クラスタリング

距離と類似度
階層型クラスタリング
K-means

クラスタリング（Clustering)とは？

- 教師データはない
- 性質が近い観測データ点をひとまとまり（これを「クラスタ」と呼ぶ）にまとめる
- 3つのキーポイント
  - 「性質が近い」ということを定量的に表すために観測データ点間の距離を定義する
  - 距離の近いものをどのようにまとめるかというアルゴリズム
  - いくつのまとまり（＝クラスタ）があるのか
    - ただし、クラスタ数はまとめるアルゴリズムと密接に関連
    - すると距離ないし類似度が大切なので、まずはそれから。

距離の定義

- 観測データ点を多次元空間中の点と定義
  - そこで2つの問題
  - 各次元は観測データ点からどのように定義するか
    - 次元のことを**feature**あるいは**素性**（そせい）と呼ぶ
    - この問題をfeature design：素性設計と呼ぶ。例えば、
      - 2つの素性の比を新たな素性とする ex 身長/体重
      - 2つの素性の連続したもの　ex　日本・銀行、　日本・沈没、
    - しかし、これは個別適用分野に応じて工夫すべし。
  - 多次元空間における2点間の距離の定義
    - ユークリッド距離ばかりではないのだ！

距離あるいは類似度の定義

- w_iをデータ点を表す素性のベクトルとする。
  - ex. あるテキストwにおいて語彙番号1の「日本」が3回、語彙番号2の「米国」が2回、語彙番号3の「中国」が1回,... 出現したとき、wの素性ベクトル: $w_i = (3,2,1,\ldots)^{\mathrm{T}}$
- w_iの第j成分w_{ij}としてTF*IDF(i,j)を使うこと有
- 距離で定義する場合と、その逆の類似度で定義する場合がある。類似度の場合は、最大値＝1、最小値＝0とすることが多い。
- いくつかの定義を次のページ以降に記す。

距離あるいは類似度の定義　1

cosine　(類似度)　$cosine(w_i, w_j) = \frac{w_i \cdot w_j}{\|w_i\| \cdot \|w_j\|}$

$w_i = (w_{i1}, w_{i2}, ..., w_{iN})$と書くと　$w_i \cdot w_j = \sum_{k=1}^{N} w_{ik} \cdot w_{jk}$(内積), $\|w_i\| = \left(\sum_{k=1}^{N} w_{ik}^{\ 2} \right)^{1/2}$

i.e.TF*IDF(i,N)

Jaccard係数（類似度）　$Jaccard(w_i, w_j) = \frac{|w_i \cap w_j|}{|w_i \cup w_j|}$

*Dice*係数（類似度）　$Dice(w_i, w_j) = \frac{2w_i \cdot w_j}{\|w_i\| + \|w_j\|}$

ユークリッド距離　$Dist(w_i, w_j) = \|w_i - w_j\| = \left(\sum_{k=1}^{N} (w_{ik} - w_{jk})^2 \right)^{1/2}$

$KL-divergence$(擬距離)　$KL(w_i \| w_j)$

w_{ik}がk番目の成分の生起確率だと思えば、KLが定義できる。

いよいよ距離の話に進むのだが、その前に情報理論の基礎概念を復習

- エントロピー: $H(x) = -\sum_{x} P(x) \log P(x)$
- 結合エントロピー: $H(x,y) = -\sum_{x}\sum_{y} P(x,y) \log P(x,y)$
- 条件付エントロピー: $H(y|x) = -\sum_{x}\sum_{y} P(x,y) \log P(y|x)$
- $$H(x,y) = -\sum_{x}\sum_{y} P(x,y) \log P(x,y) = -\sum_{x}\sum_{y} P(x,y) \log P(x|y)P(y)$$
  $$= -\sum_{x}\sum_{y} P(x,y) \log P(x|y) - \sum_{x}\sum_{y} P(x,y) \log P(y)$$
  $$= H(x|y) + H(y)$$

KL divergence: 情報理論による擬距離

- 相対エントロピー　or Kullback-Leibler divergence or KL divergence: KL(P||Q) : 分布PとQの類似性を測る尺度

$$KL(P||Q) = \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)}$$

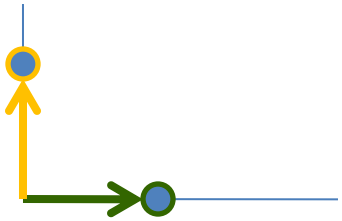
  - KL(P||P)=0
  - KL(P||Q)≠KL(Q||P)
    - 非対称なので擬距離
    - 対称性を持たせるために SymmetricKL(P||Q)=(KL(P||Q)+KL(Q||P))/2　という尺度もある
- 相互情報量 :

$$I(x,y) = KL(P(x,y) \| P(x)P(y)) = \sum P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

この部分をpointwise mutual informationとして使うこともある

距離の計算例

- x=(1,0), y=(0,1)の場合



類似度　$\cos(x,y)=\dfrac{1\cdot 0+0\cdot 1}{\sqrt{1^2+0}\sqrt{0+1^2}}=0$

類似度　$Jaccard(x,y)=\dfrac{|(1,0)\cap(0,1)|}{|(1,0)\cup(0,1)|}=\dfrac{0}{2}=0$

類似度　$Dice(x,y)=\dfrac{2\cdot(1\cdot 0+0\cdot 1)}{\sqrt{1^2+0}+\sqrt{0+1^2}}=0$

距離　$Dist(x,y)=\left((1-0)^2+(0-1)^2\right)^{1/2}=\sqrt{2}$

（擬）距離　$KL(x\,\|\,y)=1\cdot\log\dfrac{1}{0}+0\cdot\log\dfrac{0}{1}=\infty$

距離の計算例

- x=(1,1), y=(2,2)の場合

$$\cos(x,y)=\frac{1\cdot 2+1\cdot 2}{\sqrt{1^2+1^2}\sqrt{2^2+2^2}}=\frac{4}{\sqrt{2}\sqrt{8}}=1$$

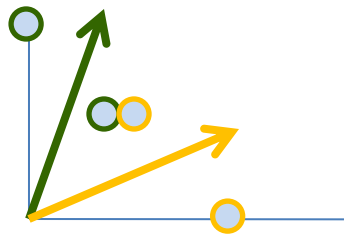
$$Jaccard(x,y)=\frac{|(1,1)\cap(2,2)|}{|(1,1)\cup(2,2)|}=\frac{2}{2}=1$$

$$Dice(x,y)=\frac{2\cdot(1\cdot 2+1\cdot 2)}{\sqrt{1^2+1^2}+\sqrt{2^2+2^2}}=\frac{8}{3\sqrt{2}}=1.88$$

$$Dist(x,y)=\left((2-1)^2+(2-1)^2\right)^{1/2}=\sqrt{2}$$

$KL(x\,\|\,y)=0$　　x, yは確率分布として正規化すれば同一分布

距離の計算例



- x=(1,2), y=(2,1)の場合

$$\cos(x,y)=\frac{1\cdot 2+1\cdot 2}{\sqrt{1^2+2^2}\sqrt{2^2+1^2}}=\frac{4}{5}$$

$$Jaccard(x,y)=\frac{|(1,2)\cap(2,1)|}{|(1,2)\cup(2,1)|}=\frac{2}{2}=1$$

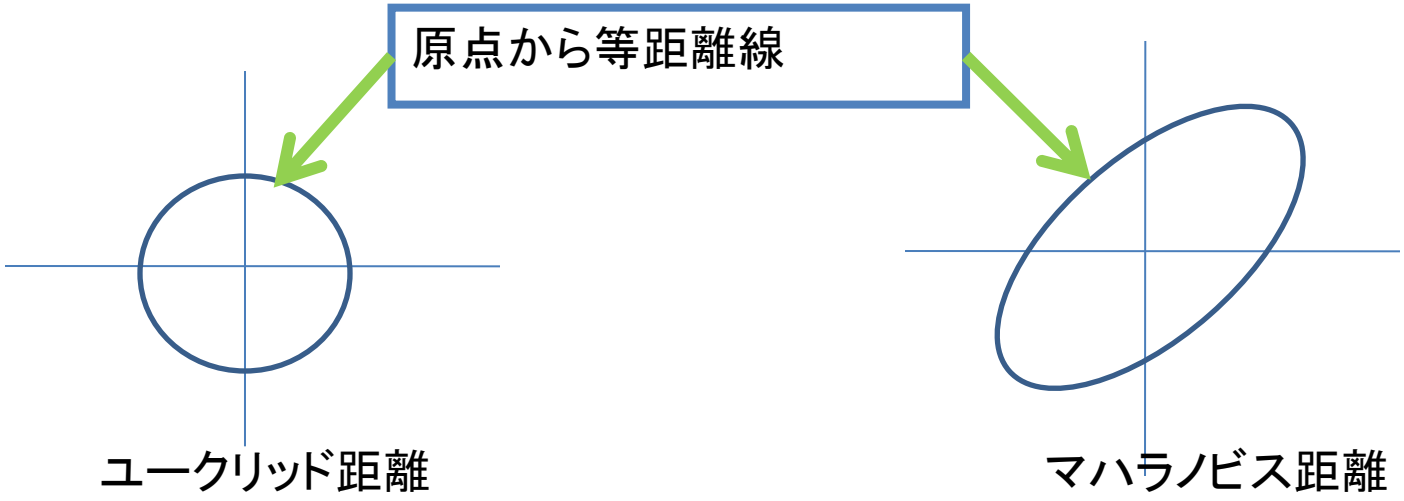
$$Dice(x,y)=\frac{2\cdot(1\cdot 2+1\cdot 2)}{\sqrt{1^2+2^2}+\sqrt{2^2+1^2}}=\frac{8}{2\sqrt{5}}=0.35$$

$$Dist(x,y)=\left((2-1)^2+(2-1)^2\right)^{1/2}=\sqrt{2}$$

$$KL(x\,\|\,y)=\frac{1}{3}\log\frac{1/3}{2/3}+\frac{2}{3}\log\frac{2/3}{1/3}=\frac{1}{3}\log\frac{1}{2}+\frac{2}{3}\log\frac{2}{1}=\frac{1}{3}\log 2=2.82$$

距離の定義　２－１

- 距離の定義１のスライドで示した距離、類似度はすべての次元の数値を平等に扱っていた。
- ただし、次元の間に相関がある場合はよい距離ではない。
- この場合に対応するのがマハラノビス距離
  - 図に示すように各次元の方向毎に異なるスケーリング＋回転



距離の定義　２－２

- マハラノビス距離d_Aの楕円の等距離線を実現するには次式の正定値行列Aを用いる

$$d_A(w_i, w_j) = (w_i - w_j)^T A (w_i - w_j) = (w_{i1} - w_{j1}, \ldots, w_{iN} - w_{iN}) A \begin{pmatrix} w_{i1} - w_{j1} \\ \vdots \\ w_{iN} - w_{iN} \end{pmatrix}$$

- Aが単位行列ならユークリッド距離の2乗
- Aが対角行列なら、次元毎のスケールが異なる
- Aが非対角行列なら回転も加わる

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} \frac{2}{\sqrt{2}} & -\frac{2}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

距離の定義　２－３

- Aは各次元のデータの分散行列:covの逆行列

$$d_A(w_i, w_j) = (w_i - w_j)^T A(w_i - w_j) = (w_{i1} - w_{j1}, \ldots, w_{iN} - w_{iN}) A \begin{pmatrix} w_{i1} - w_{j1} \\ \vdots \\ w_{iN} - w_{iN} \end{pmatrix}$$

Kを観測データ点の数とすると

$$A^{-1} = \mathrm{cov}(x_1 \cdots x_N) = E_{w_1 .. w_K}\left[\begin{pmatrix} x_1 - \overline{x_1} \\ \vdots \\ x_N - \overline{x_N} \end{pmatrix} (x_1 - \overline{x_1}, \ldots, x_N - \overline{x_N}) \right]$$

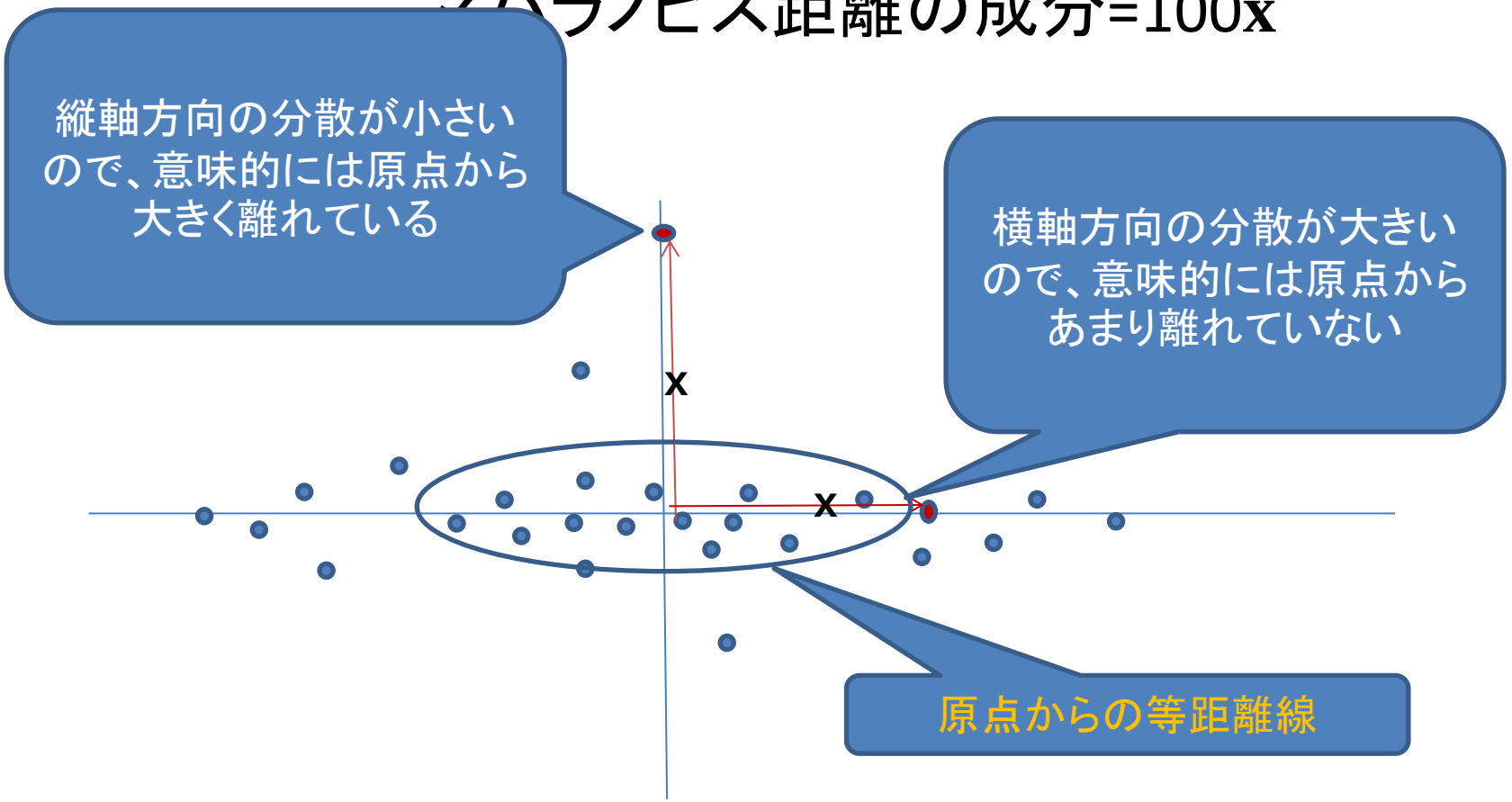
マラハノビス距離の直観的説明

下図で横軸方向の分散＝100→A_{11}=0.01

マハラノビス距離の成分=0.01x

縦軸方向の分散＝0.01　→ A_{22}=100

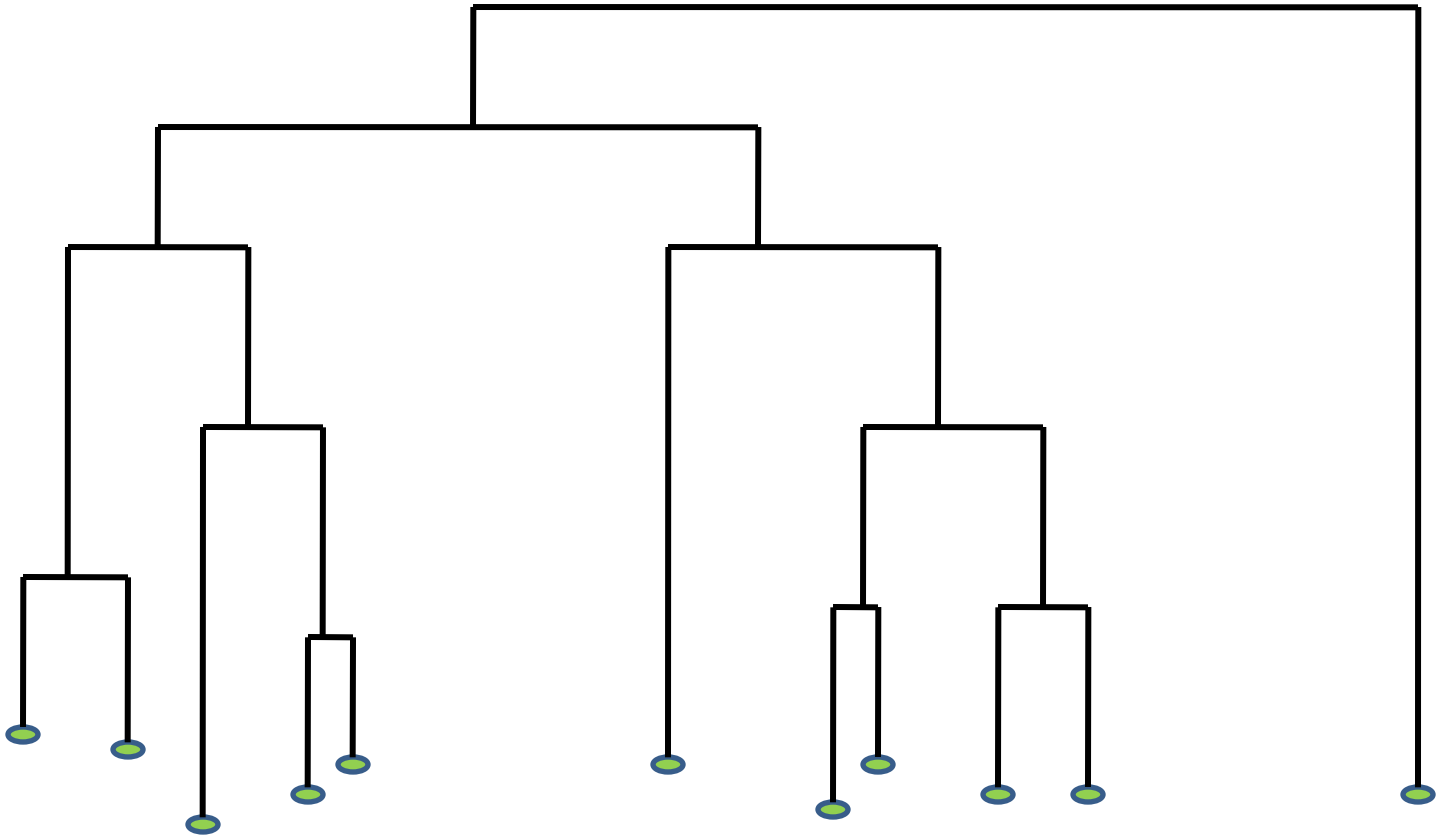
マハラノビス距離の成分=100x



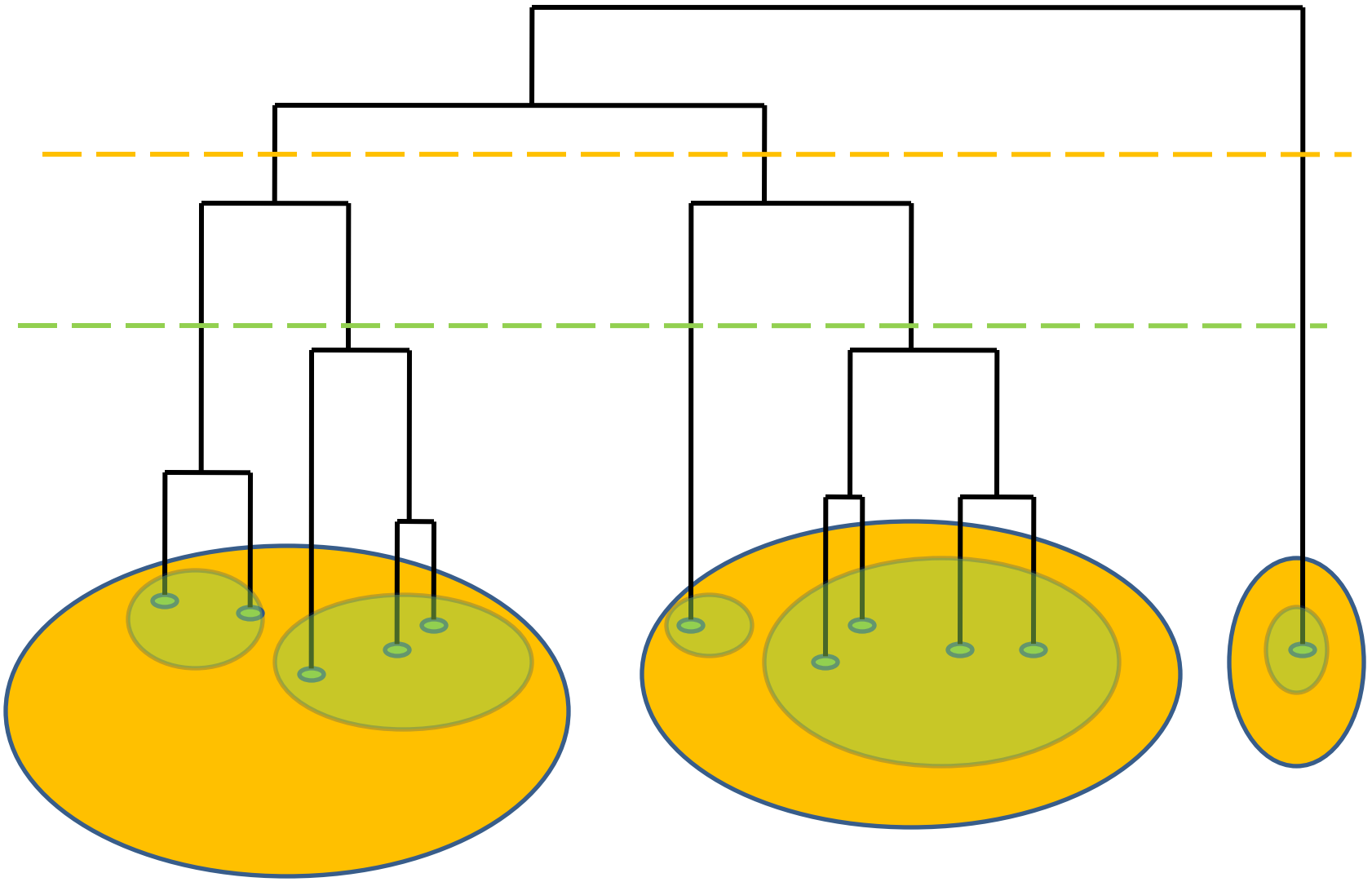
まとめるアルゴリズム

- データ点間の距離を利用してまとめあげるアルゴリズムをクラスタリングアルゴリズムと呼ぶことが多い。ここでは以下に2種について説明する。
- 階層型
- K-means

階層型クラスタリングアルゴリズムの概念：Dendrogram
(Hierarchical Agglomerative Clustering Algorithm: HAC)



階層型クラスタリングアルゴリズム：閾値で切るという直感的意味
(Hierarchical Agglomerative Clustering Algorithm: HAC)



Dendrogram形成方法が鍵

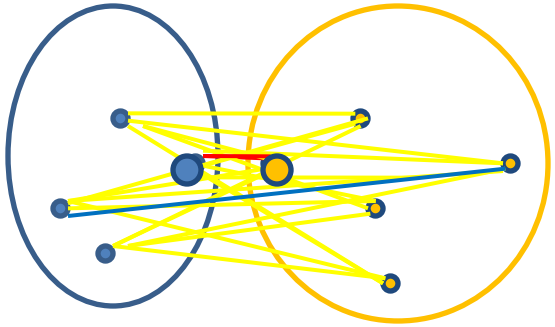
- 一番下のデータ点C_iとC_jをつなぐときは、距離$d(C_i,C_j)$が一番小さいものをつなぐ
- つながれたデータ点集合は新たなひとつの擬データ点C_kとみなす。
- （擬）データ点同士をつなぐときに閾値θにより次のように条件付ける
  - if $\min_{i,j} d(C_i,C_j) < \theta$ then new $C_k = C_i \cup C_j$
    otherwise C_i, C_j are result clusters
  - 図では高さの閾値のように書いたが、実は上記のように$d(C_i,C_j)$の閾値。θが大きいほど大きなクラスタが作られる。
- 距離$d(C_i,C_j)$の定義によっていくつかのvariationがある。

距離の定義

$$C_i = \{C_i(k), k = 1,2,..\}$$

Single link method

$$d(C_i, C_j) = \min_{l,m} \{d(C_i(l), C_j(m))\}$$



Complete link method

$$d(C_i, C_j) = \max_{l,m} \{d(C_i(l), C_j(m))\}$$

Group average link method

$$d(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_l \sum_m d(C_i(l), C_j(m))$$

- Single linkはまばらな内容のクラスタ、とくに鎖状につながったクラスタを作りやすい
- Complete link は最初は凝縮したクラスタを作りがち。最後のころに離れ離れのクラスタを無理やりつなぐ傾向あり
- Group average は、その中庸をとる。

K-means法

- 前述の階層型クラスタリングでは、閾値θで、クラスタのまとまりの良さを決めていた。しかし、クラスタ数そのものを制御できない。
- 一般的には、もともといくつのクラスタがあるか分からない。
- もし、クラスタ数がK個だと分かっているとき使えるのが、K-means法。
  - 後に述べるEMアルゴリズムの原始的な形態。

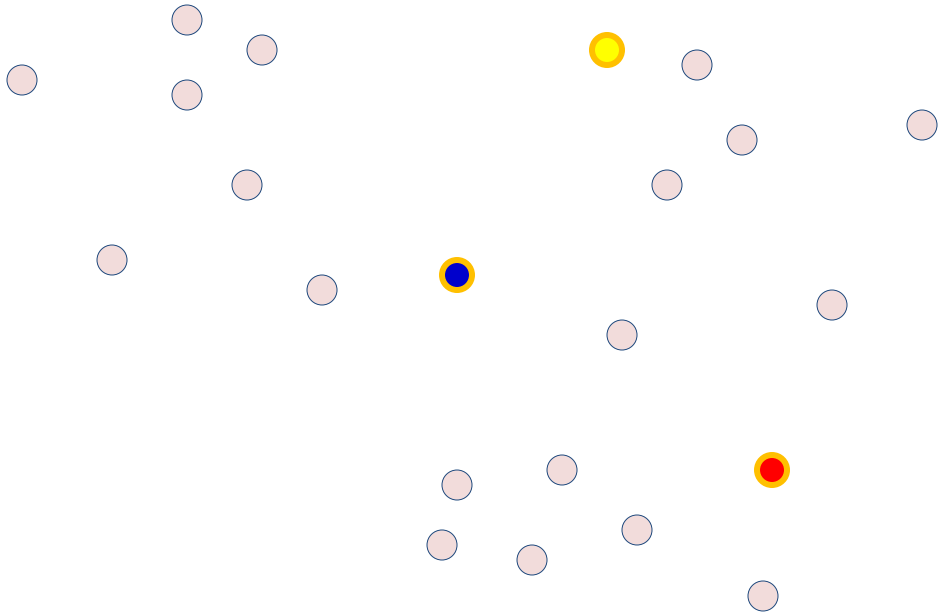
K-means　のアルゴリズム

1. データ:x[n]はN個
2. K個の点の初期値：m[k]を適当な値に決める
3. for (var n = 1; n <= N; n++)

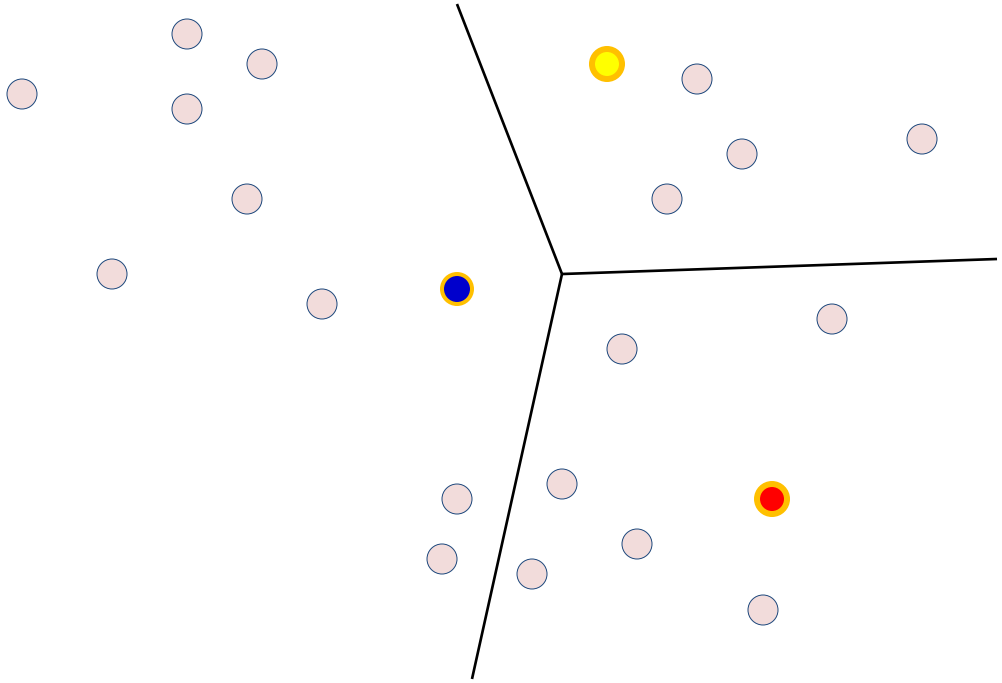
   { x[n]の所属先を　距離(x[n] , m[k])の一番小さい k　とする}
4. m[k]を、kに所属するx[n]の平均値に更新
5. 収束のテスト

   　例えば、全てのkで更新前m[k]と更新後m[k]の距離が予め決めた閾値εより小さい

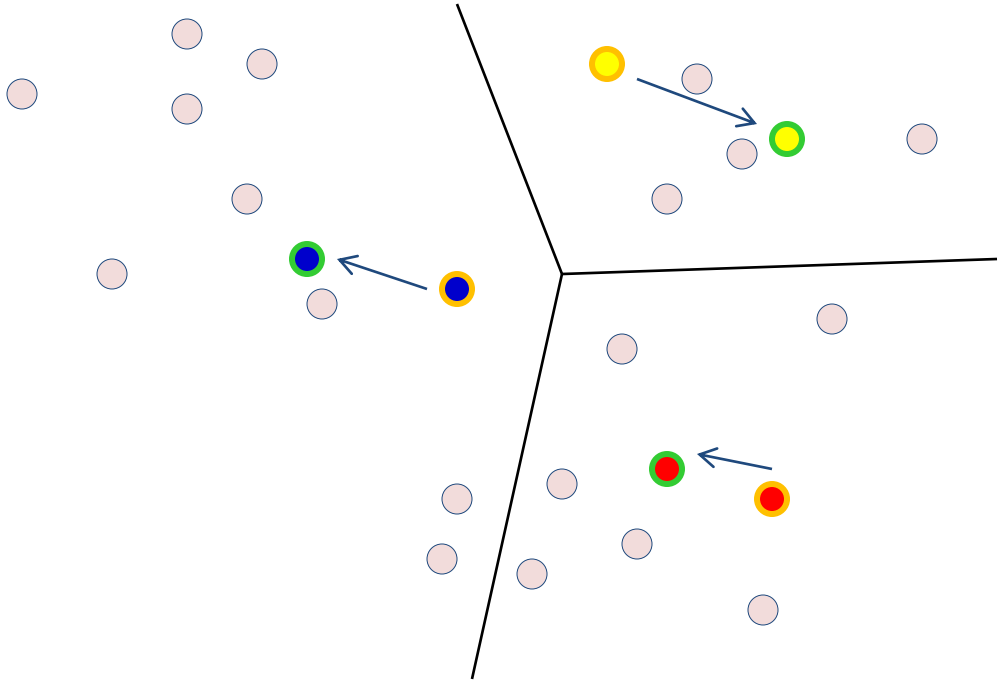
6. 収束したら終了。収束しなかったら3.へ戻る



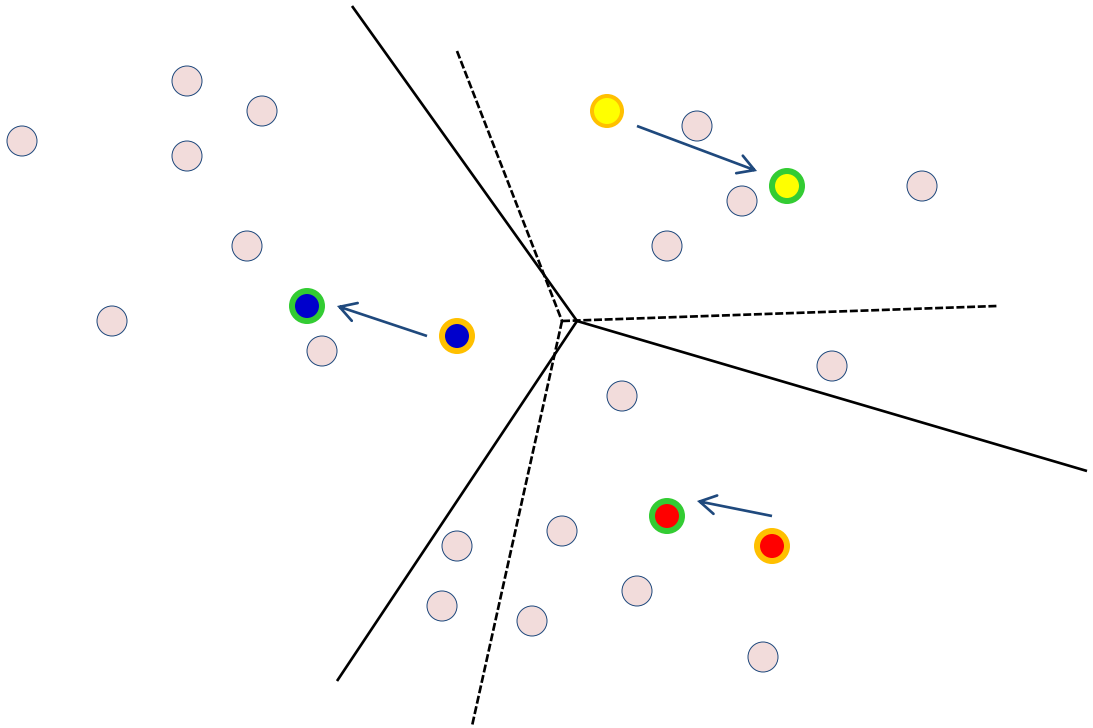
初期化



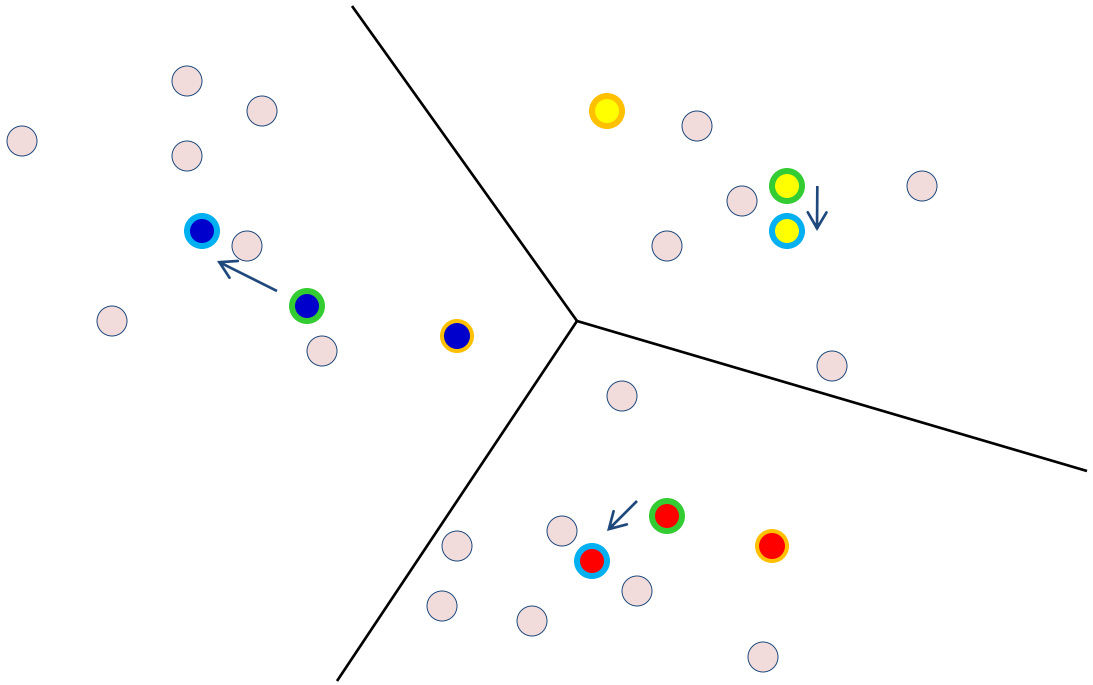
各データの所属を一番m[k]に近い決める



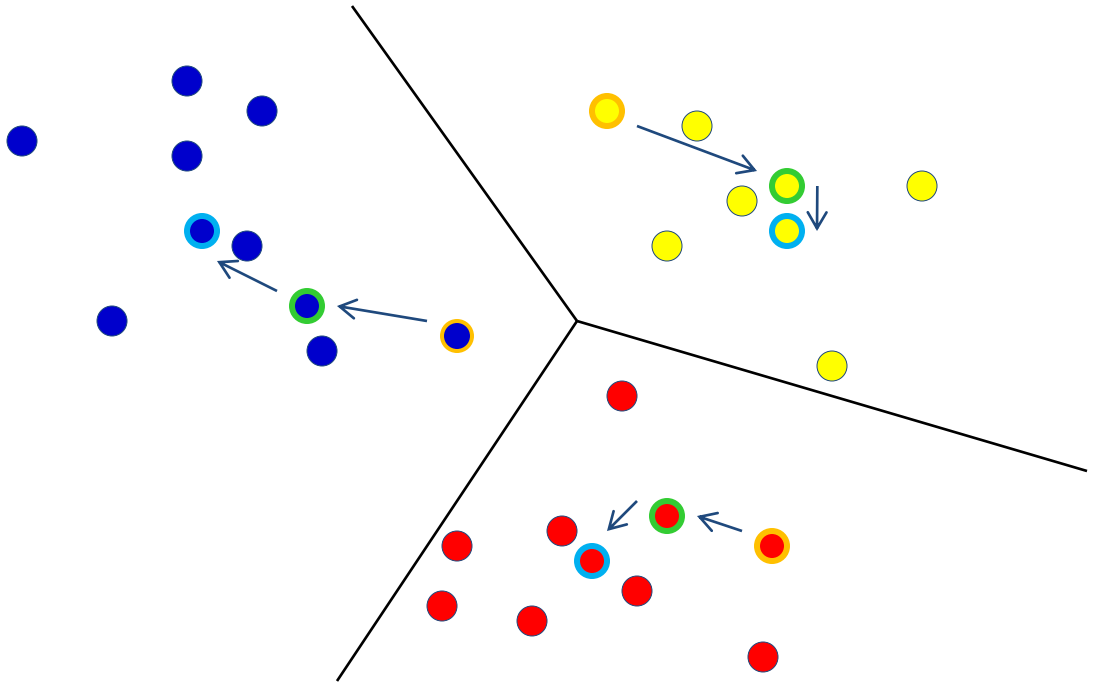
m[k]の更新



各データ点の所属の更新
破線の境界から実線の境界へ



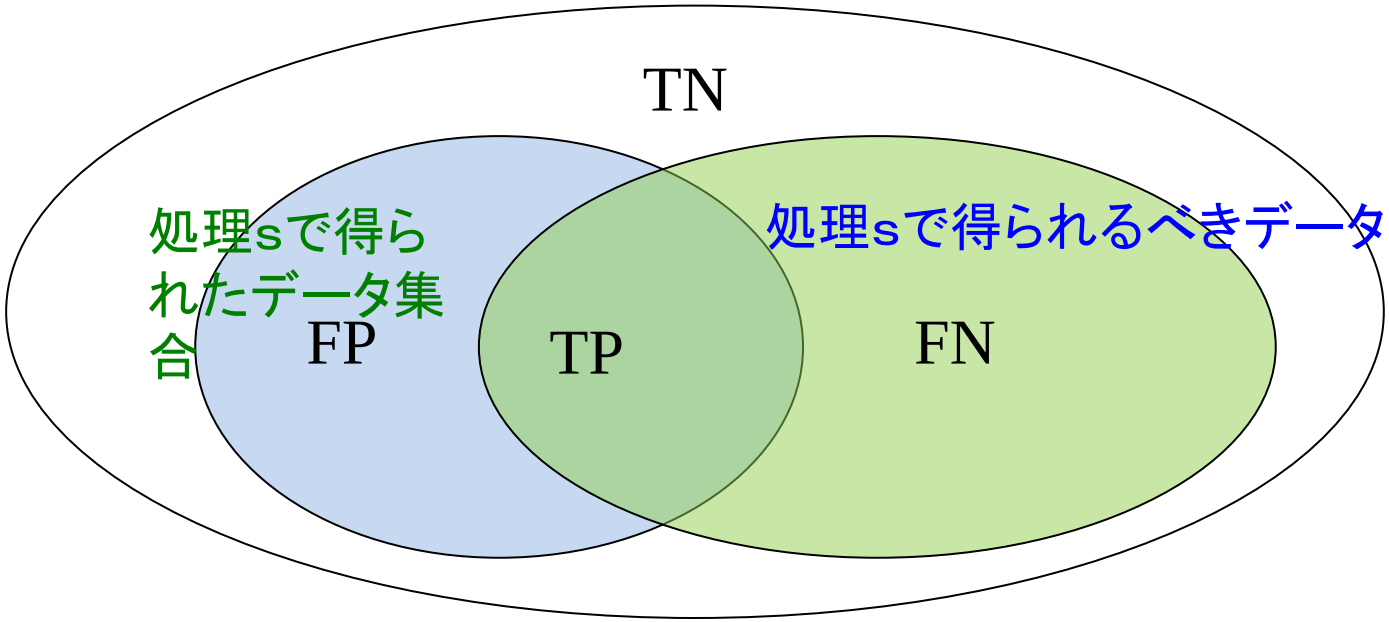
m[k]の更新



所属の更新：m[k]の変化は小さくなってくる

標準的評価法

- 精度:Precision=$\frac{TP}{TP+FP}$
- 再現率:Recall=$\frac{TP}{TP+FN}$
- F値=$\frac{2\times Precision\times Recall}{Precision+Reall}$



クラスタリングの評価：Purity

- 生成されたクラスタがどれだけ多数派で占められているかを表す尺度

N：データ数, C：真のクラス集合 $= (C_1, \ldots C_K)$,

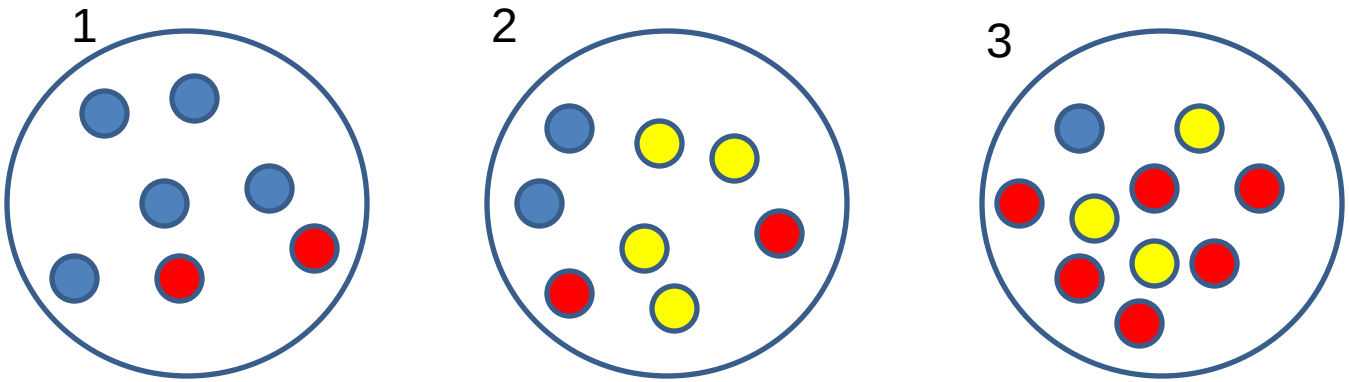
生成されたクラスタ数 $= L$

$n_{i,j}$：生成されたi番目のクラスタにおいて

j番目の真のクラスに属するデータ数

$$local \quad purity = \frac{1}{\sum_{j=1}^{L} n_{i,j}} \max\left(n_{i,j}\right)$$

$$global \quad purity = \frac{1}{\sum_{i=1}^{L}\sum_{j=1}^{K} n_{i,j}} \sum_{i=1}^{L} \max_{j}\left(n_{i,j}\right) = \frac{1}{N} \sum_{i=1}^{L} \max_{j}\left(n_{i,j}\right)$$



- local purity $purity(1)=\frac{5}{7},\ purity(1)=\frac{4}{8},\ purity(1)=\frac{6}{10}$
- global purity $purity=\frac{5+4+6}{7+8+10}=\frac{15}{25}=0.6$

- 問題点　何もしない場合
  - 全データが同一クラスタ $purity=\frac{1}{N}\max_{i,j}\left(n_{i,j}\right)$
  - 1クラスタが1データ $purity=\frac{1}{N}\sum_{i=1}^{L}\max_{j}\left(n_{i,j}\right)=\frac{1}{N}\sum_{i=1}^{L}1=\frac{N}{N}=1$

Inverse Purity

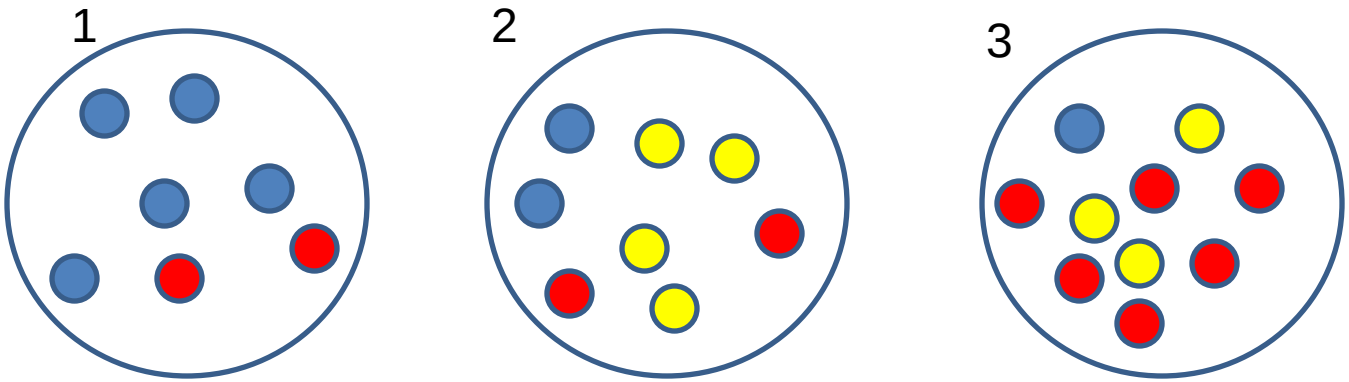
$$InversePurity = \frac{1}{N}\sum_{i=1}^{K}\left(\frac{\sum_{j=1}^{L} n_{i,j}}{\sum_{i=1}^{K} n_{i,j}}\max_{j}\left(n_{i,j}\right)\right)$$

クラスタiのデータ数

真のクラスjのデータ総数

1クラスタに1個のデータしかない場合も
Inverse Purityは1より小さい。
そこでPurityとの調和平均であるF値で評価

$$F値 = \frac{2}{\frac{1}{Purity} + \frac{1}{InversePurity}}$$



- $Purity = \frac{5+4+6}{7+8+10} = \frac{15}{25}$ ●8個、●7個、●10個

- $InversePurity = \frac{1}{25}\left(\frac{7}{8}\times 5 + \frac{8}{7}\times 4 + \frac{10}{10}\times 6\right) = 0.598$

- $F値 = \frac{2}{\frac{1}{0.6}+\frac{1}{0.598}} = 0.599$