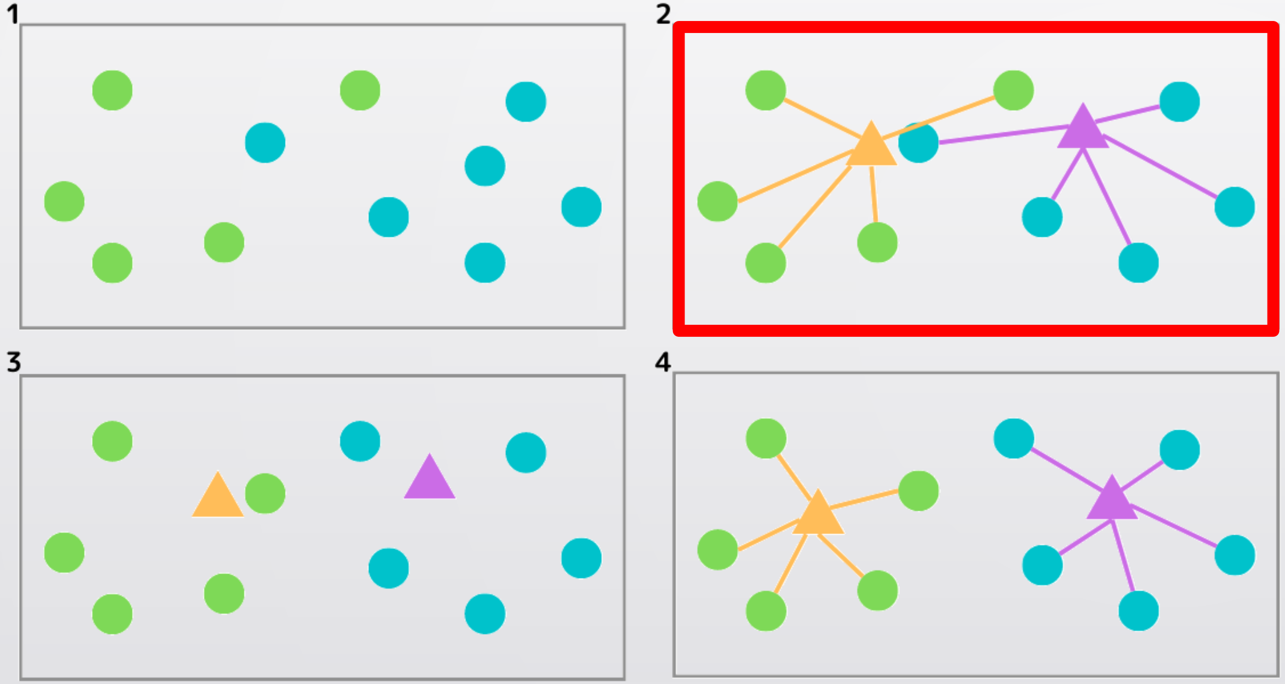
K平均クラスタリング *k*-means clustering



1. データをランダムにクラスターに割り当てる

Randomly assign data to one of the clusters

K平均クラスタリング *k*-means clustering

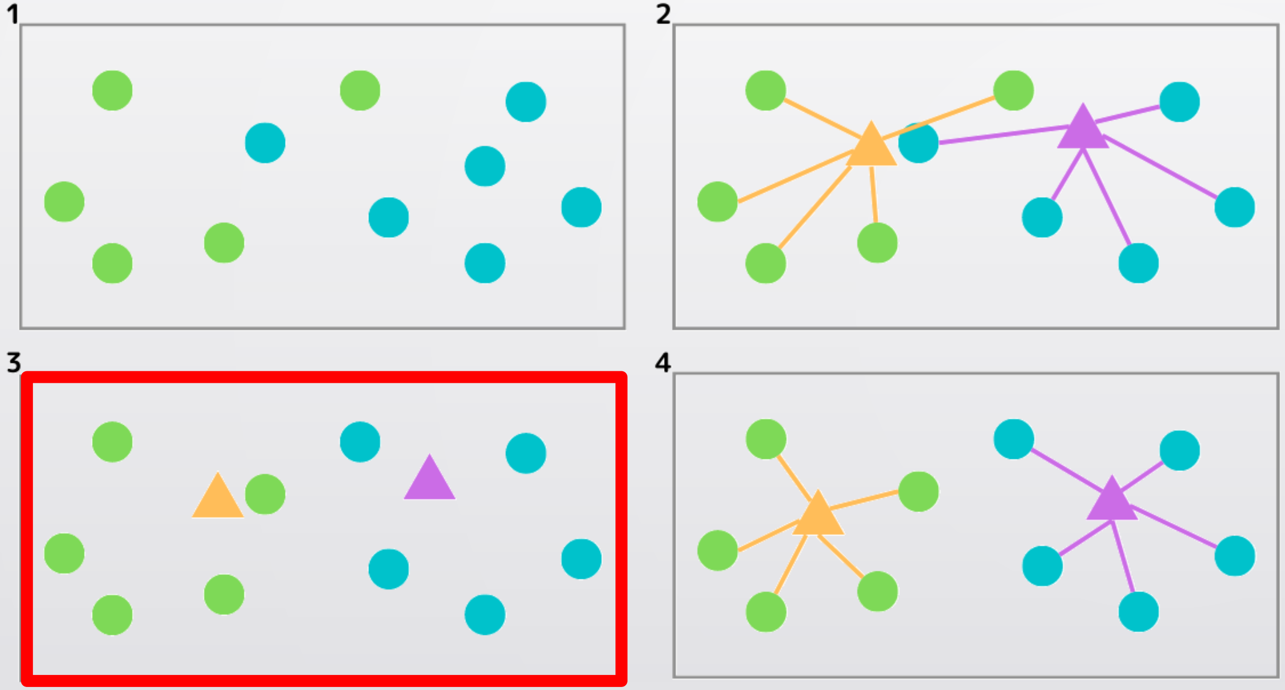


2. 各クラスターの中心を計算する
Compute the center of each cluster

3. 各クラスタ中心とデータとの距離を計算する

Compute the distance of data from center of each cluster

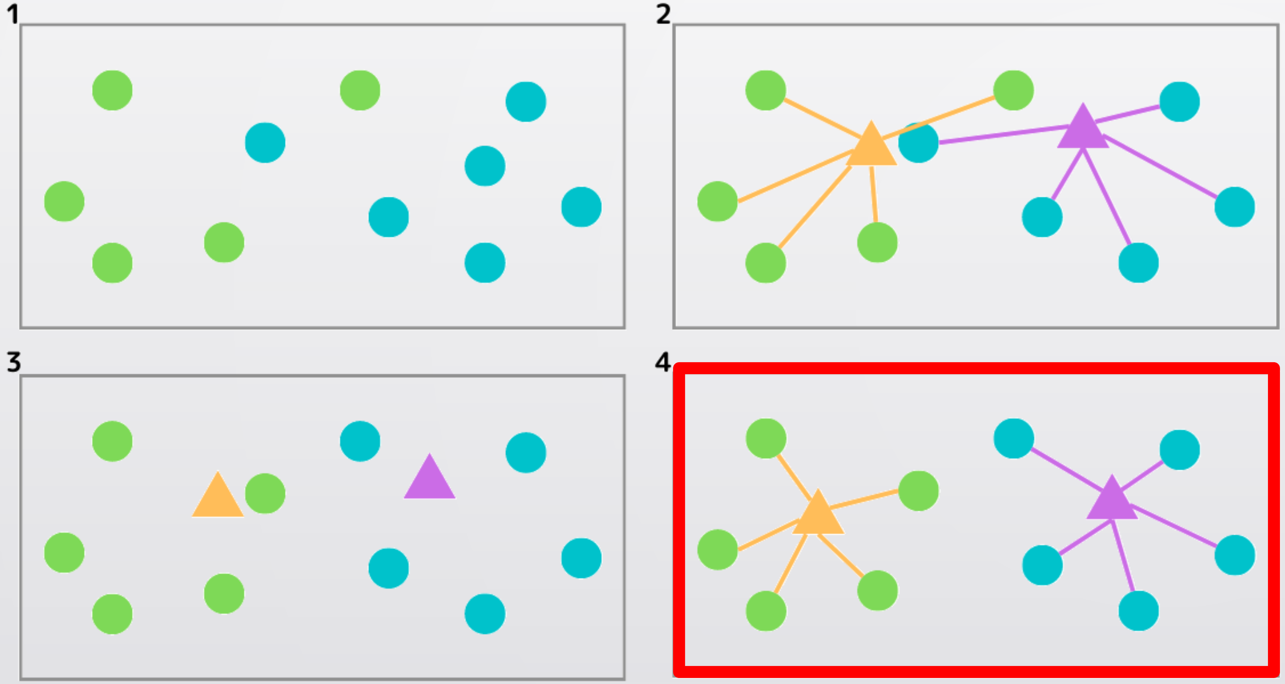
K平均クラスタリング *k*-means clustering



3. データと最も中心からの距離が近いクラスターに割り当てる

Assign data point to the cluster with smallest distance

K平均クラスタリング k-means clustering



4. データと最も中心からの距離が近いクラスターに割り当てる

Assign data point to the cluster with smallest distance

K平均クラスタリング k-means clustering

1. $\boldsymbol{\mu}_k$を固定し以下の方法で$q_{i,k}$を決定する

Fix $\boldsymbol{\mu}_k$ and determine $q_{i,k}$ following the rule below

$$q_{i,k} = \begin{cases} 1 \;\; if \;\; k = argmin_j \left\| \boldsymbol{x}_i - \boldsymbol{\mu}_j \right\|^2 \\ 0 \end{cases}$$

$\boldsymbol{x}_i$を重心$\boldsymbol{\mu}_j$との距離が一番近いクラスタに割り当てる

Assign $\boldsymbol{x}_i$ to the cluster whose centroid $\boldsymbol{\mu}_j$ is closest to $\boldsymbol{x}_i$

2. $\boldsymbol{\mu}_k$を最適化する

Optimize $\boldsymbol{\mu}_k$

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} q_{i,k} \boldsymbol{x}_i}{\sum_{i=1}^{N} q_{i,k}}$$

*K*平均クラスタリング *K*-means clustering

1-2.の手続きを収束するまで繰り返す Repeat Step 1-4 until the result converges

クラスター内の誤差平方和が閾値以下になることが収束条件である

Convergence Criterion is usually that squared-sum within cluster SSE_k becomes smaller than threshold

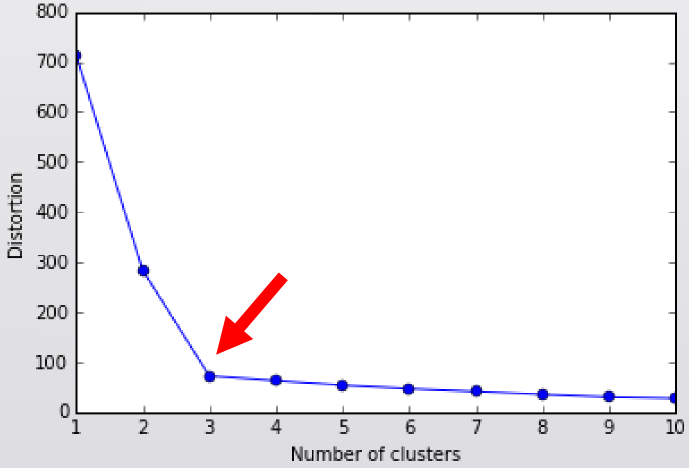
$$SSE_k = \sum \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2 \qquad \sum SSE_k \leq Threshold$$

距離は、通常、ユークリッド距離を計算する

Usually, Euclidian distance is computed as the index of distance between data point and cluster center

クラスター数の決定法: エルボー法
How to determine the number of Clusters: Elbow Method

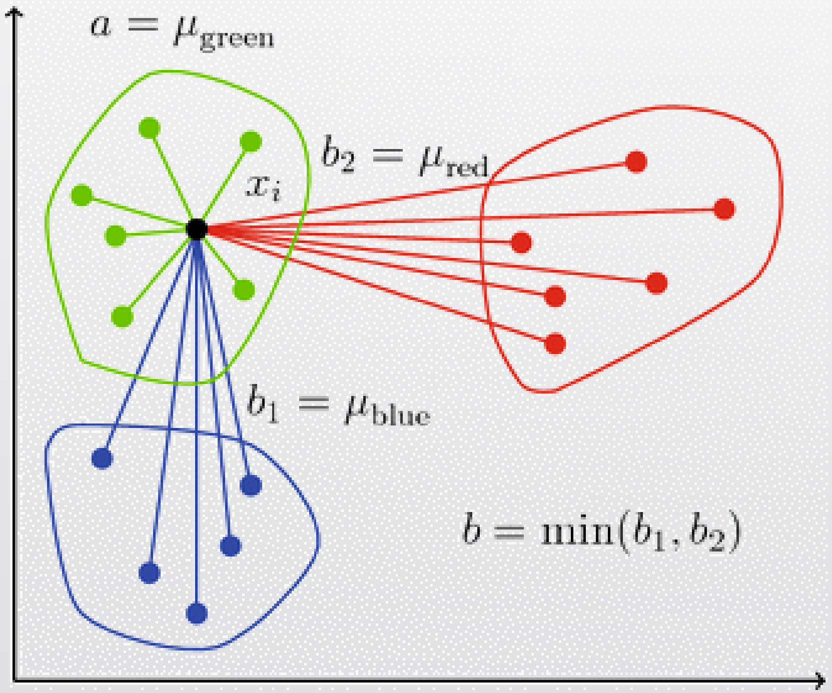
1. 様々なクラスター数でクラスタリングを行いSSEを計算する
   Compute SSE after clustering with varying number of clusters
2. SSEをプロットし、SSEの減少が平たんになるクラスター数を見つける
   Plot SSEs and find the cluster number at which decrease of SSE reaches plateau



https://qiita.com/deaikei/items/11a10fde5bb47a2cf2c2

クラスター数の決定法: シルエット係数

How to determine the number of Clusters: Silhouette Coefficient



Usha Narra et al, 2016

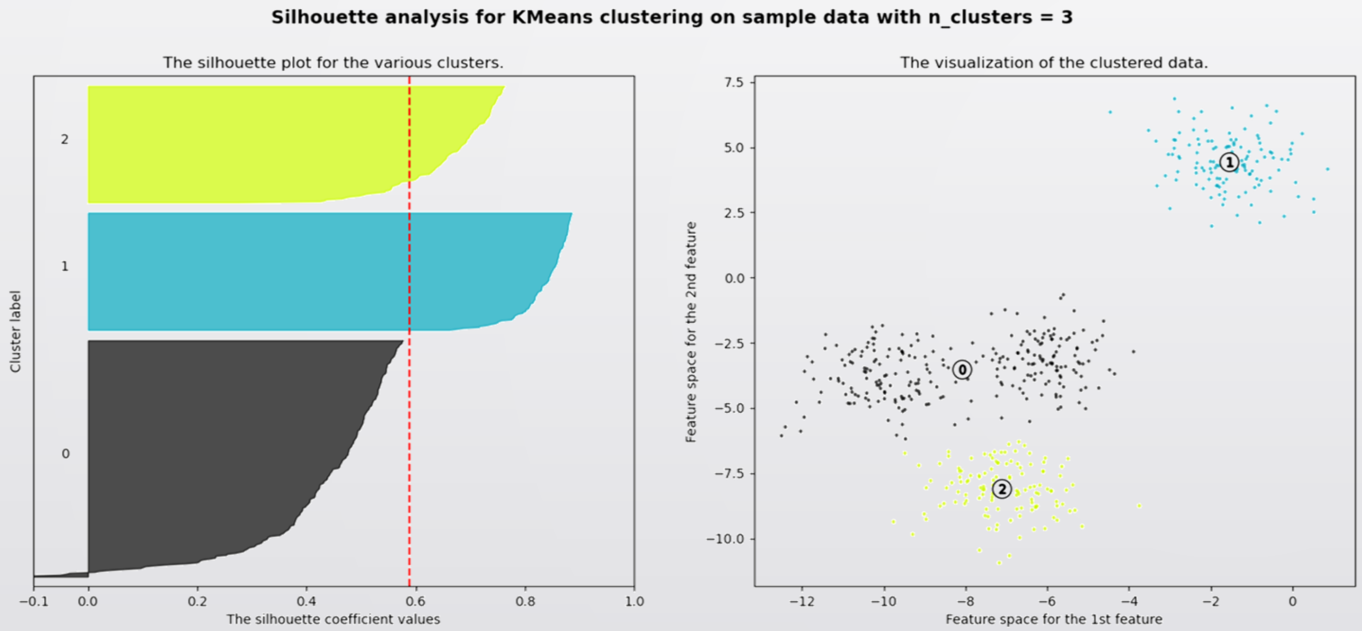
a: クラスター内の他のデータとの距離の平均
a: Mean distance from other data points within cluster

b: 最も近いクラスターのデータとの距離の平均
a: Mean distance from data points in nearest cluster

$$Silhouette\ Coefficient = \frac{b - a}{max(b, a)}$$

$[-1, 1]$の範囲で変動する
Ranges within $[-1, 1]$

クラスター数の決定法: シルエット係数
How to determine the number of Clusters: Silhouette Coefficient



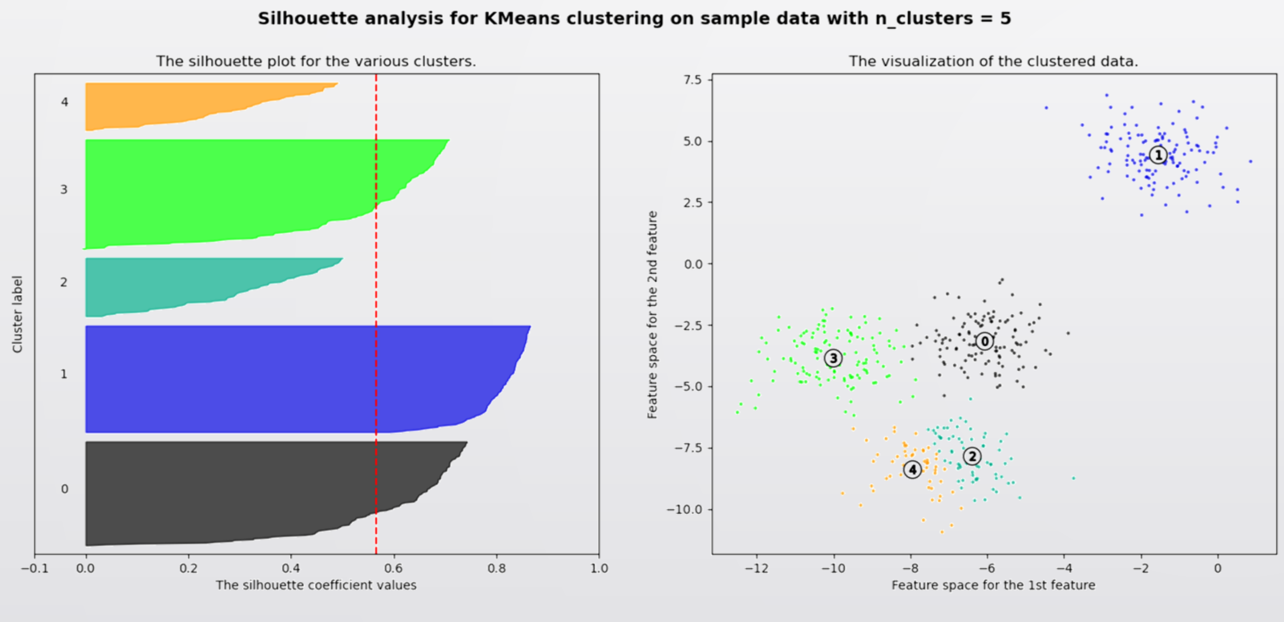
全てのデータのシルエット係数をプロットしている
Silhouette coefficient of every data point is plotted

クラスター数が不適切な時
When number of clusters is not appropriate

シルエット係数が小さなクラスターがある
There are clusters with small silhouette coefficient

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

クラスター数の決定法: シルエット係数
How to determine the number of Clusters: Silhouette Coefficient



全てのデータのシルエット係数をプロットしている
Silhouette coefficient of every data point is plotted

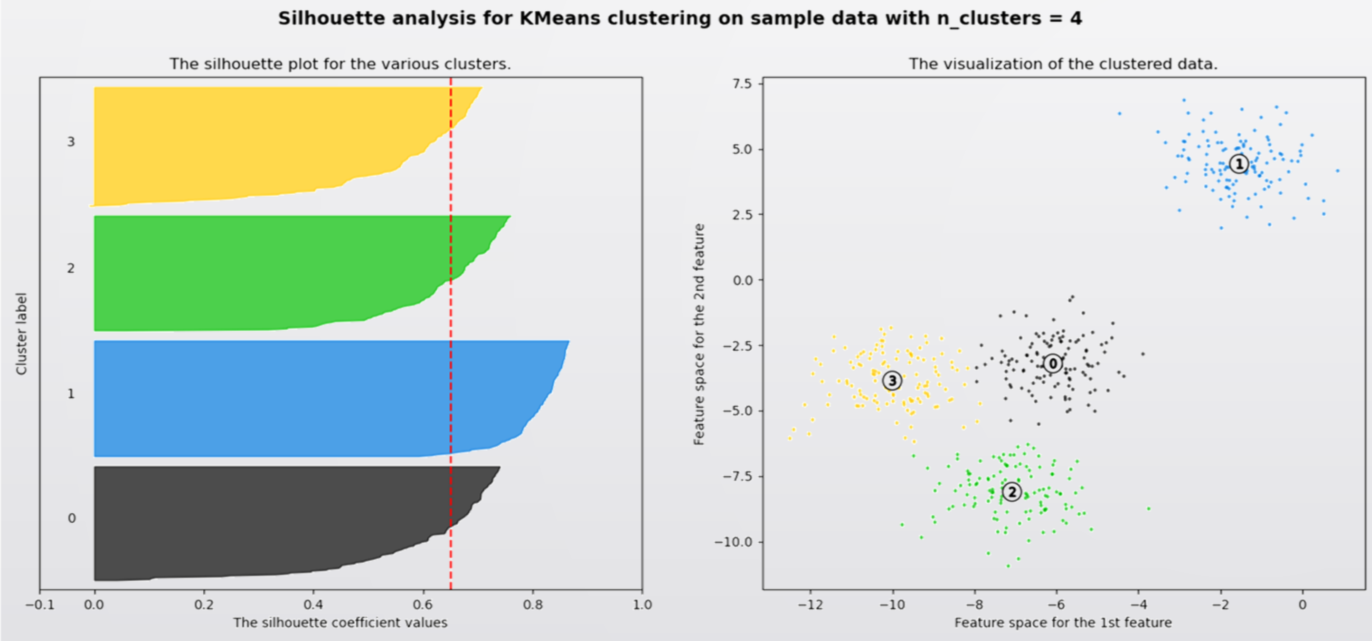
クラスター数が不適切な時
When number of clusters is not appropriate

シルエット係数が小さなクラスターがある
There are clusters with small silhouette coefficient

クラスターの大きさが不均一
Size of cluster is inhomogenous

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

クラスター数の決定法: シルエット係数
How to determine the number of Clusters: Silhouette Coefficient



全てのデータのシルエット係数をプロットしている
Silhouette coefficient of every data point is plotted

クラスター数が適切な時
When number of clusters is not appropriate

全てのクラスターのシルエット係数が大きい
Silhouette coefficient of every cluster is large enough

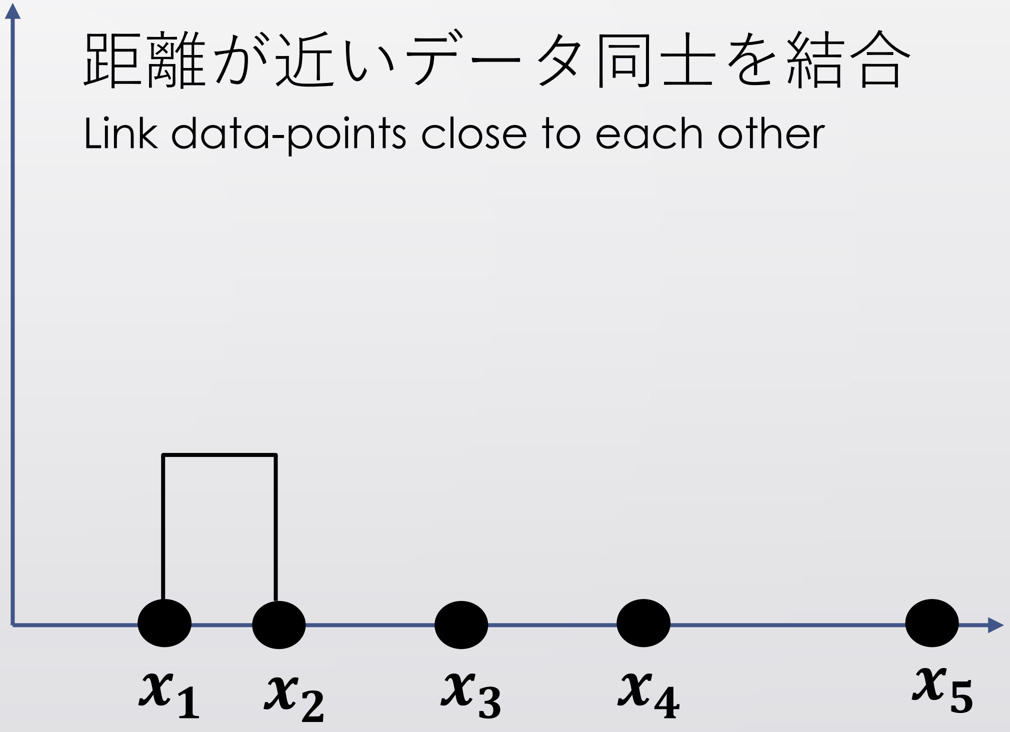
クラスターの大きさが均一
Size of cluster is homogenous

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

凝集性階層的クラスタリング
Agglomerative Hierarchical Clustering

	x_1	x_2	x_3	x_4	x_5
x_1	0				
x_2	1	0			
x_3	4	3	0		
x_4	6	5	2	0	
x_5	10	9	6	4	0

x_1	x_2	x_3	x_4	x_5
1	2	5	7	11



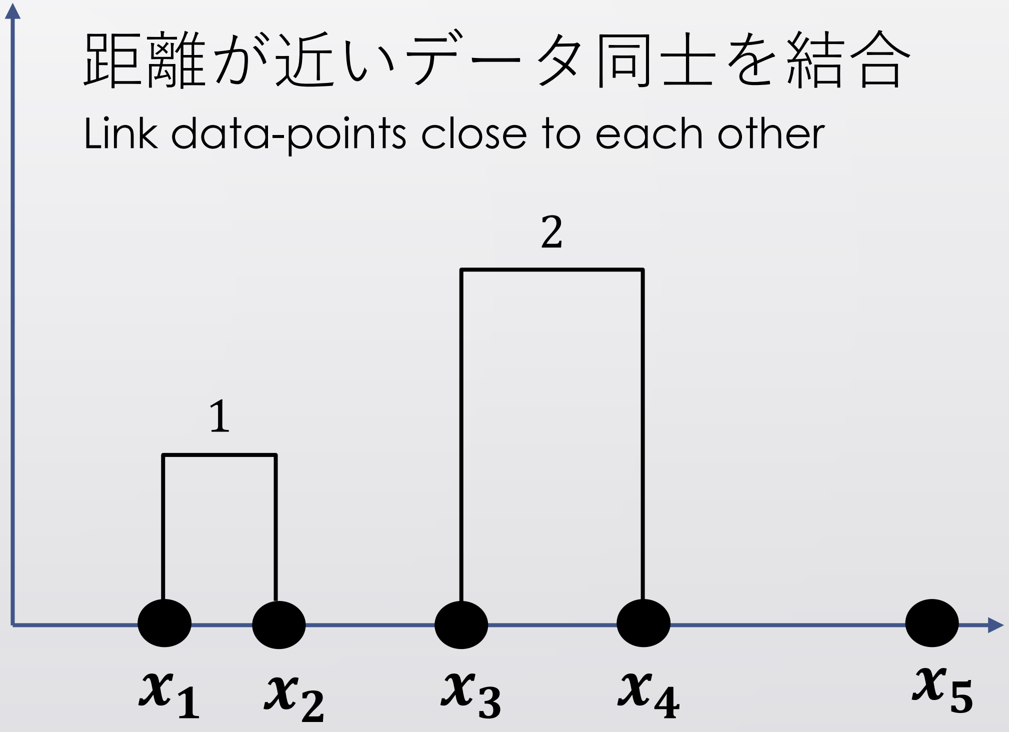
凝集性階層的クラスタリング
Agglomerative Hierarchical Clustering

	$\{x_1, x_2\}$	x_3	x_4	x_5
$\{x_1, x_2\}$	0			
x_3	3	0		
x_4	5	2	0	
x_5	9	6	4	0

x_1	x_2	x_3	x_4	x_5
1	2	5	7	11

距離が近いデータ同士を結合

Link data-points close to each other



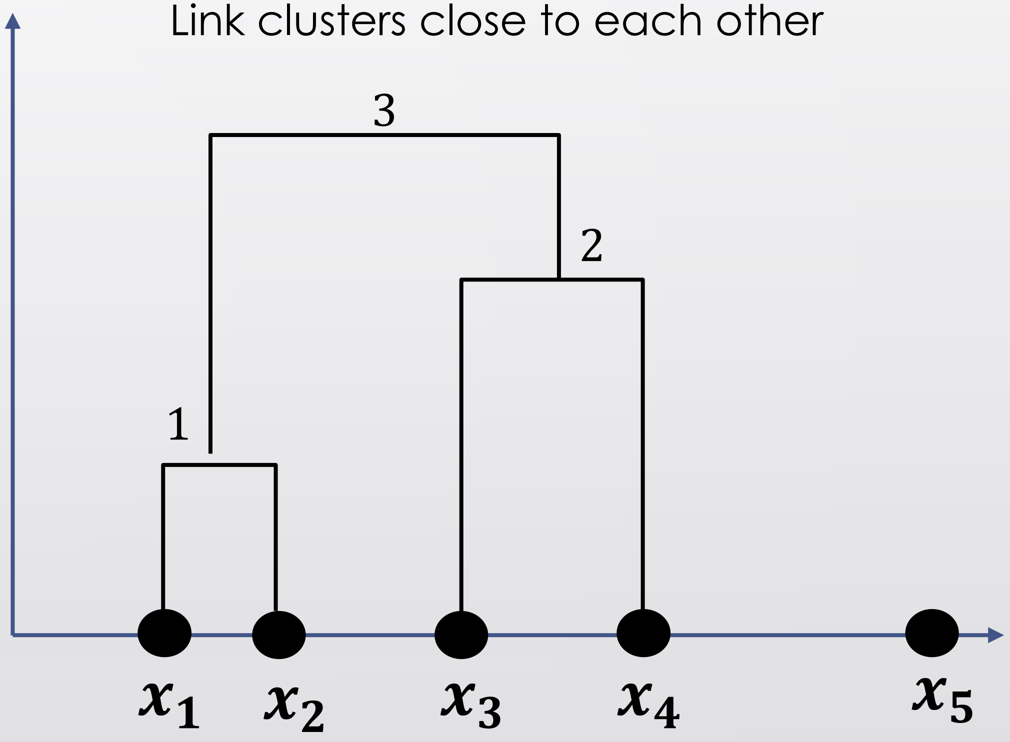
凝集性階層的クラスタリング
Agglomerative Hierarchical Clustering

距離が近いクラスター同士を結合

Link clusters close to each other

	$\{x_1, x_2\}$	$\{x_3, x_4\}$	x_5
$\{x_1, x_2\}$	0		
$\{x_3, x_4\}$	3	0	
x_5	9	4	0

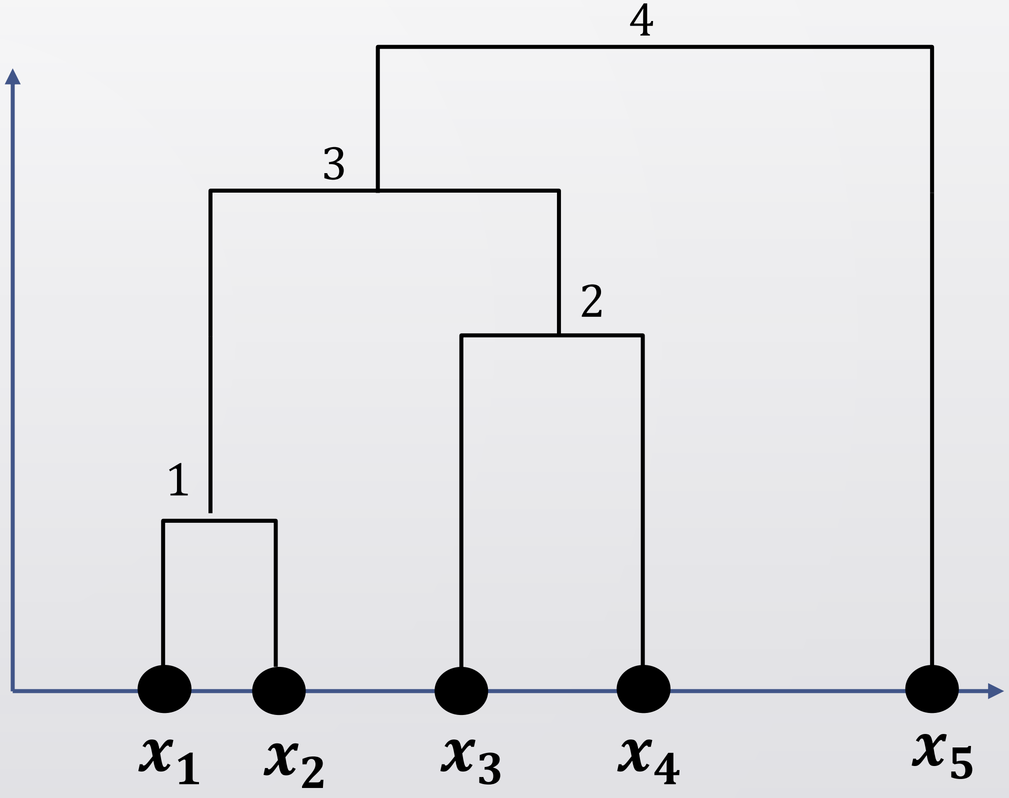
x_1	x_2	x_3	x_4	x_5
1	2	5	7	11



凝集性階層的クラスタリング
Agglomerative Hierarchical Clustering

	$\{x_1, x_2, x_3, x_4\}$	x_5
$\{x_1, x_2, x_3, x_4\}$	0	
x_5	4	0

x_1	x_2	x_3	x_4	x_5
1	2	5	7	11

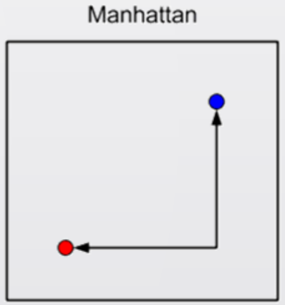


データ間の距離 Distance between Data-Points

ミンコフスキ距離 Minkowski Distance $\quad Minkowski\ Distance = \left(\sum \left|x_{i,k} - x_{j,k}\right|^p\right)^{\frac{1}{q}}$

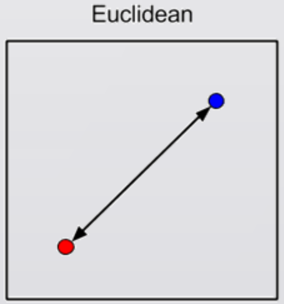
$p = 1,\ q = 1$の時 In case of $p = 1,\ q = 1$

マンハッタン距離 $\quad Manhatthan\ Distance = \sum \left|x_{i,k} - x_{j,k}\right|$



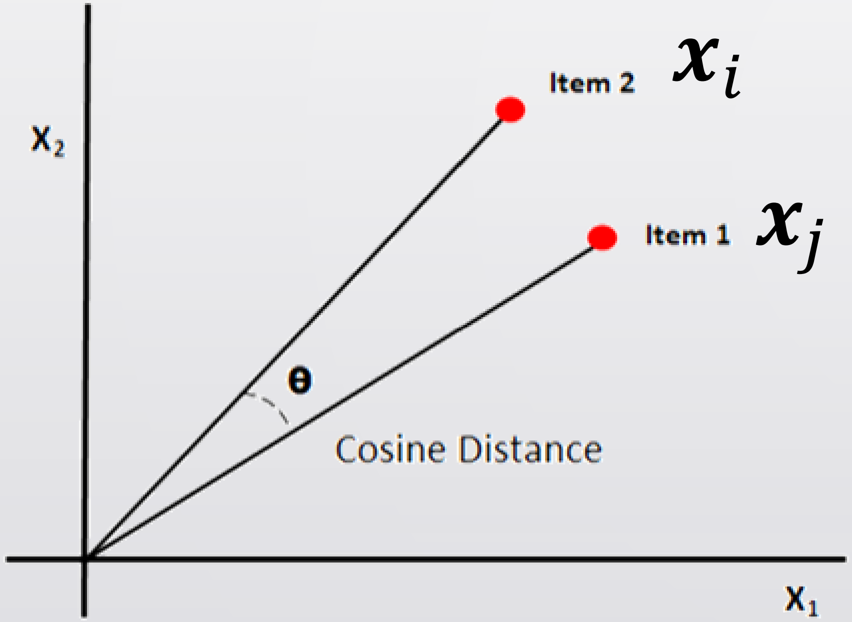
$p = 2,\ q = 2$の時 In case of $p = 2,\ q = 2$

ユークリッド距離 $\quad Euclidean\ Distance = \sqrt{\sum \left(x_{i,k} - x_{j,k}\right)^2}$



データ間の距離 Distance between Data-Points

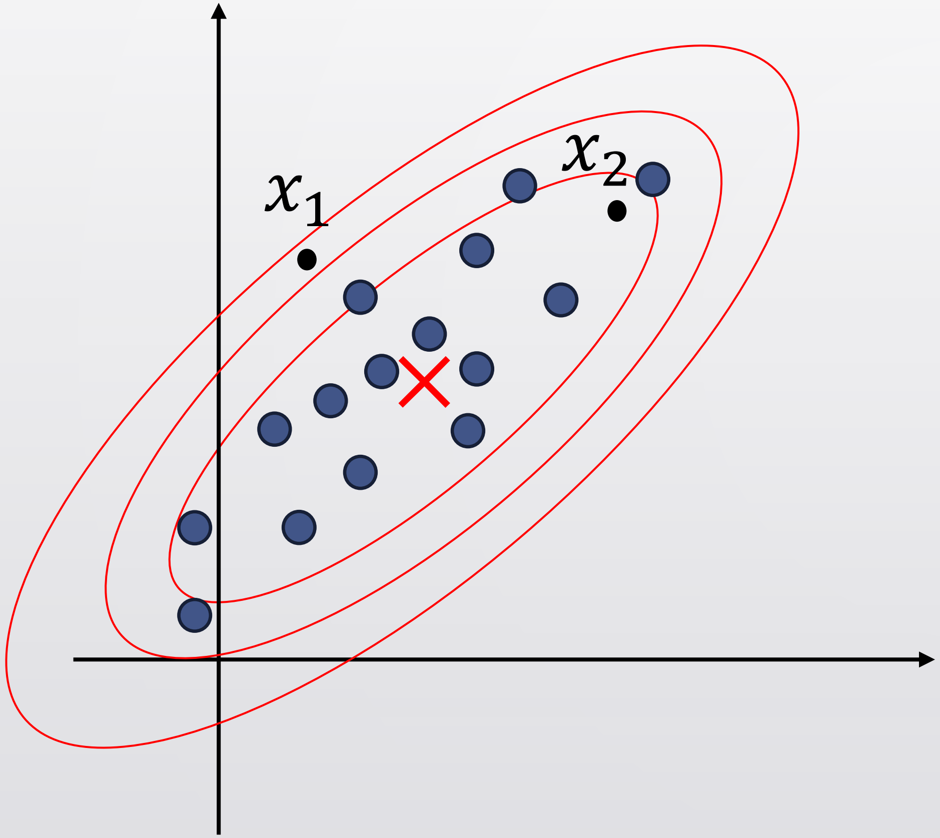
コサイン類似度 Cosine Similarity



$$D\left(\boldsymbol{x_i}, \boldsymbol{x_j}\right) = \frac{dot\left(\boldsymbol{x_i}, \boldsymbol{x_j}\right)}{\|\boldsymbol{x_i}\|\|\boldsymbol{x_j}\|} = \frac{\sum x_{i,k} \times x_{j,k}}{\sum {x_{i,k}}^2 \times \sum {x_{j,k}}^2}$$

https://www.oreilly.com/library/view/statistics-for-machine/9781788295758/eb9cd609-e44a-40a2-9c3a-f16fc4f5289a.xhtml

マハラノビス距離 Mahalanobis Distance



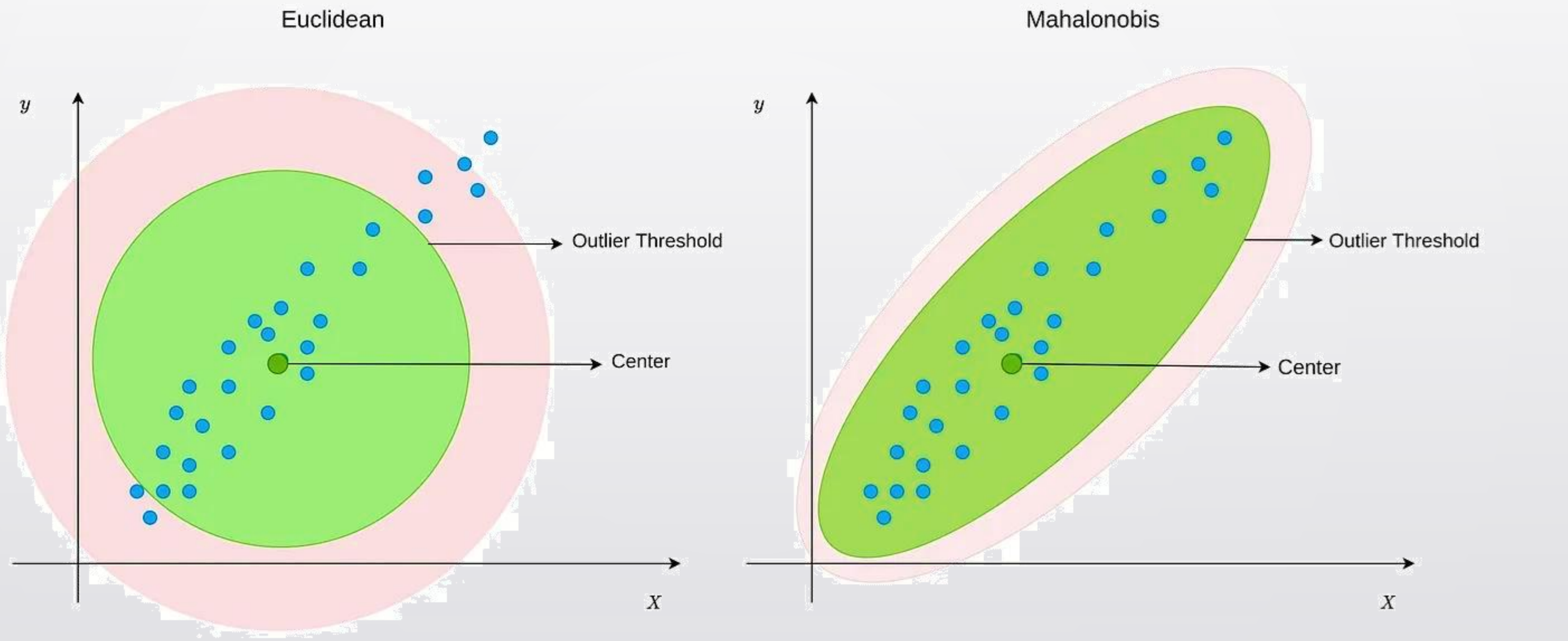
$$D_M(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\Sigma^{-1}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T}$$

Σ: 分散共分散行列
Variance-Covariance Matrix

分布の形状を考慮した分布の中心からの距離の指標

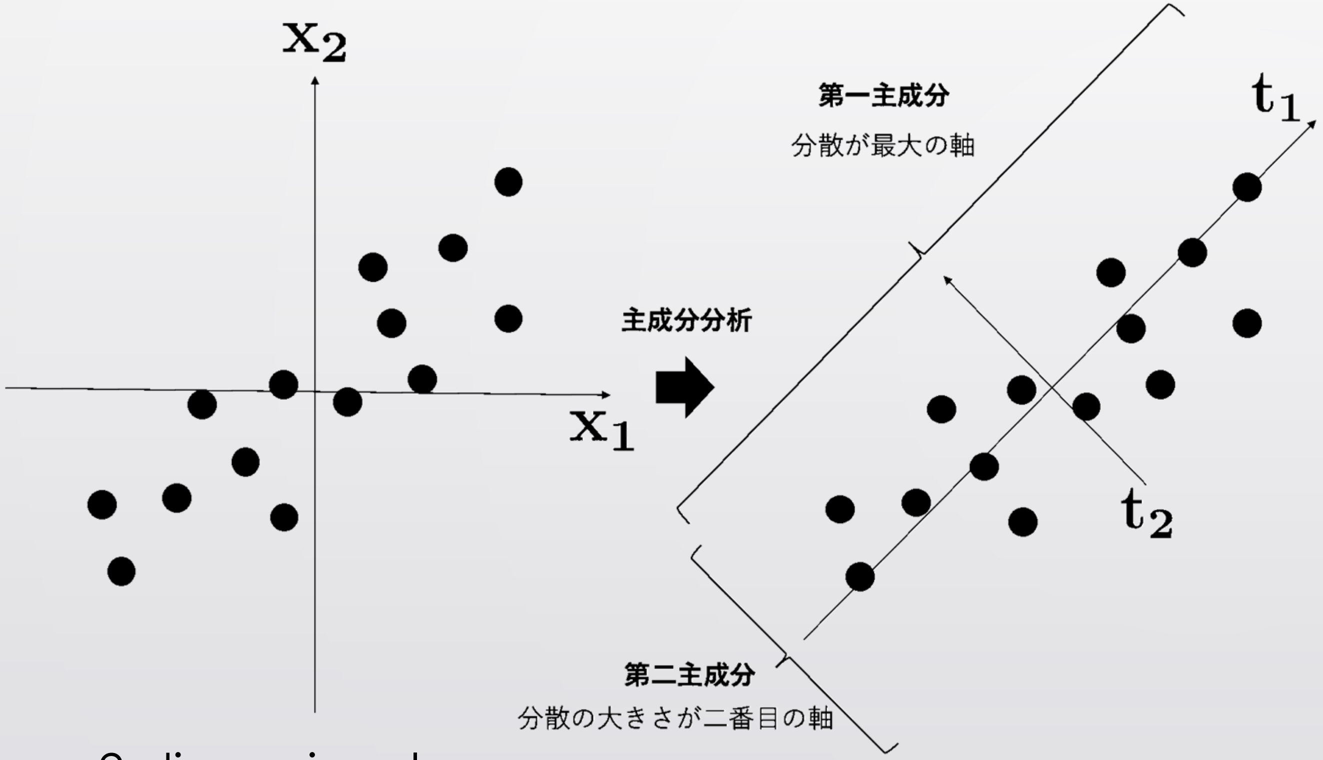
Measure of distance from distribution center adjusted by the shape of data distribution

マハラノビス距離 Mahalanobis Distance



https://bob3.hatenablog.com/entry/2023/04/22/113540

主成分 Principal Components



2-dimensional

第1主成分軸は、データの分散が最大化される方向を向いている

The first PC axis is oriented in the direction along which variance of projected data is maximized

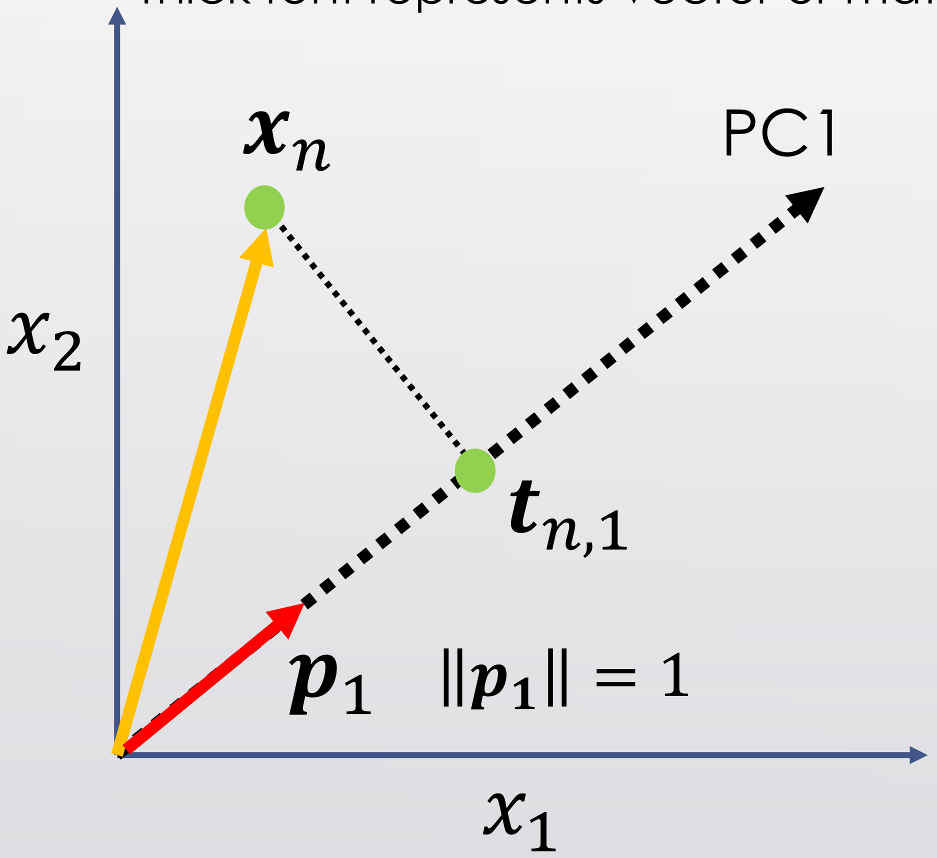
第 *j* 主成分軸は、データの分散が *j* 番目の大きさになる方向を向いている

The *j*-th PC axis is oriented in the direction along which projected data has *j*-th largest variance

https://free.kikagaku.ai/tutorial/basic_of_machine_learning/learn/machine_learning_unsupervised

第1主成分の計算 Computation of PC1

太字はベクトルか行列
Thick font represents vector or matrix



変数を中心化しておく Center the variables

観測データ$\boldsymbol{x}_n$の第1主成分軸方向への射影$t_{n,1}$を計算する

Compute the projection $t_{n,1}$ of the observed data $\boldsymbol{x}_n$ onto the first PC axis

$t_{n,1}$は$\boldsymbol{x}_n$と$\boldsymbol{p}_1$の内積

$t_{n,1}$is dot(inner) product of $\boldsymbol{x}_n$ and $\boldsymbol{p}_1$

第1主成分の計算 Computation of PC1

$\boldsymbol{V}\boldsymbol{p_1} = \lambda\boldsymbol{p_1}$　$\boldsymbol{p_1}$は$\boldsymbol{V}$の固有ベクトルである　$\boldsymbol{p_1}$is eigenvector of $\boldsymbol{V}$

$\boldsymbol{V} = \frac{1}{N}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$　$\boldsymbol{t_1}$の分散は λ に一致する　Variance of $\boldsymbol{t_1}$equals to λ

$\boldsymbol{V}$とは何か？ What is $\boldsymbol{V}$?

$$\boldsymbol{V} = \frac{1}{N}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} \quad \boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}$$

分散共分散行列 Variance-Covariance Matrix

$\boldsymbol{V}$は$\boldsymbol{X}$の分散共分散行列である $\boldsymbol{V}$ is variance-covariance matrix of $\boldsymbol{X}$

$$\boldsymbol{V} = \frac{1}{N}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} = \frac{1}{N}\begin{bmatrix} x_{1,1} & x_{2,1} & \dots & x_{N,1} \\ x_{1,2} & x_{2,2} & \dots & x_{N,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,M} & x_{2,M} & \dots & x_{N,M} \end{bmatrix}\begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{1,1}^2 & \sigma_{1,2}^2 & \dots & \sigma_{1,M}^2 \\ \sigma_{2,1}^2 & & & \vdots \\ & & \ddots & \sigma_{M-1,M}^2 \\ \vdots & & & \\ \sigma_{M,1}^2 & \sigma_{M,2}^2 & \dots & \sigma_{M,M}^2 \end{bmatrix} \qquad \sigma_{i,j}^2 = \frac{1}{N}\sum_{k=1}^{N} x_{k,i}\, x_{k,j}$$

分散共分散行列の対角化
Diagonalization of Variance-Covariance Matrix

$$\boldsymbol{P} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots \boldsymbol{p}_M)$$

$$\boldsymbol{V}\boldsymbol{p}_{\boldsymbol{i}} = \lambda_i \boldsymbol{p}_{\boldsymbol{i}} \quad \|\boldsymbol{p}_i\| = 1$$

対称行列の異なる固有値に対する固有ベクトルは直交するので
Since eigenvectors of symmetric matrix corresponding to different eigen values are orthogonal

$$\boldsymbol{p}_i \boldsymbol{p}_j^T = \begin{cases} 1 \ (\lambda_i = \lambda_j) \\ 0 \ (\lambda_i \neq \lambda_j) \end{cases}$$

分散共分散行列の対角化
Diagonalization of Variance-Covariance Matrix

$$\boldsymbol{VP} = \boldsymbol{V}(\boldsymbol{p}_1, \boldsymbol{p}_2, \dots \boldsymbol{p}_M) = (\lambda_1 \boldsymbol{p}_1, \lambda_2 \boldsymbol{p}_2, \dots \lambda_M \boldsymbol{p}_M)$$

$$\boldsymbol{p}_i^T \boldsymbol{VP} = \left(\lambda_1 \boldsymbol{p}_i^T \boldsymbol{p}_1, \lambda_2 \boldsymbol{p}_i^T \boldsymbol{p}_2, \dots \lambda_M \boldsymbol{p}_i^T \boldsymbol{p}_M\right) = (0,0, \dots \lambda_i \dots, 0)$$

$$\boldsymbol{P}^T \boldsymbol{VP} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \lambda_M \end{bmatrix}$$

マハラノビス距離の意味 Meaning of Mahalanobis Distance

$\boldsymbol{x}_1 = (x_{11}, x_{12}) \quad \boldsymbol{x}_2 = (x_{21}, x_{22})$

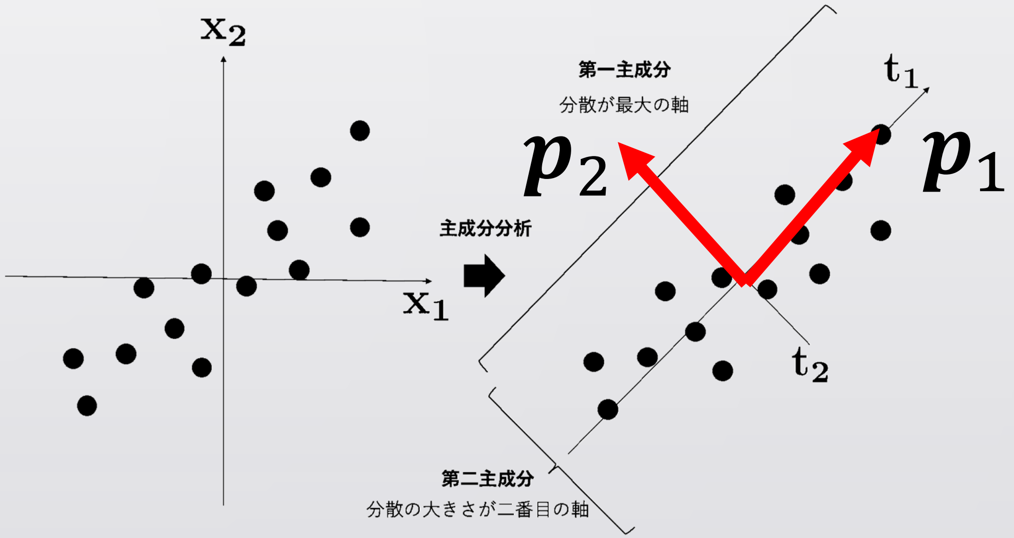
$\boldsymbol{x}_i$の$\boldsymbol{p}_k$への射影を計算 Project $\boldsymbol{x}_i$ onto $\boldsymbol{p}_k$

$\boldsymbol{u}_1 = (u_{11}, u_{12}) = \boldsymbol{x}_1(\boldsymbol{p}_1, \boldsymbol{p}_2)$

$\boldsymbol{u}_2 = (u_{21}, u_{22}) = \boldsymbol{x}_2(\boldsymbol{p}_1, \boldsymbol{p}_2)$

$(\boldsymbol{p}_1, \boldsymbol{p}_2)$を基底とする座標系では $\boldsymbol{x}_1$は$\boldsymbol{u}_1$と表現される

$\boldsymbol{x}_1$ is expressed as $\boldsymbol{u}_1$ in a coordinate system defined by basis vectors of $(\boldsymbol{p}_1, \boldsymbol{p}_2)$



マハラノビス距離の意味 Meaning of Mahalanobis Distance

$$\boldsymbol{u}_1 = (u_{11}, u_{12}) = \boldsymbol{x}_1(\boldsymbol{p}_1, \boldsymbol{p}_2) \quad \boldsymbol{u}_2 = (u_{21}, u_{22}) = \boldsymbol{x}_2(\boldsymbol{p}_1, \boldsymbol{p}_2)$$

$$\begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1\boldsymbol{p}_1 & \boldsymbol{x}_1\boldsymbol{p}_2 \\ \boldsymbol{x}_2\boldsymbol{p}_1 & \boldsymbol{x}_2\boldsymbol{p}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} (\boldsymbol{p}_1, \boldsymbol{p}_2)$$

$$U = \begin{bmatrix} u_{1,1} & u_{1,2} & \dots & u_{1,M} \\ u_{2,1} & u_{2,2} & \dots & u_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N,1} & u_{N,2} & \dots & u_{N,M} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} (\boldsymbol{p}_1, \boldsymbol{p}_2 \dots \boldsymbol{p}_M) = XP$$

マハラノビス距離の意味 Meaning of Mahalanobis Distance

$$U = \begin{bmatrix} u_{1,1} & u_{1,2} & \dots & u_{1,M} \\ u_{2,1} & u_{2,2} & \dots & u_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N,1} & u_{N,2} & \dots & u_{N,M} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} (\boldsymbol{p}_{1,} \boldsymbol{p}_2 \dots \boldsymbol{p}_M) = XP$$

Σ

Uの分散共分散行列は Variance-covariance matrix $\Sigma$of U is

$$\Sigma = \frac{1}{N}\boldsymbol{U^T U} = \frac{1}{N}(XP)^T XP = \frac{1}{N} P^T X^T XP = \boldsymbol{P^T\ VP} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \lambda_M \end{bmatrix}$$

マハラノビス距離の意味 Meaning of Mahalanobis Distance

固有ベクトルで線型変換した後、マハラノビス距離を計算する

Calculate Mahalanobis distance after linear transformation by eigen vectors

$$D_M(\boldsymbol{u}_i) = \sqrt{(\boldsymbol{u_i} - \overline{\boldsymbol{u}})\Sigma^{-1}(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T}$$

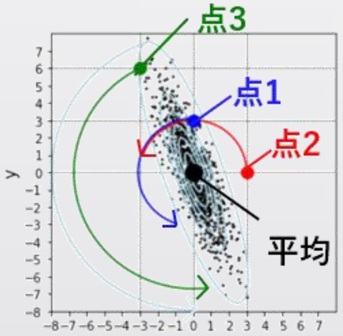
λ_iはデータの$\boldsymbol{p}_i$方向の分散と一致

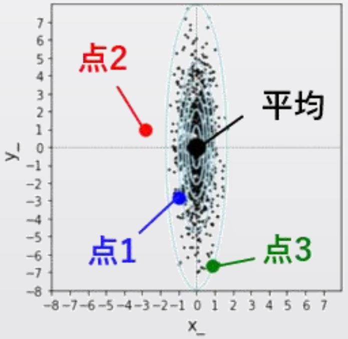
λ_i equals to variance of data along the direction of $\boldsymbol{p}_i$

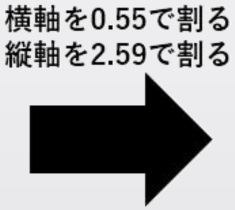
$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\lambda_1} & 0 & \ldots & 0 \\ 0 & \frac{1}{\lambda_2} & & \vdots \\ & & \ddots & 0 \\ \vdots & & & \\ 0 & \ldots & 0 & \frac{1}{\lambda_M} \end{bmatrix}$$

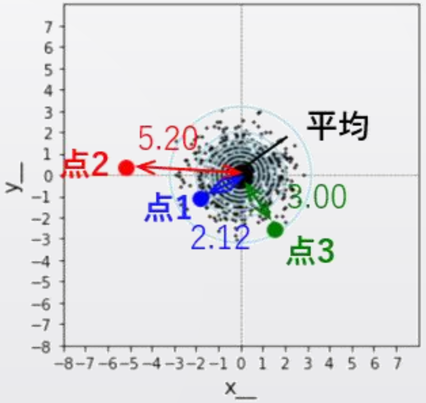
マハラノビス距離の意味 Meaning of Mahalanobis Distance

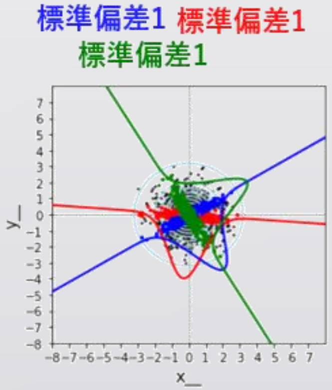
$$D_M(\boldsymbol{u}_i) = \sqrt{(\boldsymbol{u}_i - \overline{\boldsymbol{u}})\Sigma^{-1}(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T}$$











すべての方向の分散を一致させた後、ユークリッド距離を計算

Compute Euclidian distance after equalizing variance across all the directions

固有ベクトルで線型変換

Linear transformation by eigen vectors

https://qiita.com/yutera12/items/db425fafce2d87a25a1f

Jaccard 係数 Jaccard Coefficient

集合の類似度の指標 Measure of similarity between two sets

ベクトル間の類似度の指標としても用いることが出来る
Can be used as a measure of similarity between vectors

$$x = \{x_1, x_2 \dots x_n\} \quad y = \{y_1, y_2 \dots y_n\}$$

$$Jaccard(x, y) = \frac{|x \cap y|}{|x \cup y|} \qquad Jaccard(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}\boldsymbol{y}^{\boldsymbol{T}}}{\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}} + \boldsymbol{y}\boldsymbol{y}^{\boldsymbol{T}} - \boldsymbol{x}\boldsymbol{y}^{\boldsymbol{T}}}$$

Dice 係数 Dice Coefficient

集合の類似度の指標 Measure of similarity between two sets

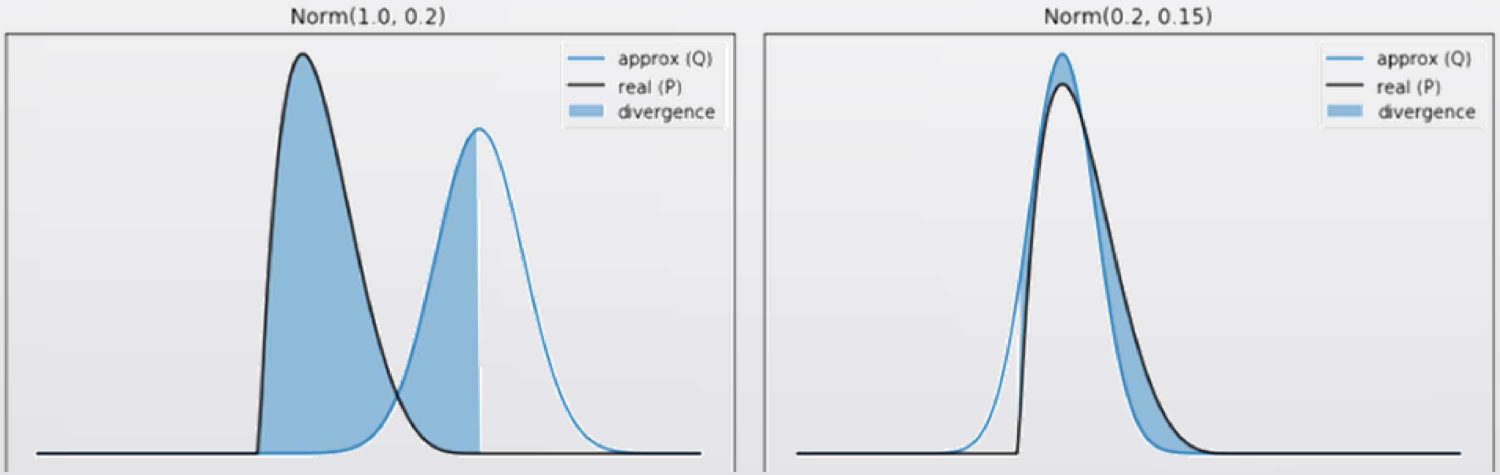
ベクトル間の類似度の指標としても用いることが出来る
Can be used as a measure of similarity between vectors

$$x = \{x_1, x_2 \dots x_n\} \quad y = \{y_1, y_2 \dots y_n\}$$

$$Dice(x, y) = \frac{2|x \cap y|}{|x| + |y|} \qquad Dice(\boldsymbol{x}, \boldsymbol{y}) = \frac{2\boldsymbol{x}\boldsymbol{y}^{\boldsymbol{T}}}{\|\boldsymbol{x}\| + \|\boldsymbol{y}\|}$$

KLダイバージェンス Kullback-Leibler Divergence

分布同士の類似度の評価指標 Measure of similarity between two distributions



$$KL(p(x)|q(x)) = \int p(x)\, log\left(\frac{p(x)}{q(x)}\right)$$

https://jessicastringham.net/2018/12/27/KL-Divergence/