データマイニング

Data Mining

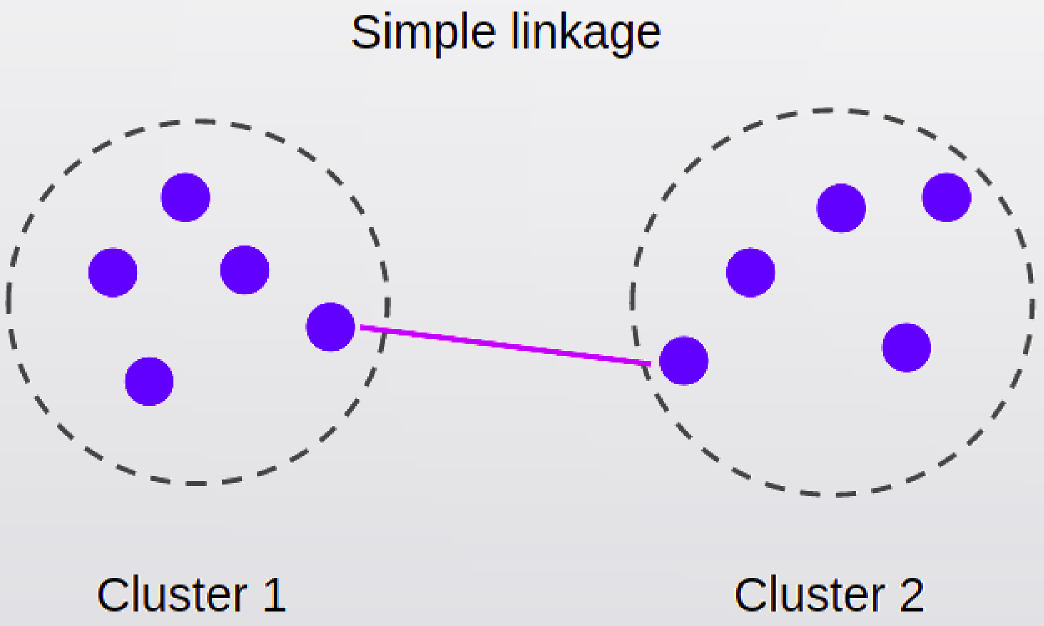
12: クラスタリング② Clustering

土居 裕和 Hirokazu Doi

長岡技術科学大学 Nagaoka University of Technology

クラスター間の距離 Distance between Clusters

単リンク法 Single Linkage



https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/

$$D(A,B) = min_{\boldsymbol{x}\in A,\boldsymbol{y}\in B}\, d(\boldsymbol{x},\boldsymbol{y})$$

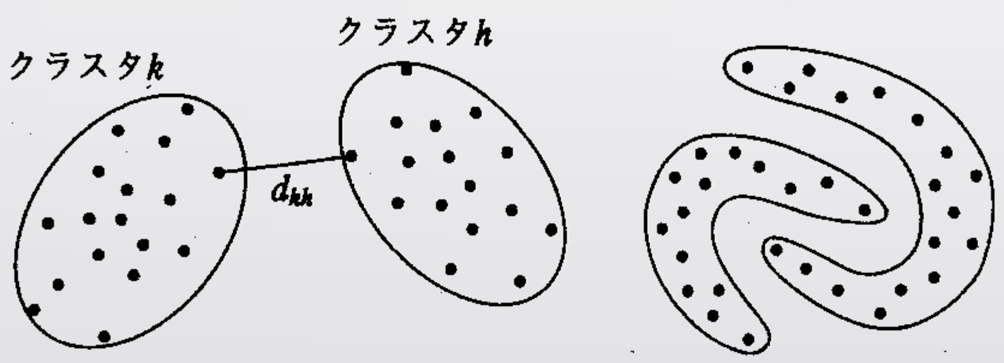
各クラスターのデータの内、最も近いデータ間の距離を、クラスター間の距離とする

Distance between clusters is defined as the distance between their closest members

単リンク法 Single Linkage

・大きなクラスターが形成されやすい　Large cluster is likely to be formed

・近いデータ同士が別のクラスターに含まれてしまう連鎖効果が起きやすい
Neighboring data points tend to be included in separate clusters (chain effect)



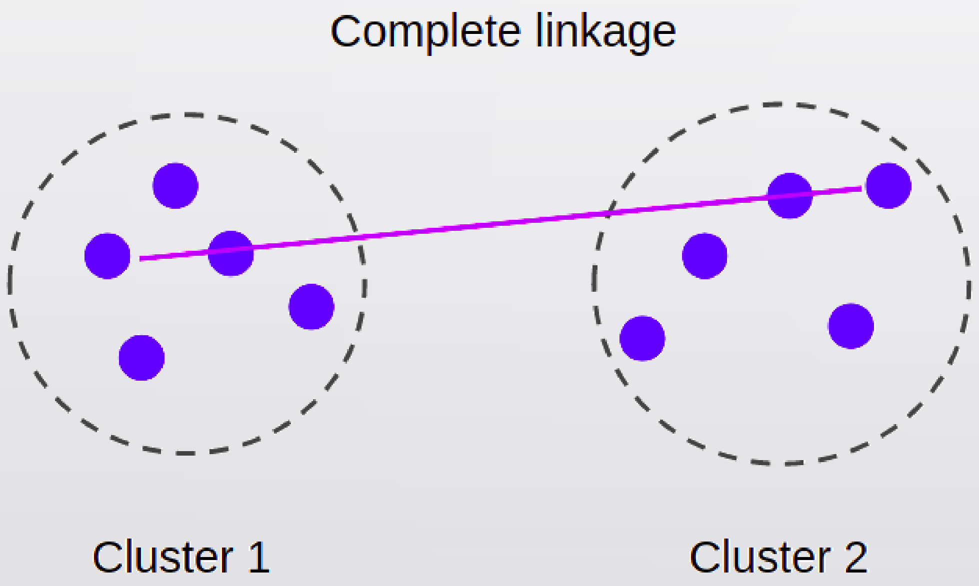
(a)　クラスタ間距離の定義　　(b)　連鎖効果の例

図1.2.5　最短距離法におけるクラスタ間距離の定義と連鎖効果

https://www.is.kochi-u.ac.jp/kyoko/edu/image/c.html

クラスター間の距離 Distance between Clusters

完全リンク法 Complete Linkage



$$D(A,B) = max_{\boldsymbol{x}\in A,\boldsymbol{y}\in B}\, d(\boldsymbol{x},\boldsymbol{y})$$

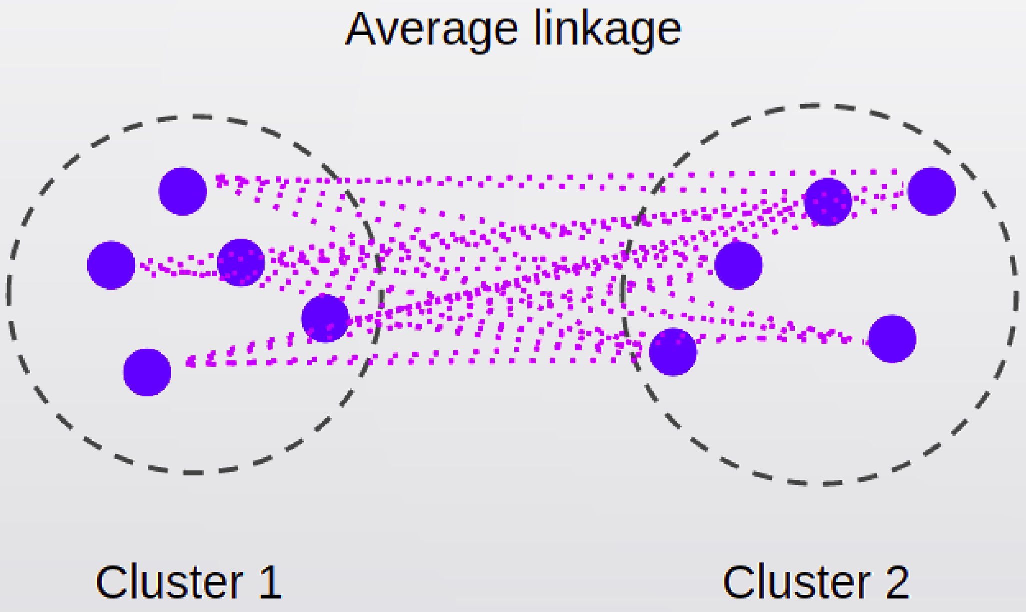
各クラスターのデータの内、最も遠いデータ間の距離を、クラスター間の距離とする

Distance between clusters is defined as the distance between their farthest members

https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/

クラスター間の距離 Distance between Clusters

平均リンク法 Average Linkage



$$D(A,B) = \frac{1}{N_A N_B} \sum_{\boldsymbol{x} \in A} \sum_{\boldsymbol{y} \in B} d(\boldsymbol{x}, \boldsymbol{y})$$

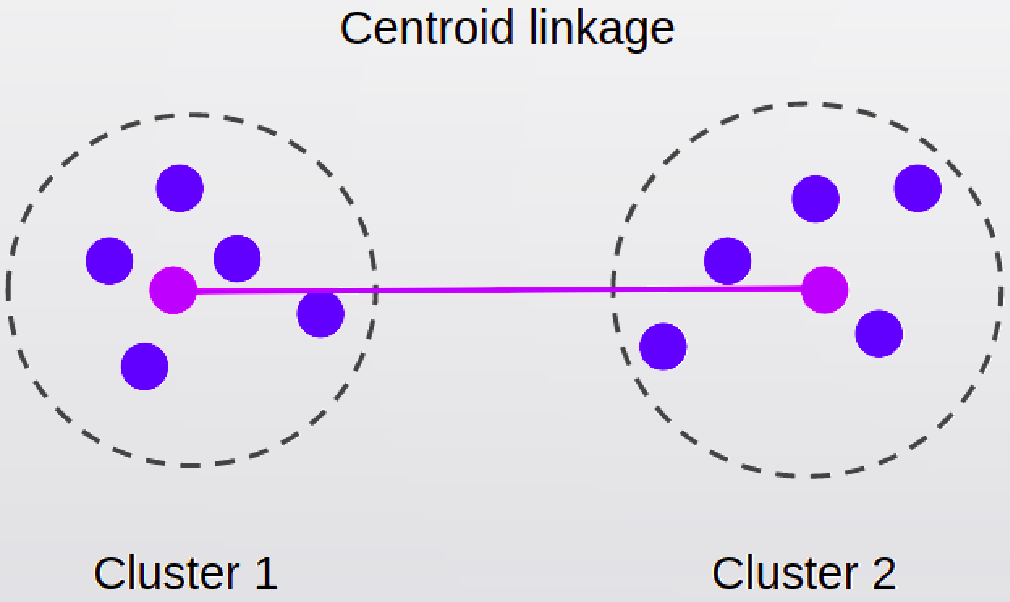
各クラスターのすべてのデータペアの平均

Distance between clusters is defined as average distance of all the between-cluster data pairs

https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/

クラスター間の距離 Distance between Clusters

中心リンク法 Centroid Linkage



$$D(A, B) = d(\boldsymbol{\mu}_A, \boldsymbol{\mu}_B)$$

各クラスターの中心間の距離

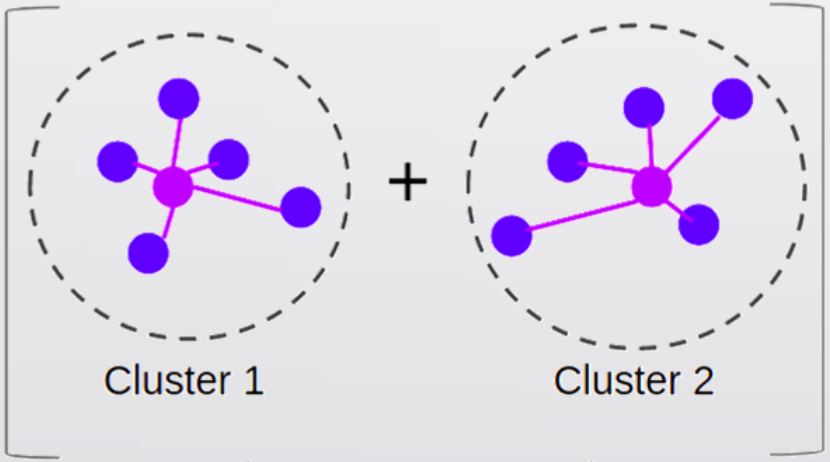
Distance between clusters is defined as the distance between cluster centers

https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/

ウォード法 Ward Linkage

Δが最小になるようなクラスター同士を結合する

Link clusters with minimum Δ



クラスター内SSEの合計を計算する

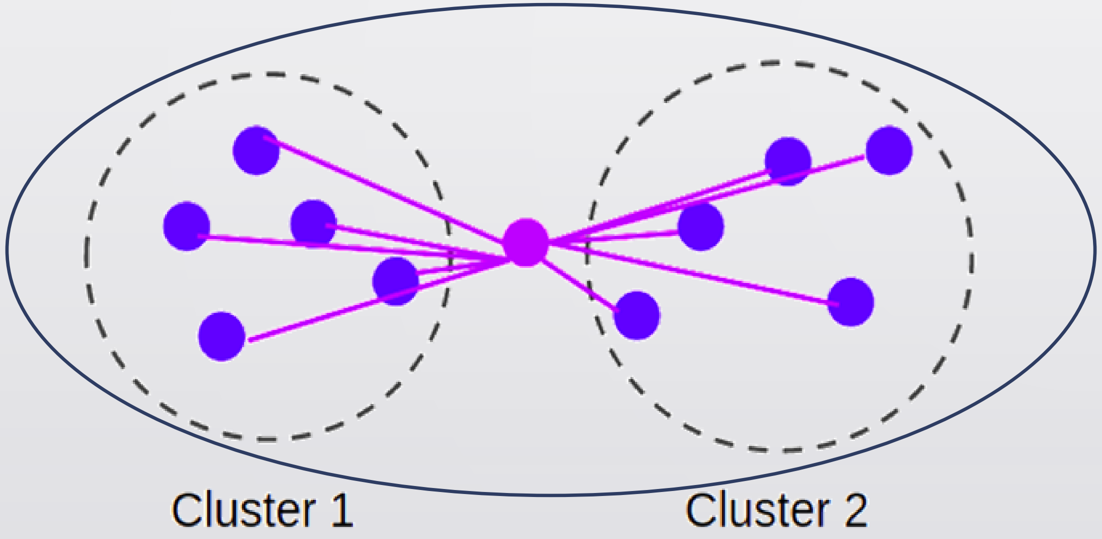
Compute sum of intra-cluster sum of squared error (SSE)

$$\sum_{\boldsymbol{x}\in A} d(\boldsymbol{x}, \boldsymbol{\mu_A})^2 + \sum_{\boldsymbol{y}\in B} d(\boldsymbol{y}, \boldsymbol{\mu_B})^2$$

ウォード法 Ward Linkage

Δが最小になるようなクラスター同士を結合する

Link clusters with minimum Δ



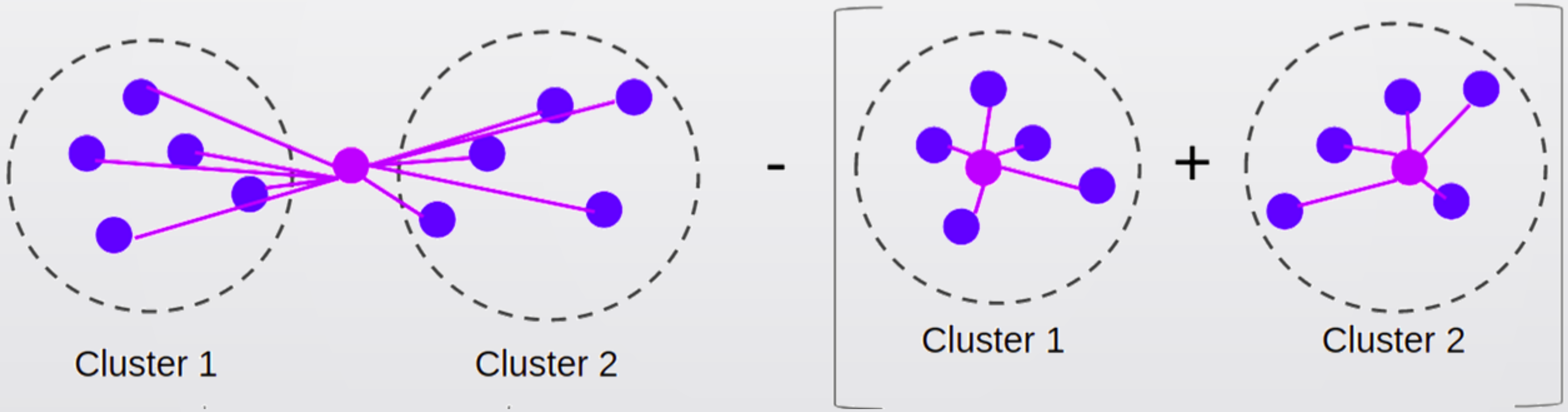
クラスターを結合した時のSSEを計算する

Compute within-cluster sum of squared error (SSE) when the two clusters are joined to form single cluster

$$\sum_{\boldsymbol{x} \in AB} d(\boldsymbol{x}, \boldsymbol{\mu_{AB}})^2$$

ウォード法 Ward Linkage

Δが最小になるようなクラスター同士を結合する
Link clusters with minimum Δ

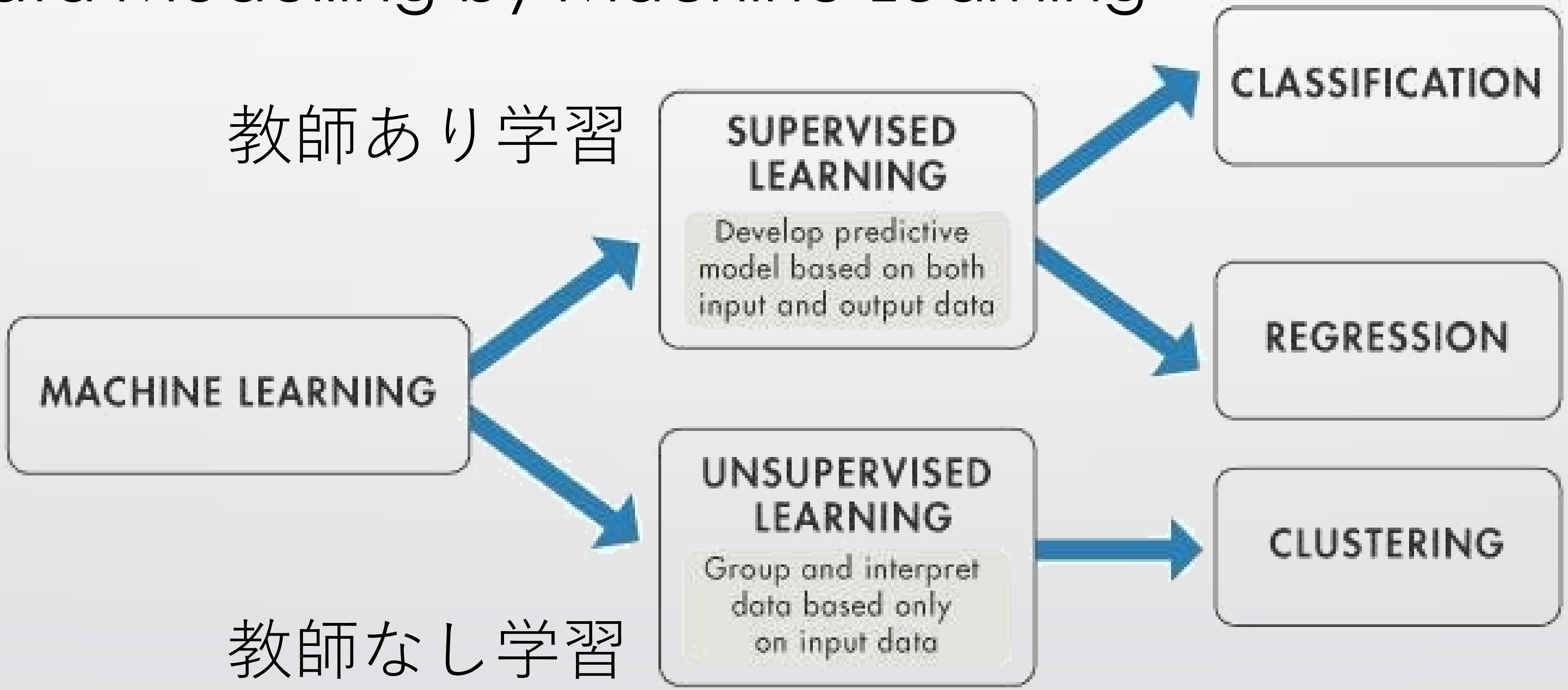


$$\Delta = SSE\ after\ linkage\ - SSE\ before\ linkage$$

Δを情報ロスと呼ぶ Δ is referred to as “Information loss”

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

ソフトクラスタリング Soft Clustering

ハードクラスタリング Hard Clustering

各データは一つのクラスターにしか所属できない

Each data belongs to single cluster

ソフトクラスタリング Soft Clustering

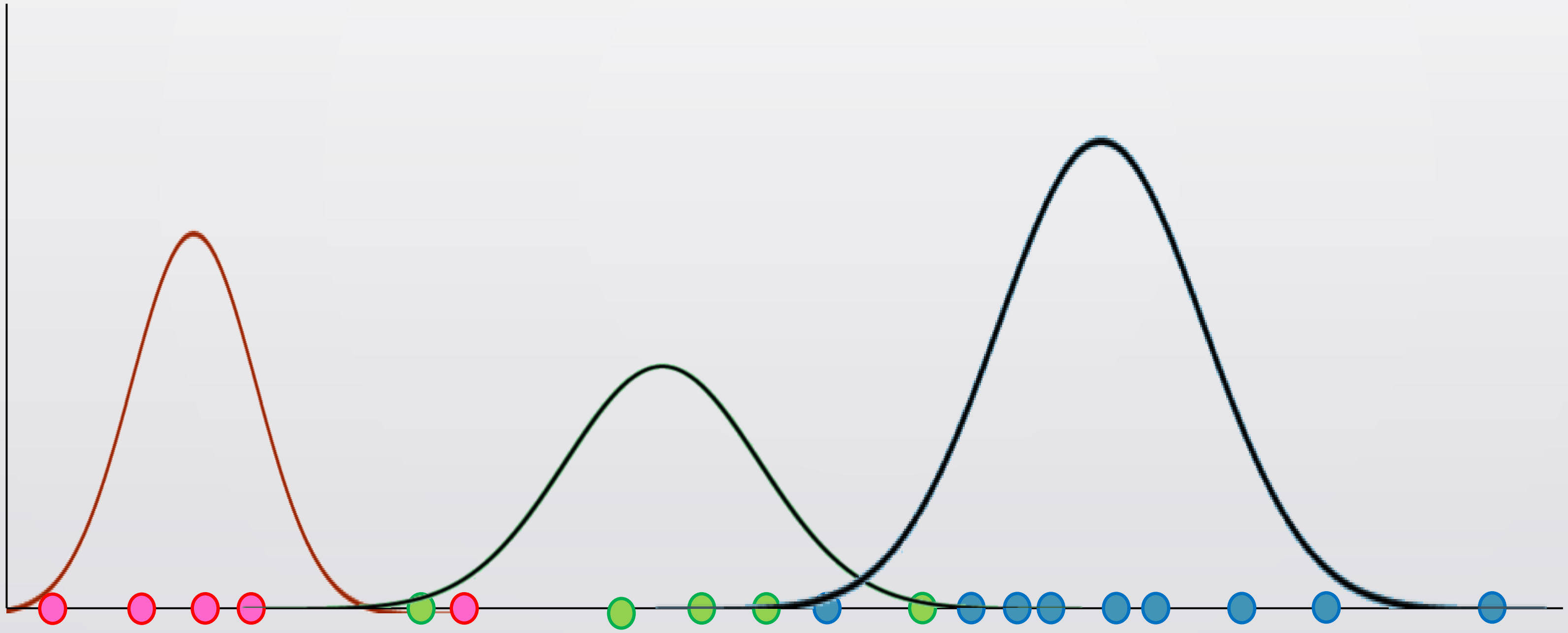
各データが複数のクラスに所属しうる

Each data can belong to multiple classes

混合ガウス分布モデル Gaussian Mixture Model

観測されたデータが複数のガウス分布の重ね合わせから生成されたと仮定する

Assume that observations are generated by multiple overlapping Gaussian distributions



多次元ガウス分布 Multidimensional Gaussian Distribution

正規分布を多次元に拡張した分布

Probability density distribution obtained by extending normal distribution to multi-dimensional space

$$N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

$\boldsymbol{\Sigma}$：分散共分散行列 Variance-covariance matrix

$|\boldsymbol{\Sigma}|$：分散共分散行列の行列式 Determinant of variance-covariance matrix

$\boldsymbol{\Sigma}^{-1}$: 分散共分散行列の逆行列 Inverse matrix of variance-covariance matrix

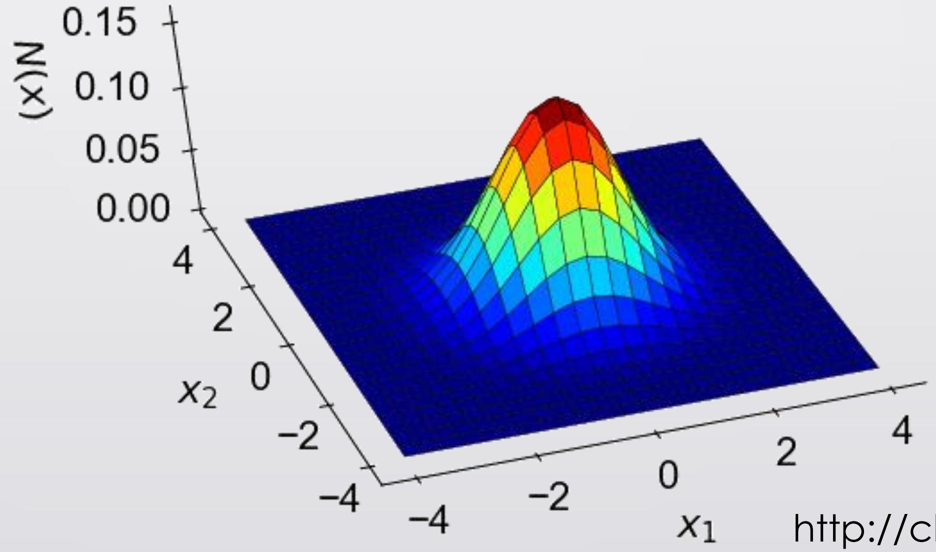
多次元ガウス分布 Multidimensional Gaussian Distribution

$$N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

$$\boldsymbol{x} = (x_1, x_2)$$

$$\boldsymbol{\mu} = (\mu_1, \mu_2)$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{pmatrix}$$



http://chasen.org/~daiti-m/gpbook/

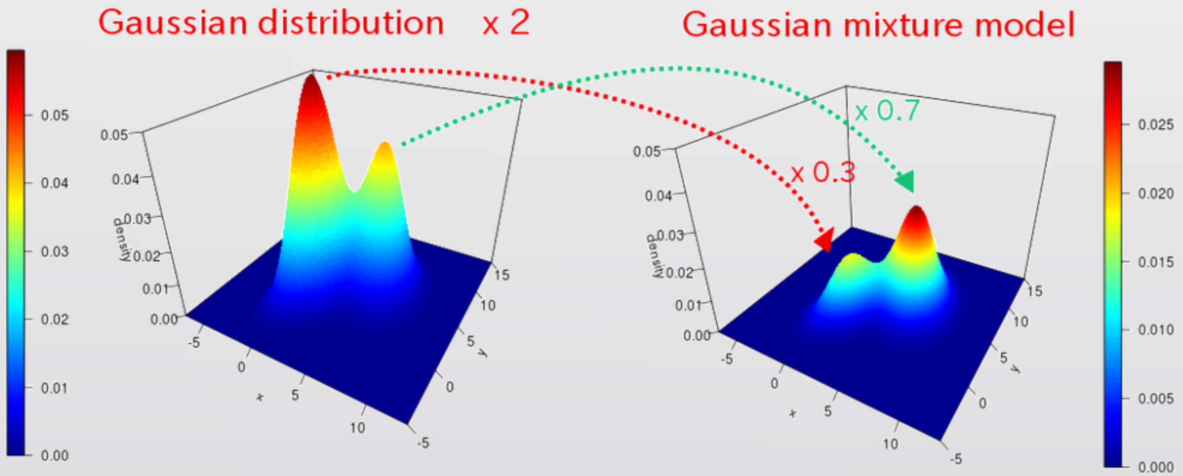
混合ガウス分布 Gaussian Mixture Distribution

M個の正規分布の重ね合わせにより確率分布を表現する

Represent probability distribution as weighted mixture of M normal distributions

$$p(\boldsymbol{x}) = \sum_{m=1}^{M} \pi_m N(\boldsymbol{x}|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m) \quad 0 \leq \pi_m \leq 1 \quad \sum_{m=1}^{M} \pi_m = 1$$

π_m：混合比 Mixing Ratio



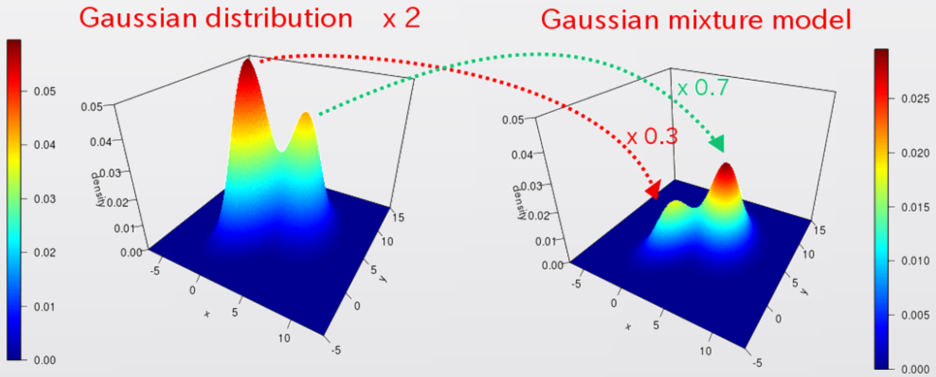
https://work-in-progress.hatenablog.com/entry/2018/11/08/224826

潜在変数 Latent Variable (隠れ変数 Hidden Variable)

観測データからは直接得ることが出来ない情報

Information that cannot be obtained directly from observations

$$p(\boldsymbol{x}) = \sum_{m=1} \pi_m N(\boldsymbol{x}|\boldsymbol{\mu}_{\boldsymbol{m}}, \boldsymbol{\sigma}_{\boldsymbol{m}})$$



潜在変数を使って、どの分布からデータが生成されたかを表現する

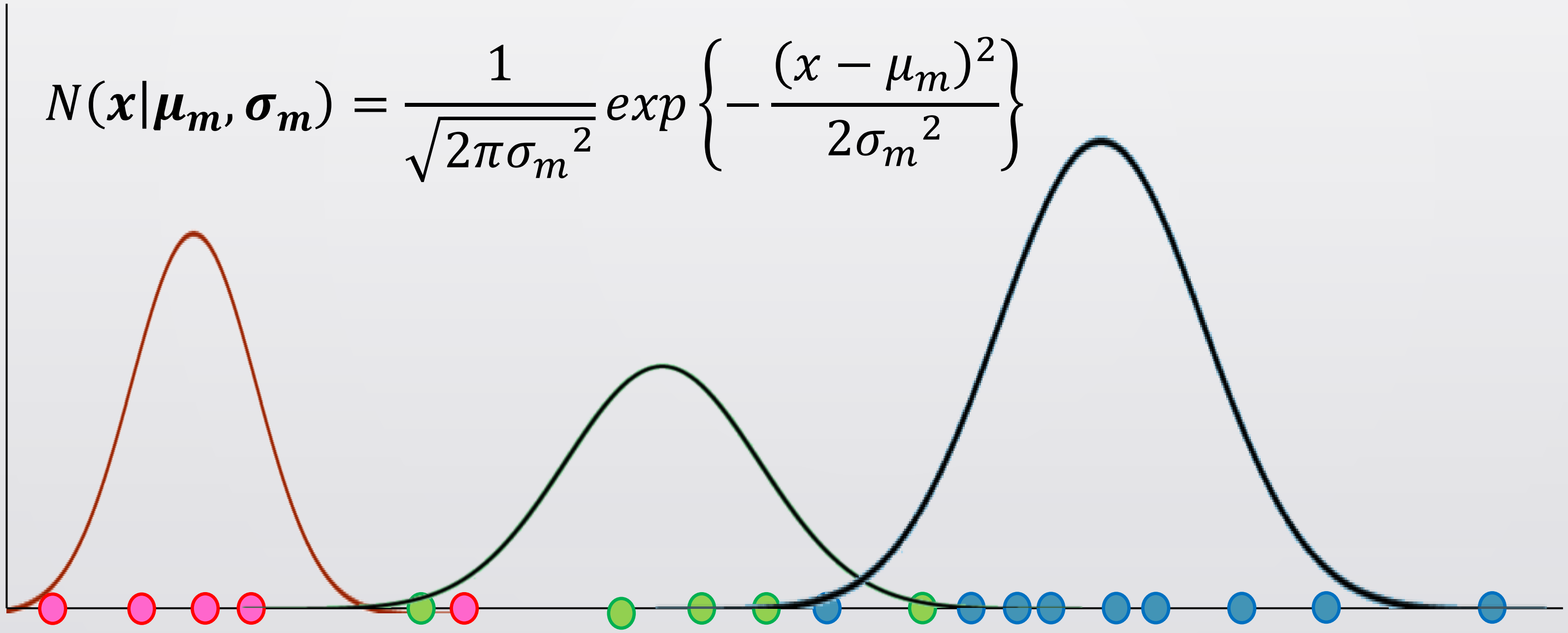
Latent variable represents from which gaussian distribution a data is generated¥

一次元混合ガウスモデル 1-Dimensional GMM

観測されたデータが複数のガウス分布の重ね合わせから生成されたと仮定する

Assume that observations are generated by multiple overlapping Gaussian distributions

$$N(\boldsymbol{x}|\boldsymbol{\mu_m}, \boldsymbol{\sigma_m}) = \frac{1}{\sqrt{2\pi{\sigma_m}^2}} exp\left\{-\frac{(x-\mu_m)^2}{2{\sigma_m}^2}\right\}$$



一次元混合ガウスモデル 1-Dimensional GMM

π_m：混合比　Mixing Ratio　$\sum_{m=1}^{M} \pi_m = 1 \quad 0 \leq \pi_m \leq 1$

$\boldsymbol{z}$：潜在変数 Latent Variables

$$z_m \in \{0, 1\} \quad \boldsymbol{z} = (z_1, z_2, z_3, z_4, \cdots, z_{M-1}, z_M)$$

$$\sum_{m=1}^{M} z_m = 1 \qquad \boldsymbol{ex})\ \boldsymbol{z} = (0, 0, 0, 1, \cdots, 0, 0, 0)$$

一次元混合ガウスモデル 1-Dimensional GMM

$$N(x|\mu_m, \sigma_m) = \frac{1}{\sqrt{2\pi{\sigma_m}^2}} exp\left\{-\frac{(x-\mu_m)^2}{2{\sigma_m}^2}\right\}$$

$$p(x) = \sum_{m=1}^{M} p(x|z_m = 1)p(z_m = 1) = \sum_{m=1}^{M} N(x|\mu_m, \sigma_m)\pi_m$$

一次元混合ガウスモデル 1-Dimensional GMM

$$p(z_m = 1|x) = \frac{p(x|z_m = 1)p(z_m = 1)}{p(x)} = \frac{p(x|z_m = 1)p(z_m = 1)}{\sum_{m=1}^{M} p(x|z_m = 1)\pi_m}$$

観測されたxが分布mから生成された事後確率

Posterior probability that observation x is generated by distribution m

潜在変数の期待値は、その事後確率と一致する

Expected value of latent variable corresponds to its posterior probability

$$E[z_m] = p(z_m = 1|x)$$

完全データ Complete Data

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2 \cdots \boldsymbol{x}_N\} \quad \boldsymbol{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2 \cdots \boldsymbol{z}_N\} \quad \boldsymbol{z}_n = (z_{n1}, z_{n2} \cdots z_{nM})$$

$$\boldsymbol{Y} = \{\boldsymbol{X}, \boldsymbol{Z}\} = \{\boldsymbol{x}_1, \boldsymbol{x}_2 \cdots \boldsymbol{x}_N, \boldsymbol{z}_1, \boldsymbol{z}_2 \cdots \boldsymbol{z}_N\}$$

$$p(\boldsymbol{x}_n, z_{nm} = 1 | \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi})$$

$$= p(\boldsymbol{x}_n | z_{nm} = 1) p(z_{nm} = 1 | \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi})$$

$$= N(\boldsymbol{x}_n | \mu_m, \sigma_m) \pi_m$$

MLEによるパラメータ推定Parameter Estimation by MLE

$$p(\boldsymbol{x}_n, z_{nm} = 1 | \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) = N(\boldsymbol{x}_n | \mu_m, \sigma_m) \pi_m$$

$$p(\boldsymbol{Y} | \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{m=1}^{M} [N(\boldsymbol{x}_n | \mu_m, \sigma_m) \pi_m]^{z_{nm}}$$

$$\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi} = argmax_{\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}}\, p(\boldsymbol{Y} | \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi})$$

観測される$\boldsymbol{X}$と対応する$\boldsymbol{Z}$の同時確率を最大化するパラメータセットを求める

Search for parameter set that maximizes observations $\boldsymbol{X}$ and corresponding $\boldsymbol{Z}$

MLEによるパラメータ推定Parameter Estimation by MLE

$$\log p(\boldsymbol{Y}|\,\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}\,) = \sum_{n=1}^{N}\sum_{m=1}^{M} {\color{red}z_{nm}}\, logN(\boldsymbol{x}_n|\mu_m,\sigma_m) + \sum_{n=1}^{N}\sum_{m=1}^{M} {\color{red}z_{nm}} log\pi_m$$

潜在変数は直接的に観察できない

Latent variables are not directly observable

潜在変数の期待値＝事後確率で置き換えてパラメータ推定

Estimate parameters by replacing latent variables with their expected values

Q関数 Q function

$$\log p(\boldsymbol{Y}|\,\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}\,) = \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\, logN(\boldsymbol{x}_n|\mu_m,\sigma_m) + \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm} log\pi_m$$

$$\Downarrow$$

$$Q = \sum_{n=1}^{N}\sum_{m=1}^{M} E[z_{nm}] \quad logN(\boldsymbol{x}_n|\mu_m,\sigma_m) + \sum_{n=1}^{N}\sum_{m=1}^{M} E[z_{nm}] \quad log\pi_m$$

EM アルゴリズム Expectation-Maximizing Algorithm

潜在変数を含むモデルの代表的なパラメータ推定法

Algorithm for parameter estimation of models including latent variables

E Step　θからzを計算

Compute z using θ

M Step　zとxでθを更新

Update θ based on z and x

x: 観測 Observations

θ: 確率密度関数のパラメータセット

Parameter set of probability distribution functions

z: 潜在変数 Latent Variables

E-ステップ Expectation-Step

現在のパラメータセット$\theta^{(t)}$を用いてzの期待値を求める

Compute expected value of z based on current parameter set $\theta^{(t)}$

$$\boldsymbol{\theta}^{(t)} = \left\{\boldsymbol{\mu}^{(t)}, \boldsymbol{\sigma}^{(t)}, \boldsymbol{\pi}^{(t)}\right\}$$

$$\boldsymbol{z} = (z_1, z_2 \cdots z_m \cdots z_M)$$

$$E[z_m] = \frac{N\left(x \middle| \mu_m{}^{(t)}, \sigma_m{}^{(t)}\right)\pi_m{}^{(t)}}{\sum_{m=1}^{M} N\left(x \middle| \mu_m{}^{(t)}, \sigma_m{}^{(t)}\right)\pi_m{}^{(t)}}$$

M-ステップ Maximization-Step

$$p\left(\boldsymbol{x},\boldsymbol{z}\middle|\boldsymbol{\theta}^{(t)}\right)=\prod_{n=1}^{N}\prod_{m=1}^{M}p\left(\boldsymbol{x}\middle|\boldsymbol{z},\boldsymbol{\theta}^{(t)}\right)p\left(\boldsymbol{z}\middle|\boldsymbol{\theta}^{(t)}\right)$$

$$=\prod_{n=1}^{N}\prod_{m=1}^{M}\left[p\left(x_n\middle|z_{n,m},\boldsymbol{\theta}^{(t)}\right)p\left(z_m\middle|\boldsymbol{\theta}^{(t)}\right)\right]^{z_{n,m}}$$

$$=\prod_{n=1}^{N}\prod_{m=1}^{M}\left[N\left(x_n\middle|{\mu_m}^{(t)},{\sigma_m}^{(t)}\right){\pi_m}^{(t)}\right]^{z_{n,m}}$$

M-ステップ Maximization-Step

Q関数を最大化するようパラメータセット$\theta^{(t)}$を更新

Update parameter set $\theta^{(t)}$so that Q function is maximized

$$p\left(\boldsymbol{x},\boldsymbol{z}\middle|\boldsymbol{\theta}^{(t)}\right)=\prod_{n=1}^{N}\prod_{m=1}^{M}\left[N\left(x_n\middle|{\mu_m}^{(t)},{\sigma_m}^{(t)}\right){\pi_m}^{(t)}\right]^{z_{n,m}}$$

$$Q\left(\boldsymbol{\theta}^{(t)}\right)=\sum_{n=1}^{N}\sum_{m=1}^{M}E\left[z_{n,m}\right]\quad\left[log\left({\pi_m}^{(t)}\right)+log\left(N\left(x_n\middle|{\mu_m}^{(t)},{\sigma_m}^{(t)}\right)\right)\right]$$

M-ステップ Maximization-Step

$$Q\left(\boldsymbol{\theta}^{(t)}\right) = \sum_{n=1}^{N}\sum_{m=1}^{M} E\left[z_{n,m}\right] \quad \left[log\left(\pi_m{}^{(t)}\right) + log\left(N\left(x_n \middle| \mu_m{}^{(t)}, \sigma_m{}^{(t)}\right)\right)\right]$$

$$E[z_m] = p(z_m = 1|x) = \frac{p(x|z_m = 1)p(z_m=1)}{\sum_{m=1}^{M} p(x|z_m = 1)\pi_m}$$

$$\frac{\partial Q}{\partial \pi_m} = 0 \qquad \frac{\partial Q}{\partial \mu_m} = 0 \qquad \frac{\partial Q}{\partial \sigma_m} = 0$$

EMアルゴリズム Expectation-Maximizing Algorithm

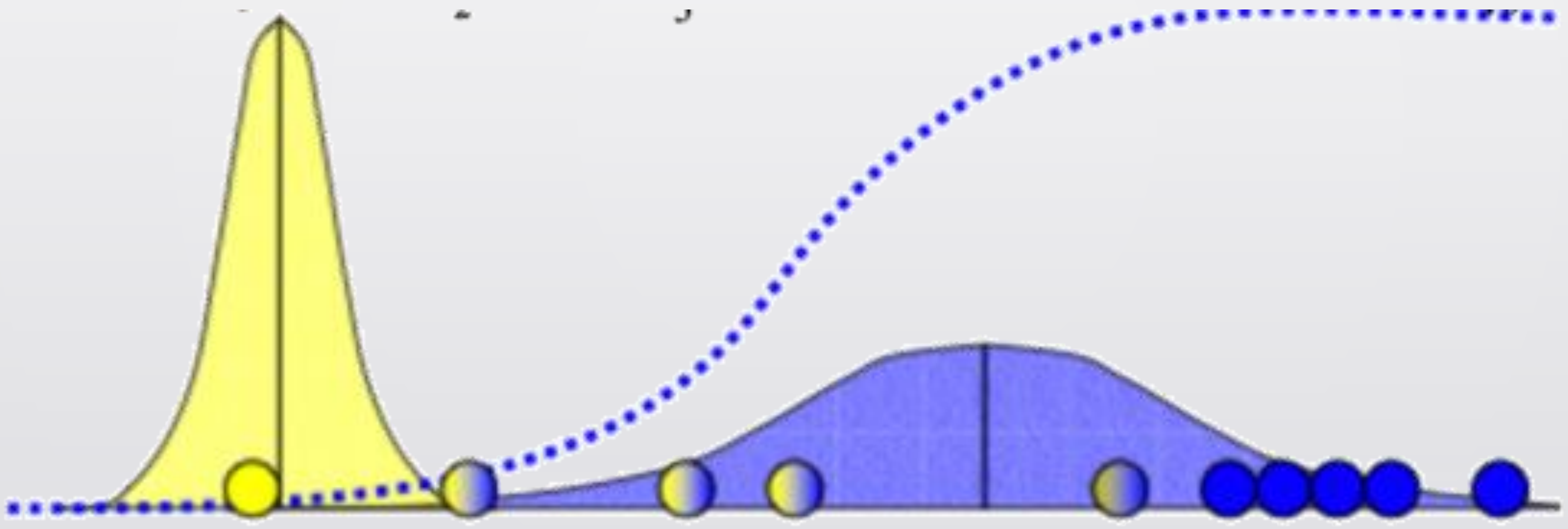
1.θを初期化する　Initialize θ

https://courses.cs.washington.edu/courses/cse416/22sp/lectures/12/12.pdf

EMアルゴリズム Expectation-Maximizing Algorithm

2. E-ステップ：現在のパラメータセット$\theta^{(t)}$を用いてzの期待値を求める

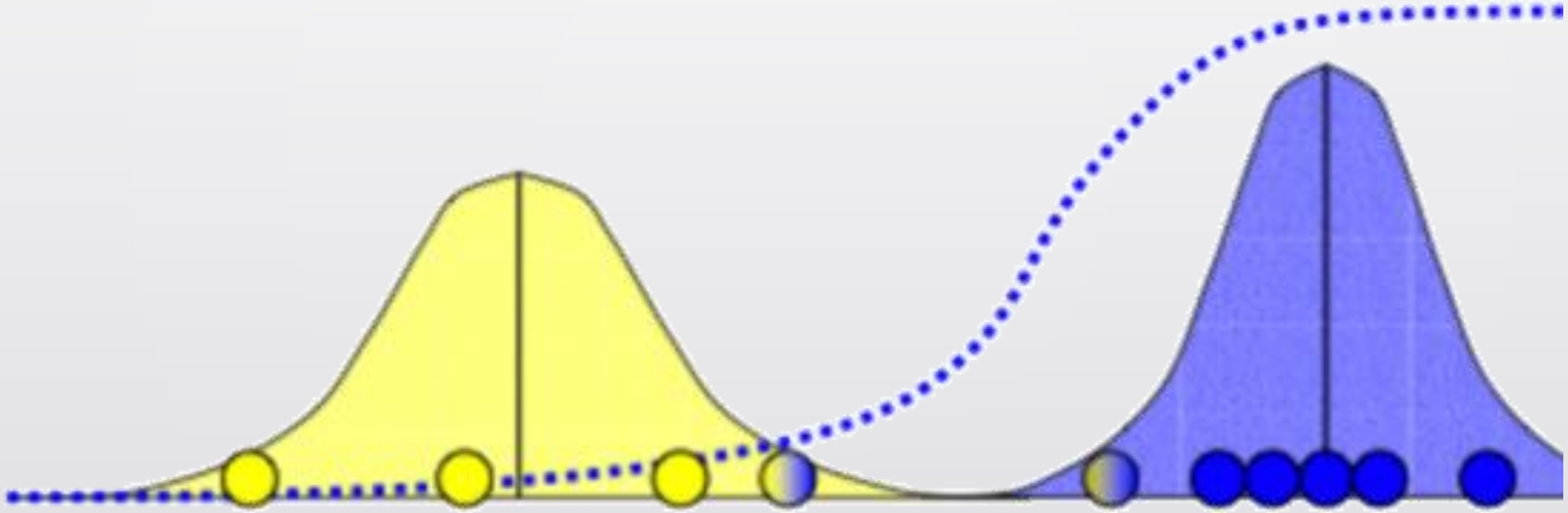
データの色が$E[z_{n,m}]$を表す



https://courses.cs.washington.edu/courses/cse416/22sp/lectures/12/12.pdf

EMアルゴリズム Expectation-Maximizing Algorithm

3. M-ステップ：パラメータセット$\theta^{(t)}$を更新する　Update θ



https://courses.cs.washington.edu/courses/cse416/22sp/lectures/12/12.pdf

クラスタリングの評価 Evaluation of Clustering Results

	C^1	C^2
a	90	10
b	20	80

C^k: k番目のクラスター k-th cluster

$n_{m,k}$: C^kのうちm番目のクラスC_mに属するデータの数
Number of data belonging to class C_m within C^k

$|C^k|$: クラスターC^kに属するデータの数
Number of data belonging to cluster C^k

純度 Purity

局所的純度 Local Purity

$$Purity = \frac{max_m n_{m,k}}{|C^k|}$$

最大多数派のクラスのデータがクラスターに占める割合

Proportion of data of majority class

大域的純度 Global Purity

$$Purity = \frac{\sum_k max_m n_{m,k}}{\sum_k |C^k|}$$

	C^1	C^2
a	90	10
b	20	80

逆純度 Inverse Purity

クラスターの純度は２つのテーブルで同じ

Purity of the clusters is the same across the tables below

	C^1	C^2
a	90	10
b	20	80

	C^1	C^2
a	90	40
b	20	5

$$Purity = \frac{max_m n_{m,k}}{|C^k|}$$

逆純度 Inverse Purity

$M_k = \ argmax_m n_{m,k}$ クラスターC^kにおいて最大多数派のクラス

Majority class within cluster C_k

$|C_{M_k}|$: クラスM_kに属するデータの総数

Total number of data belonging to class M_k

$$|C_{M_k}| = \sum_{k=1}^{K} n_{M_k,k}$$

逆純度 Inverse Purity

$$Inverse\ Purity = \frac{1}{N}\sum_{k=1}^{K}\frac{max_m n_{m,k}}{|C_{M_k}|}|C^k|$$

$M_1 = a$

	C^1	C^2
a	90	10
b	20	80

$M_2 = b$

	C^1	C^2
a	90	10
b	20	80

逆純度 Inverse Purity

$$Inverse\ Purity = \frac{1}{N}\sum_{k=1}^{K} \frac{max_m n_{m,k}}{|C_{M_k}|}|C^k|$$

$M_1 = a$

	C^1	C^2
a	90	40
b	20	5

$M_2 = a$

	C^1	C^2
a	90	40
b	20	5

F値 *F*-value

$$Purity = \frac{\sum_k max_m n_{m,k}}{\sum_k |C^k|} \qquad Inverse\ Purity = \frac{1}{N}\sum_{k=1}^{K} \frac{max_m n_{m,k}}{|C_{M_k}|}|C^k|$$

$$F = \frac{1}{\frac{1}{2}\left(\frac{1}{Purity} + \frac{1}{Inverse\ Purity}\right)} = \frac{2\ Purity \cdot Inverse\ Purity}{Purity + Inverse\ Purity}$$