

次元削減：PCAとLDA

主成分分析(PCA)

線形判別分析(LDA)

Pairwise距離と重みの導入



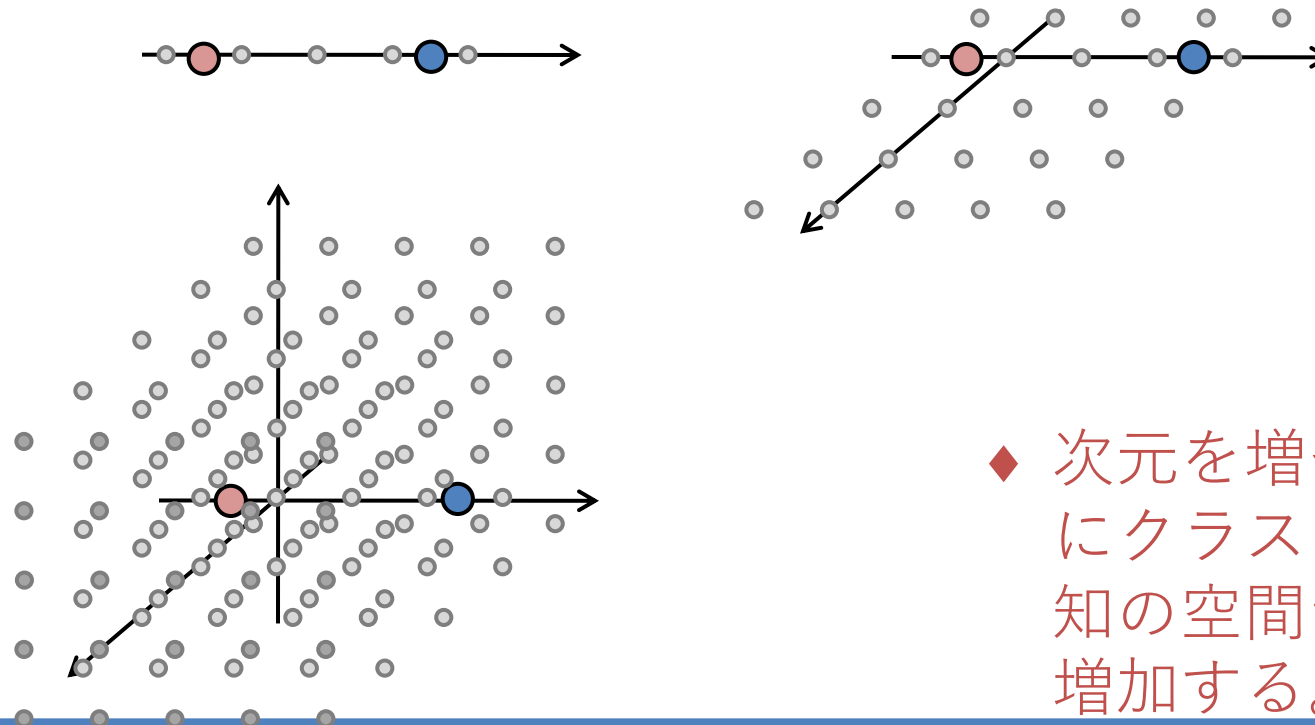
§

パターン認識における次元削減の必要性

□ 次元の呪い

- 特徴空間の次元が高くなると、信頼できる境界を得るために必要な学習データの量が幾何級数的に増大する。

⇒ 次元削減（情報圧縮）の必要性



- ◆ 次元を増やすたびに、周囲にクラスを定めていない未知の空間が、幾何級数的に増加する。



次元削減（情報圧縮）

次元削減（情報圧縮）： 高次元空間上のデータを，低次元の空間に，ある基準のもとで 写像する。

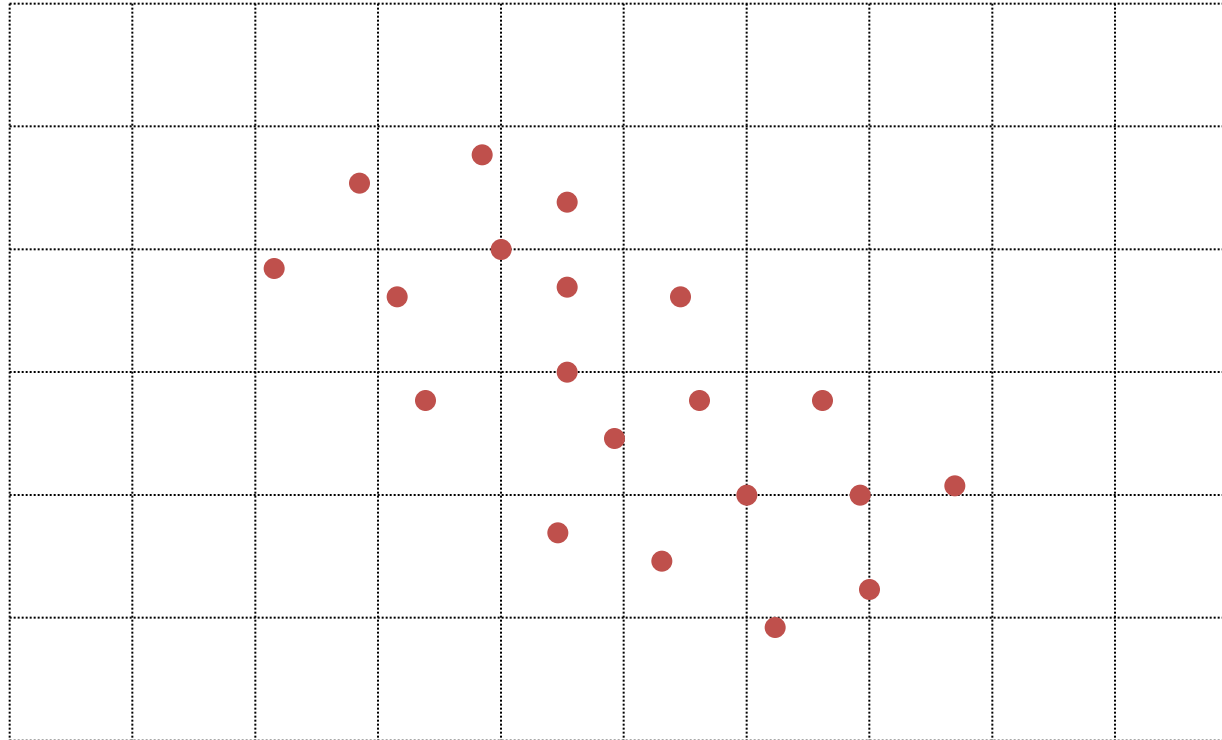
主成分分析： 低次元空間への写像に伴う情報のロスを最小に抑える線形写像。

線形判別分析： 低次元空間への写像で，複数のクラスの分離度を最大化する線形写像。



§

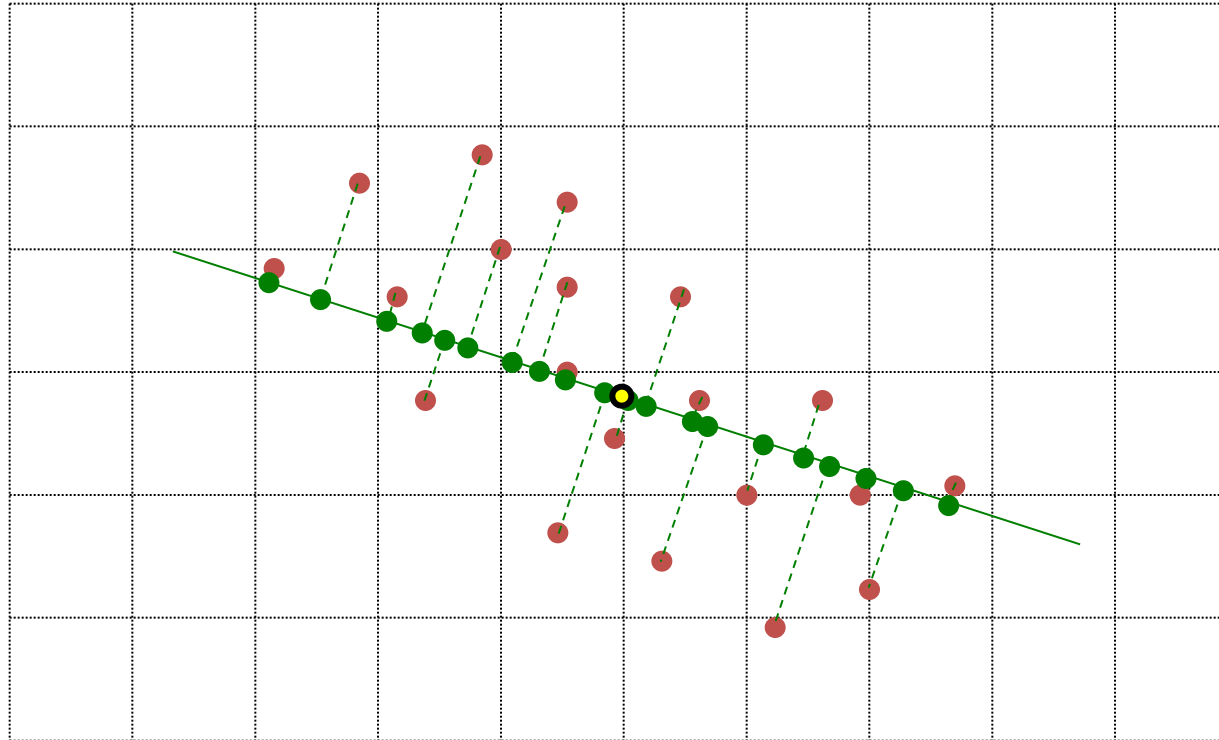
主成分分析



□ R^2 の空間に分布するデータが与えられたとする。



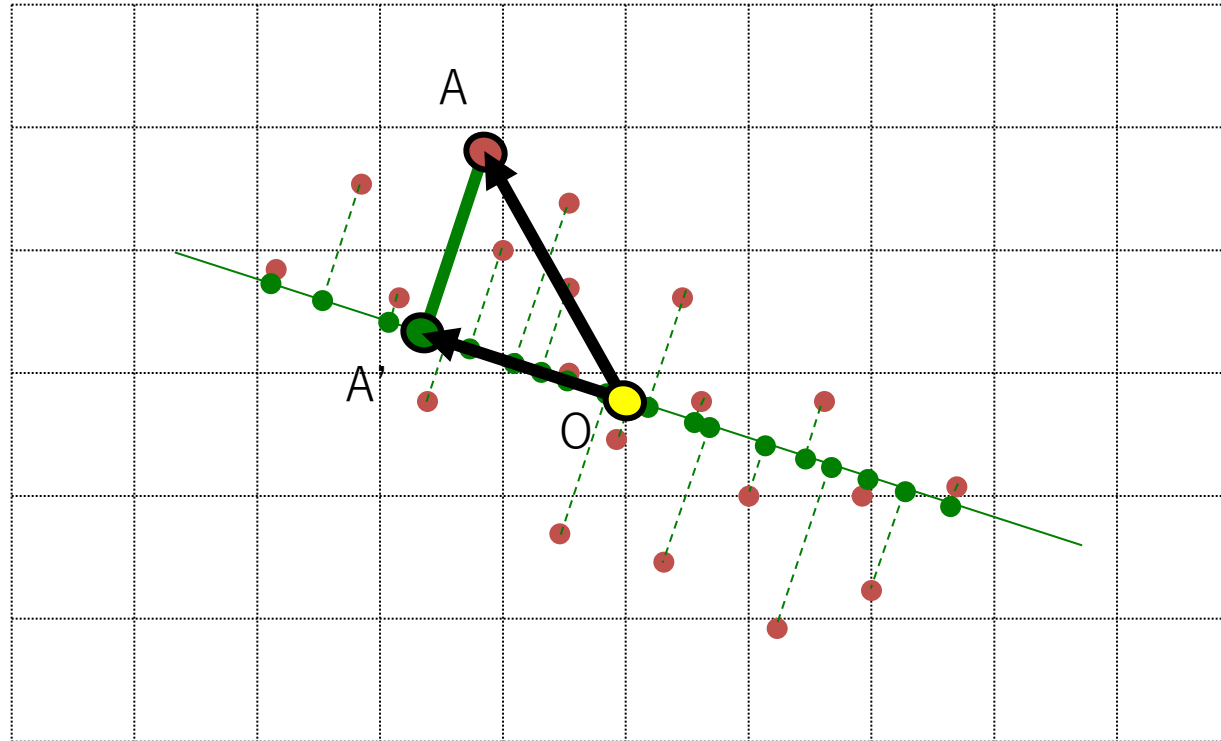
主成分分析



- この空間上に分布するデータを，直線状の部分空間に写像することを考える。



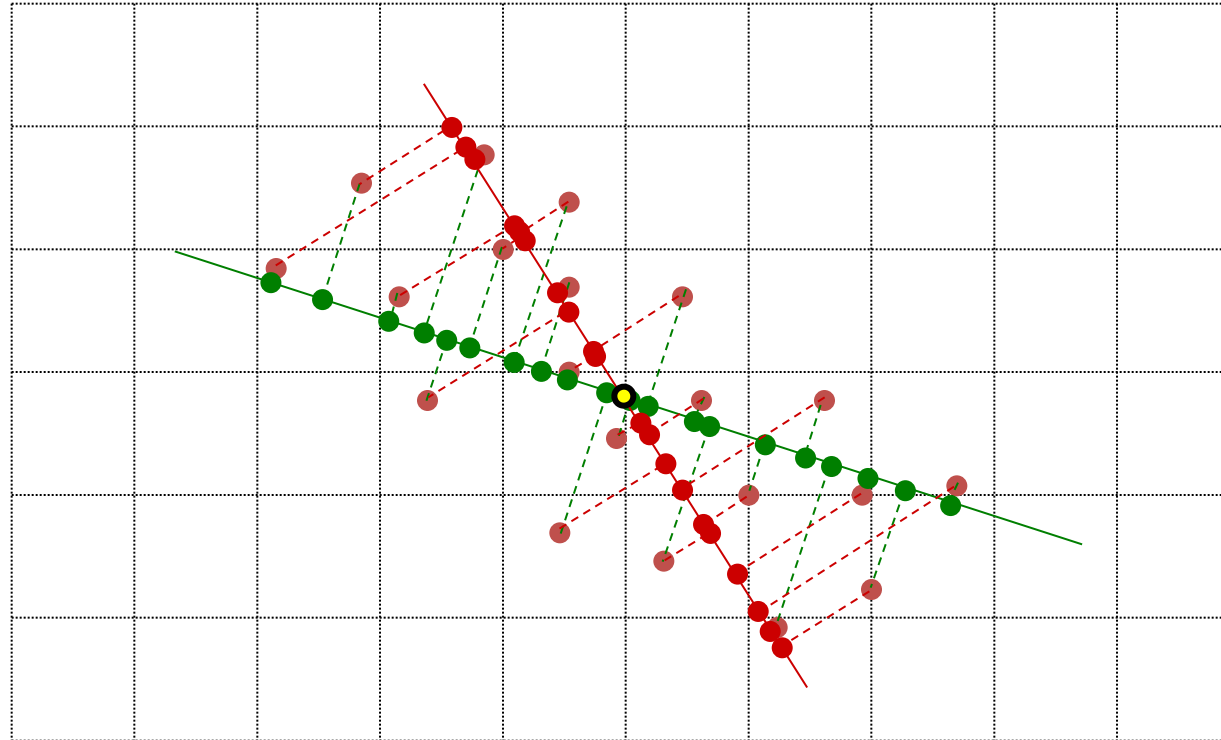
主成分分析



- $A-A'$ がロス
- ロスの最小化と、データの分散の最大化は等価。



主成分分析



- 部分空間を変えれば，分散も変わる。
- 最も，射影成分の分散の大きくなる部分空間を求める



第 1 主成分の求め方

いま P 次元のベクトルが, N 個あるとする。

$$\mathbf{x}_n = (x_{n1} \ x_{n2} \ \cdots \ x_{nP})^T, \quad n = 1, 2, 3, \dots, N$$

簡単のため, 各変数についてその平均値は 0 とする。(0 でない場合は, 平均値を引いて新たに変数を定義し直すことで, 一般性を失わない。)

このとき, 観測データ全体は次のデータ行列 X で与えられる。

$$X = (\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_N)^T = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1P} \\ x_{21} & x_{22} & \cdots & x_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{nP} \end{pmatrix}$$



第1主成分の求め方

求める主成分を,

$$\mathbf{w} = (w_1 \ w_2 \ \cdots \ w_P)^T, \quad \|\mathbf{w}\|^2 = \mathbf{w}^T \mathbf{w} = 1,$$

とする。今, n 番目のデータ

$$\mathbf{x}_n = (x_{n1} \ x_{n2} \ \cdots \ x_{nP})^T,$$

の主成分への射影 z_n を考えると,

$$z_n = \mathbf{x}_n^T \mathbf{w}$$

$\mathbf{z} = (z_1 \ z_2 \ \cdots \ z_N)^T$ とすれば,

$$\mathbf{z} = X\mathbf{w}$$



第1主成分の求め方

z_n の分散 σ^2 は,

$$\begin{aligned}\sigma^2 &= \frac{1}{N} \mathbf{z}^T \mathbf{z} = \frac{1}{N} (\mathbf{X}\mathbf{w})^T \mathbf{X}\mathbf{w} \\ &= \frac{1}{N} \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} = \mathbf{w}^T \mathbf{C} \mathbf{w}\end{aligned}$$

ただし, \mathbf{C} は, 共分散行列.

$$\mathbf{C} = \frac{1}{N} \mathbf{X}^T \mathbf{X} = (c_{ij}), \quad c_{ij} = \frac{1}{N} \sum_{k=1}^N x_{ki} x_{kj}$$

平均値がゼロでない一般的な場合は,

$$\mathbf{C} = \frac{1}{N} \mathbf{X}^T \mathbf{X} - \boldsymbol{\mu} \boldsymbol{\mu}^T = (c_{ij}), \quad c_{ij} = \frac{1}{N} \sum_{k=1}^N (x_{ki} - \mu_i)(x_{kj} - \mu_j)$$



補足：共分散行列

$\mathbf{X} = (X_1 \ X_2 \ \cdots \ X_M)^T$ ($i = 1, 2, \dots, N$) : X_i は確率変数

$\boldsymbol{\mu} = (\mu_1 \ \mu_2 \ \cdots \ \mu_M)^T$, $\mu_i = E[X_i]$: 母平均

に対し,

$$c_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)]$$

$$\mathbf{C} = (c_{ij}) = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]$$

を共分散行列と呼ぶ。

このとき,

$$\mathbf{C} = E[\mathbf{X}\mathbf{X}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T$$

が成立する。



補足：共分散行列

$$\mathbf{x}_i = (x_{i1} \ x_{i2} \ \cdots \ x_{iM})^T \ (i = 1, 2, \dots, N)$$

$$\mathbf{X} = (\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_N)^T : \text{データ行列} \quad (\text{前ページの}\mathbf{X}\text{と異なることに注意})$$

$$\mathbf{X}' = (\mathbf{x}_1 - \boldsymbol{\mu} \ \mathbf{x}_2 - \boldsymbol{\mu} \ \cdots \ \mathbf{x}_N - \boldsymbol{\mu})^T : \text{平均値補正後のデータ行列}$$

$$\boldsymbol{\mu} = (\mu_1 \ \mu_2 \ \cdots \ \mu_M)^T : \text{母平均}$$

のとき,

共分散行列は,

$$\mathbf{C} = \frac{1}{N} \mathbf{X}'^T \mathbf{X}' = \frac{1}{N} \mathbf{X}^T \mathbf{X} - \boldsymbol{\mu} \boldsymbol{\mu}^T \qquad \mathbf{C} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^T - \boldsymbol{\mu} \boldsymbol{\mu}^T$$

となる。



$$X' = \begin{pmatrix} x_{11} - \mu_1 & x_{12} - \mu_2 & \cdots & x_{1P} - \mu_P \\ x_{21} - \mu_1 & x_{22} - \mu_2 & \cdots & x_{jP} - \mu_P \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} - \mu_1 & x_{Ni} - \mu_2 & \cdots & x_{NP} - \mu_P \end{pmatrix}$$

$$C = \frac{1}{N} X'^T X'$$

$$= \frac{1}{N} \begin{pmatrix} x_{11} - \mu_1 & x_{21} - \mu_1 & \cdots & x_{N1} - \mu_1 \\ \vdots & \vdots & \ddots & \vdots \\ x_{1i} - \mu_i & x_{2i} - \mu_i & \cdots & x_{Ni} - \mu_i \\ \vdots & \vdots & \ddots & \vdots \\ x_{1P} - \mu_P & x_{2P} - \mu_P & \cdots & x_{NP} - \mu_P \end{pmatrix} \begin{pmatrix} x_{11} - \mu_1 & \cdots & x_{1j} - \mu_j & \cdots & x_{1P} - \mu_P \\ x_{21} - \mu_1 & \cdots & x_{2j} - \mu_j & \cdots & x_{2P} - \mu_P \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{N1} - \mu_1 & \cdots & x_{Nj} - \mu_j & \cdots & x_{NP} - \mu_P \end{pmatrix}$$

$$= (c_{ij})$$



$\mu = 0$ のとき,

$$X = \begin{pmatrix} \boxed{x_{11} \quad x_{12} \quad \cdots \quad x_{1P}} \\ \boxed{x_{21} \quad x_{22} \quad \cdots \quad x_{2P}} \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ \boxed{x_{N1} \quad x_{N2} \quad \cdots \quad x_{NP}} \end{pmatrix}$$

$$\begin{aligned} C &= \frac{1}{N} X^T X = \frac{1}{N} \begin{pmatrix} \boxed{x_{11}} & \boxed{x_{21}} & \cdots & \boxed{x_{N1}} \\ \vdots & \vdots & \ddots & \vdots \\ \boxed{x_{1i}} & \boxed{x_{2i}} & \cdots & \boxed{x_{Ni}} \\ \vdots & \vdots & \ddots & \vdots \\ \boxed{x_{1P}} & \boxed{x_{2P}} & \cdots & \boxed{x_{NP}} \end{pmatrix} \begin{pmatrix} \boxed{x_{11}} \quad \cdots \quad \boxed{x_{1j}} \quad \cdots \quad x_{1P} \\ \boxed{x_{21}} \quad \cdots \quad \boxed{x_{2j}} \quad \cdots \quad x_{2P} \\ \vdots \quad \ddots \quad \vdots \quad \ddots \quad \vdots \\ \boxed{x_{N1}} \quad \cdots \quad \boxed{x_{Nj}} \quad \cdots \quad x_{NP} \end{pmatrix} \\ &= \boxed{(c_{ij})} \end{aligned}$$



補足：共分散行列

$$\mathbf{x}_i = (x_{i1} \ x_{i2} \ \cdots \ x_{iM})^T \ (i = 1, 2, \dots, N)$$

$$D = (\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_N)^T \quad : \text{データ行列}$$

$$\bar{\boldsymbol{\mu}} = (\bar{\mu}_1 \ \bar{\mu}_2 \ \cdots \ \bar{\mu}_M)^T \quad : \text{標本平均}$$

のとき,

共分散行列は,

$$C = \frac{1}{N-1} X^T X - \bar{\boldsymbol{\mu}} \bar{\boldsymbol{\mu}}^T$$

となる。

(標本平均を用いる場合であっても, N で割ることもある。この授業では, 主に N で割る形をとることが多い。)



補足: 標本平均で分散を求めるとき なぜ N でなく $N - 1$ で割るのか

μ : 母平均, $\bar{\mu}$: 標本平均

$$\sum_{i=1}^N (x_i - \bar{\mu})^2 = \sum_{i=1}^N (x_i - \mu - \bar{\mu} + \mu)^2 = \sum_{i=1}^N (x_i - \mu)^2 - N(\bar{\mu} - \mu)^2$$

$$= \sum_{i=1}^N (x_i - \mu)^2 - \frac{1}{N} \left(\sum_{i=1}^N (x_i - \mu) \right)^2$$

標本平均の母平均
からのずれを補正
する必要がある

$$= \sum_{i=1}^N (x_i - \mu)^2 - \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 - \frac{1}{N} \sum_{i=1}^N \sum_{j \neq i}^N (x_i - \mu)(x_j - \mu)$$

$$\therefore \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{\mu})^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

偏差が無相関であれば期待値0



第1主成分の求め方

問題は, w で $\sigma^2 = w^T C w$ を最大化すること。

$$\max_w w^T C w$$

ただし, $w^T w = 1$ でなければならない !!

$$\max_w w^T C w$$

$$\text{s. t. } w^T w = 1$$

(1)



第 1 主成分の求め方

Lagrange の未定乗数法により,

$$L(\mathbf{w}, \lambda) = \mathbf{w}^T C \mathbf{w} - \lambda(\mathbf{w}^T \mathbf{w} - 1) \quad (2)$$

と置いて, $\frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, \lambda) = 0$ を解くと,

$$\frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, \lambda) = 2(C - \lambda I)\mathbf{w} = 0 \quad (3)$$

となつて, 解くべき問題は固有値問題に帰着する。

$$(C - \lambda I)\mathbf{w} = 0, \quad C\mathbf{w} = \lambda\mathbf{w}, \quad \det(C - \lambda I) = 0, \quad (4)$$



第 1 主成分の求め方

$$(C - \lambda I)\mathbf{w} = 0$$

$$\det(C - \lambda I) = 0$$

λ が満たすべき条件は、共分散行列 C の固有方程式となる。

よって、

射影の分散の最大値を与える主成分 \mathbf{w} は共分散行列の固有ベクトルのひとつであり、

λ はその固有ベクトルに対応した固有値であることがわかる。



第 1 主成分の求め方

ここで, $(C - \lambda I)\mathbf{w} = 0$ を満たす固有ベクトル \mathbf{w} , 固有値 λ はそれぞれ, P 個ある。

このうち, どれが射影の分散の最大値を与えるかを考える。

$$C\mathbf{w} = \lambda\mathbf{w}$$

であって,

$$\sigma^2 = \mathbf{w}^T C \mathbf{w}$$

であるから,

$$\sigma^2 = \mathbf{w}^T C \mathbf{w} = \lambda \mathbf{w}^T \mathbf{w}$$



第1主成分の求め方

ここで,

$$\mathbf{w}^T \mathbf{w} = 1$$

に注意すれば,

$$\sigma^2 = \lambda \quad (5)$$

であるから, 射影の分散の値は固有値に等しいことが分かる。

よって, 求める主成分は最大固有値に対応した固有ベクトルとして与えられることが分かる。



第2以降の主成分の求め方

帰納法によって、第 m 主成分を求める。

今、共分散行列 C の固有値のうち、大きなものから順に $m - 1$ 個に対する固有ベクトルとして、 $m - 1$ の主成分 $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k, \dots, \mathbf{w}_{m-1}$ が求まっているものとする。

これらのベクトルは、

$$(C - \lambda I)\mathbf{w}_k = 0, \quad \mathbf{w}_k^T \mathbf{w}_k = 1 \quad (6)$$

を満たし、互いに直交しているものとする。

$$\mathbf{w}_k^T \mathbf{w}_{k'(\neq k)} = 0 \quad (7)$$



第2以降の主成分の求め方

Lagrange の未定乗数法により,

$$J_m = \mathbf{w}_m^T C \mathbf{w}_m - \lambda_m (\mathbf{w}_m^T \mathbf{w}_m - 1) - \sum_{k=1}^{m-1} \mu_k \mathbf{w}_m^T \mathbf{w}_k \quad (8)$$

と置いて, $\frac{\partial J_m}{\partial \mathbf{w}_m} = 0$ を解くと,

$$C \mathbf{w}_m - \lambda_m \mathbf{w}_m - \sum_{k=1}^{m-1} \mu_k \mathbf{w}_k = 0 \quad (9)$$

左から \mathbf{w}_j^T ($j = 1, 2, \dots, m-1$) をかけると,

$$\mathbf{w}_j^T C \mathbf{w}_m - \lambda_m \mathbf{w}_j^T \mathbf{w}_m - \sum_{k=1}^{m-1} \mu_k \mathbf{w}_j^T \mathbf{w}_k = 0$$

$$\mathbf{w}_j^T C \mathbf{w}_m - \mu_j = 0$$

$$\mu_j = \mathbf{w}_j^T C \mathbf{w}_m = \lambda_j \mathbf{w}_j^T \mathbf{w}_m^T = 0 \quad (10)$$



第2以降の主成分の求め方

$$\mu_j = 0, (j = 1, 2, \dots, m - 1)$$

以上によって、最適化のための評価関数として、以下の関係式が導かれる。

$$J_m = \mathbf{w}_m^T C \mathbf{w}_m - \lambda_m (\mathbf{w}_m^T \mathbf{w}_m - 1) \quad (11)$$

これは、第1主成分を求めるための評価関数と同じである。

よって、求める第 m 主成分 \mathbf{w}_m は、共分散行列 C の m 番目に大きい固有値に対応した固有ベクトルとして与えられることが分かる。



寄与率

元の空間における分散 : σ_0^2

第 m 次の部分空間における分散 : σ_m^2

□ 寄与率 : 第 m 次部分空間で表現できる分布の広がり の程度

$$\text{寄与率} = \frac{\sigma_m^2}{\sigma_0^2}$$

□ 累積寄与率 : 第1～ M 次部分空間で表現できる分布の広がり の程度

$$\text{累積寄与率} = \frac{\sum_{m=1}^M \sigma_m^2}{\sigma_0^2}$$



例題

4つのデータ,
$$\mathbf{x}_1 = \begin{pmatrix} 5 \\ 0 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} -3 \\ -4 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix} -5 \\ 0 \end{pmatrix}$$

が与えられている。

- 1) このデータの共分散行列 C を求めよ。
- 2) ベクトル $\mathbf{w} = (w_1 \ w_2)^T$ を用いて, $z_i = \mathbf{x}_i^T \mathbf{w}$ と座標変換するとき, $\|\mathbf{w}\| = 1$ の条件の下で, z_i の分散を最も大きくするように, \mathbf{w} を定めることにするとき, C と \mathbf{w} の関係を導け。
- 3) 2) の条件を満たす \mathbf{w} を実際に求めよ。
- 4) データの散布図を描け。この散布図に重ねて \mathbf{w} を描け。



1)

$$\mathbf{x}_1 = \begin{pmatrix} 5 \\ 0 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} -3 \\ -4 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix} -5 \\ 0 \end{pmatrix}$$

$$N = 4, \quad \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$X = \begin{pmatrix} 5 & -3 & 3 & -5 \\ 0 & -4 & 4 & 0 \end{pmatrix}^T$$

$$\begin{aligned} C &= \frac{1}{4} \begin{pmatrix} 5 & -3 & 3 & -5 \\ 0 & -4 & 4 & 0 \end{pmatrix} \begin{pmatrix} 5 & 0 \\ -3 & -4 \\ 3 & 4 \\ -5 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 17 & 6 \\ 6 & 8 \end{pmatrix} \end{aligned}$$



$$2) \quad L(\mathbf{w}, \lambda) = \mathbf{w}^T C \mathbf{w} - \lambda(\mathbf{w}^T \mathbf{w} - 1)$$

$$\frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, \lambda) = (C - \lambda I) \mathbf{w} = 0$$

$$3) \quad \det(C - \lambda I) = 0, \quad \begin{vmatrix} 17 - \lambda & 6 \\ 6 & 8 - \lambda \end{vmatrix} = 0$$

$$(17 - \lambda)(8 - \lambda) - 6 \cdot 6 = (\lambda - 20)(\lambda - 5) = 0$$

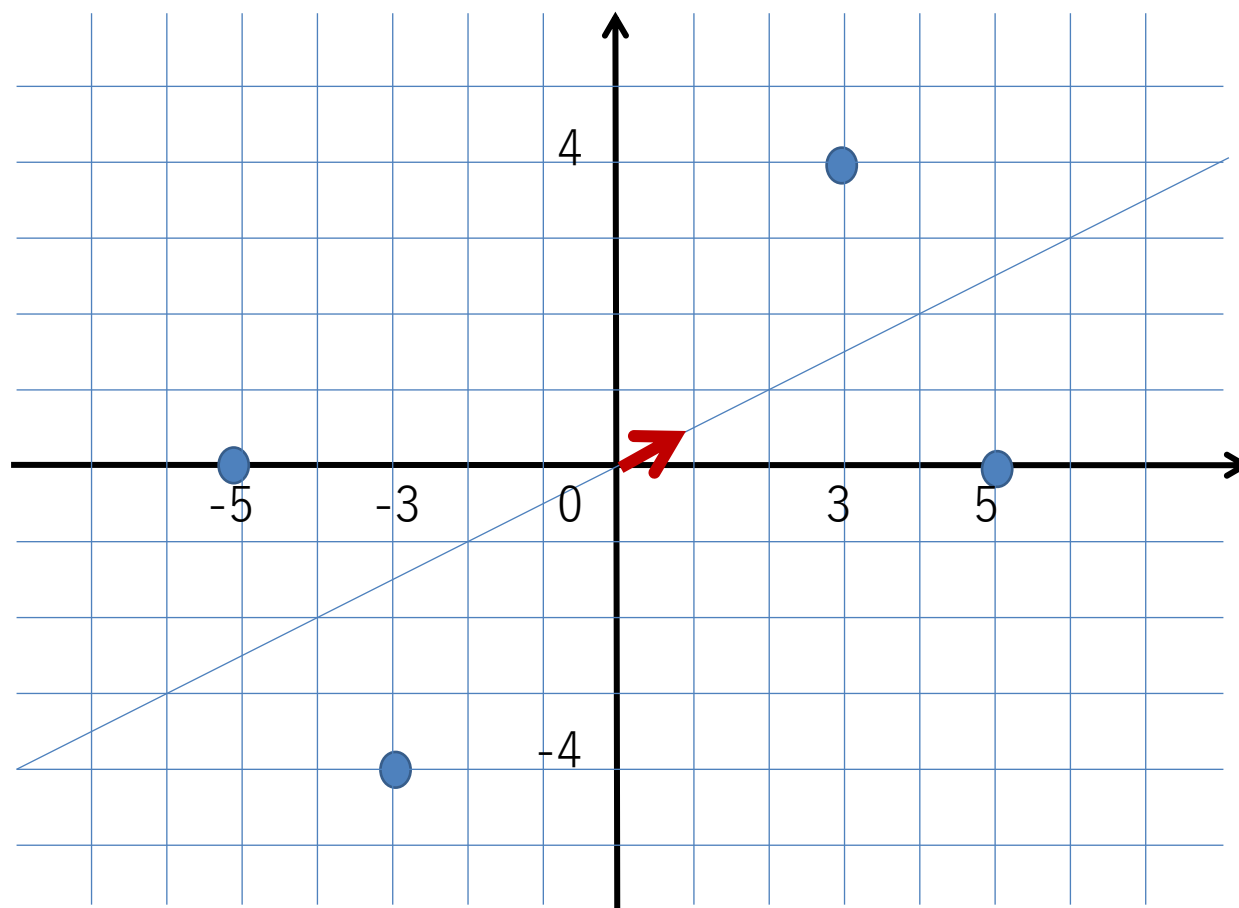
最大固有値は $\lambda = 20$

最大固有値に対応した固有ベクトルは, $(17 - 20 \quad 6) = k(-1, 2)$ に直交する単位ベクトル

$$\mathbf{w} = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$



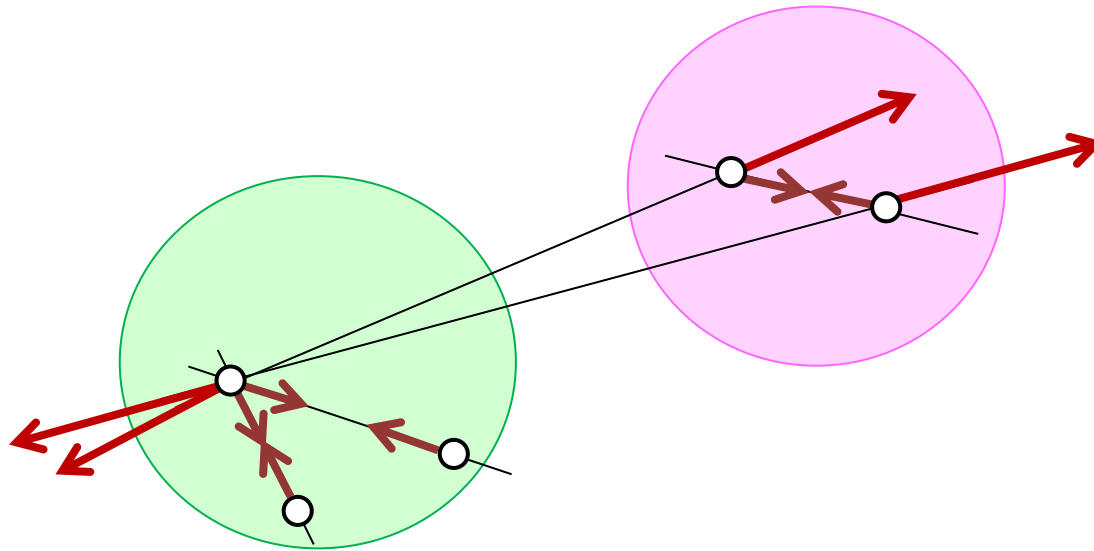
4)



§

線形判別分析(LDA : Linear Discriminant Analysis)

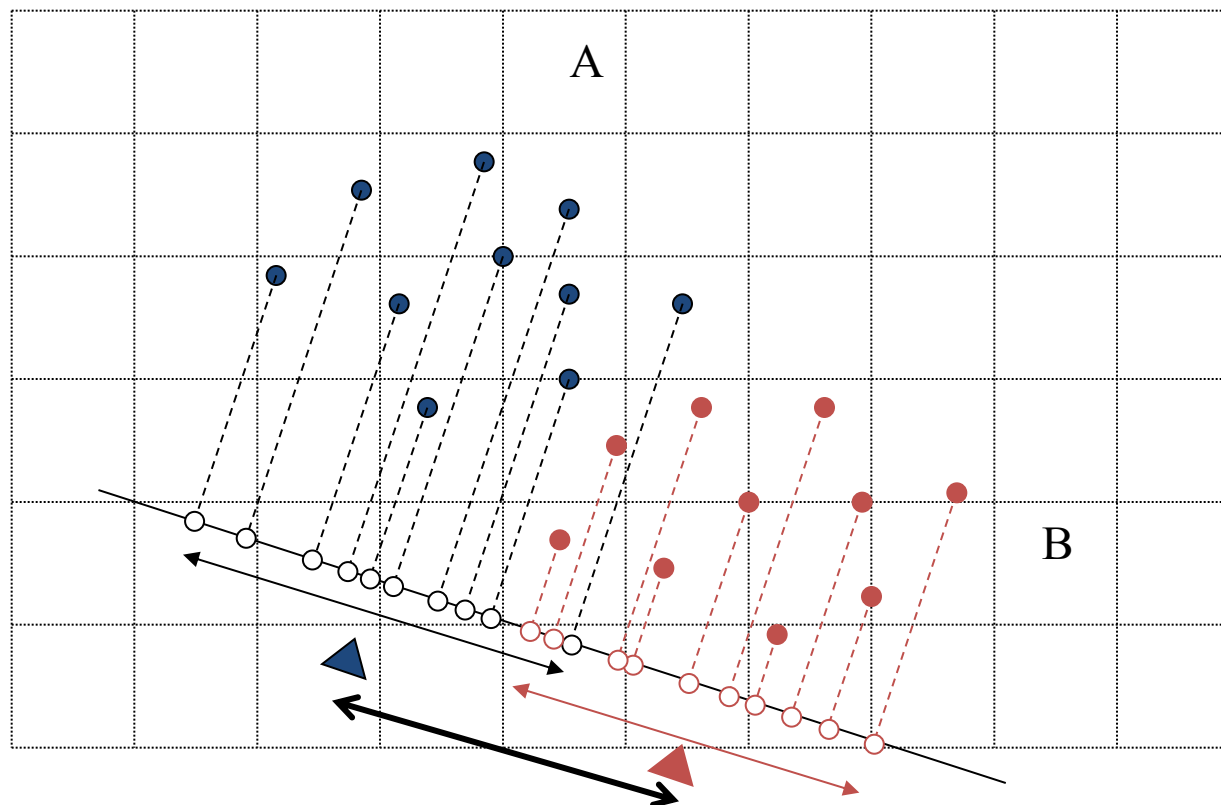
級内分散 C_W と級間分散 C_B の比を最大にする座標空間を求める。



クラスと同じデータは近くに,
クラスの違うデータは遠くに配置する。



2クラスの場合のLDA



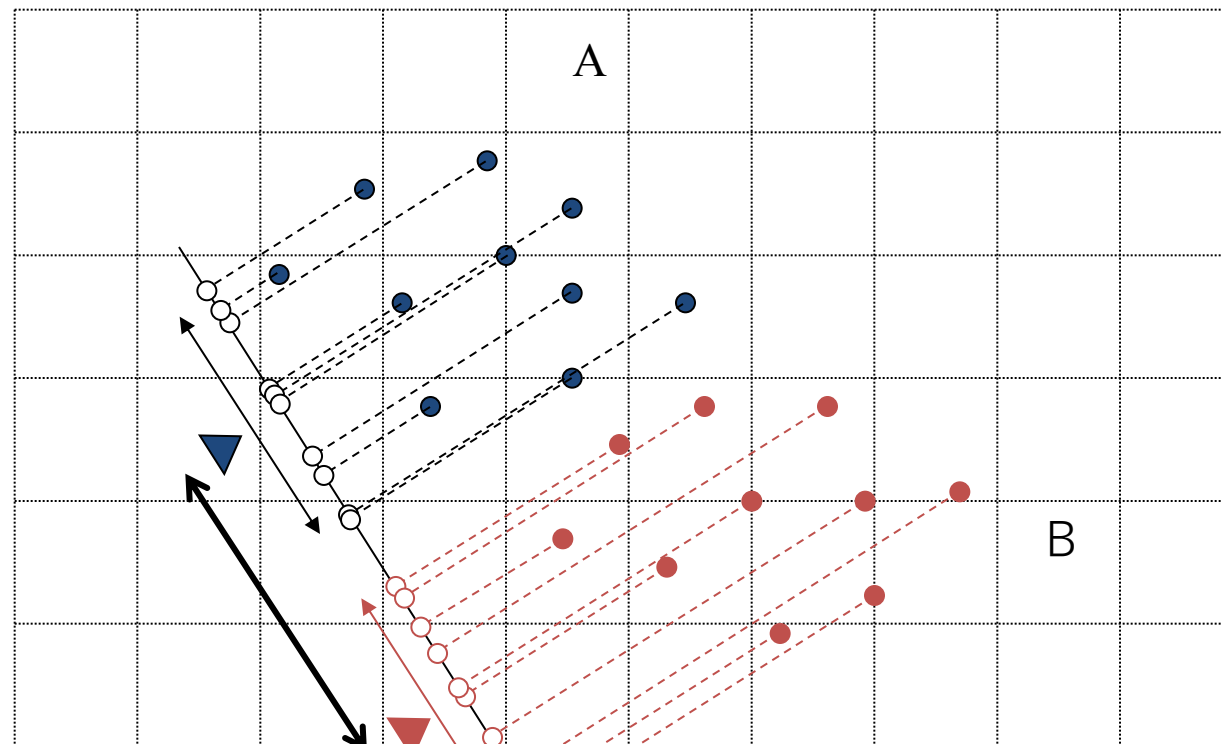
$$y = \mathbf{w}^T \mathbf{x}$$

$$\mu' = E[y], \quad \sigma' = E[(y - \mu')^2]$$

$$J(\mathbf{w}) = \frac{(\mu'_A - \mu'_B)^2}{\sigma'^2_A + \sigma'^2_B}$$



2クラスの場合のLDA



$$J(\mathbf{w}) = \frac{(\mu'_A - \mu'_B)^2}{\sigma'^2_A + \sigma'^2_B}$$

射影ベクトル \mathbf{w} を変えると、 J の値が変わる。

J が大きいとき、識別は容易になる。

→ J を最大化するベクトル \mathbf{w} を求める。



\boldsymbol{x} : 特徴ベクトル

N_i : クラス i のデータ数

$\boldsymbol{\mu}_i$: クラス i の平均ベクトル

X_i : クラス i に対する元空間における特徴ベクトルの集合

とすると、クラス i の共分散行列 C_i , 変動行列 S_i は、次式で与えられる。

$$C_i = \frac{1}{N_i} \sum_{\boldsymbol{x} \in X_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \quad (1)$$

$$S_i = N_i C_i = \sum_{\boldsymbol{x} \in X_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \quad (2)$$



また, \mathbf{x} をベクトル \mathbf{w} に射影して作った1次元の空間において,
 y : 射影して作られた \mathbf{x} の像

$$y = \mathbf{x}^T \mathbf{w}$$

$\tilde{\mu}_i$: 射影先の空間におけるクラス i のデータの平均値

Y_i : 射影先の空間におけるクラス i のデータの集合

とすれば, 射影先の空間におけるクラス i のデータの分散 $\tilde{\sigma}_i$, 変動 \tilde{S}_i は, 次式で与えられる。

$$\tilde{\sigma}_i = \frac{1}{N_i} \sum_{y \in Y_i} (y - \tilde{\mu}_i)^2 = \frac{1}{N_i} \sum_{\mathbf{x} \in X_i} \mathbf{w}^T (\mathbf{x} - \boldsymbol{\mu}_i) (\mathbf{x} - \boldsymbol{\mu}_i)^T \mathbf{w} = \mathbf{w}^T \mathbf{C}_i \mathbf{w} \quad (3)$$

$$\tilde{S}_i = N_i \tilde{\sigma}_i = N_i \mathbf{w}^T \mathbf{C}_i \mathbf{w} = \mathbf{w}^T \mathbf{S}_i \mathbf{w} \quad (4)$$




クラス内変動 S_W および クラス間変動 S_B を以下のように定義する。

$$S_W = S_1 + S_2 = \sum_{i=1,2} \sum_{\mathbf{x} \in X_i} (\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T \quad (5)$$

$$S_B = \sum_{i=1,2} N_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \quad (6)$$

ただし, $\boldsymbol{\mu}$ は全データの平均値である。このとき, 射影先でクラス内変動 \tilde{S}_W , クラス間変動 \tilde{S}_B は以下のようになる。

$$\tilde{S}_W = \tilde{S}_1 + \tilde{S}_2 = \mathbf{w}^T S_1 \mathbf{w} + \mathbf{w}^T S_2 \mathbf{w} = \mathbf{w}^T S_W \mathbf{w} \quad (7)$$

$$\begin{aligned} \tilde{S}_B &= \sum_{i=1,2} N_i (\tilde{\mu}_i - \tilde{\mu})^2 = \sum_{i=1,2} N_i \mathbf{w}^T (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \mathbf{w} \\ &= \mathbf{w}^T S_B \mathbf{w} \end{aligned} \quad (8)$$


J を以下のように定義する。

$$J = \frac{\tilde{S}_B}{\tilde{S}_W} = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \quad (9)$$

この J を最大にする \mathbf{w} を求める問題を解く。
これは,

$$\mathbf{w}^T S_W \mathbf{w} = \text{const.} \quad (10)$$

なる条件の下で,

$$\mathbf{w}^T S_B \mathbf{w} \quad (11)$$

を最大にする問題と等価である。



E を以下のように定義する。

$$E = \mathbf{w}^T S_B \mathbf{w} - \lambda(\mathbf{w}^T S_w \mathbf{w} - \text{const.}) \quad (12)$$

\mathbf{w} で偏微分して 零 と置くと,

$$\begin{aligned} S_B \mathbf{w} - \lambda S_w \mathbf{w} &= 0 \\ (S_w^{-1} S_B - \lambda I) \mathbf{w} &= 0 \end{aligned} \quad (13)$$

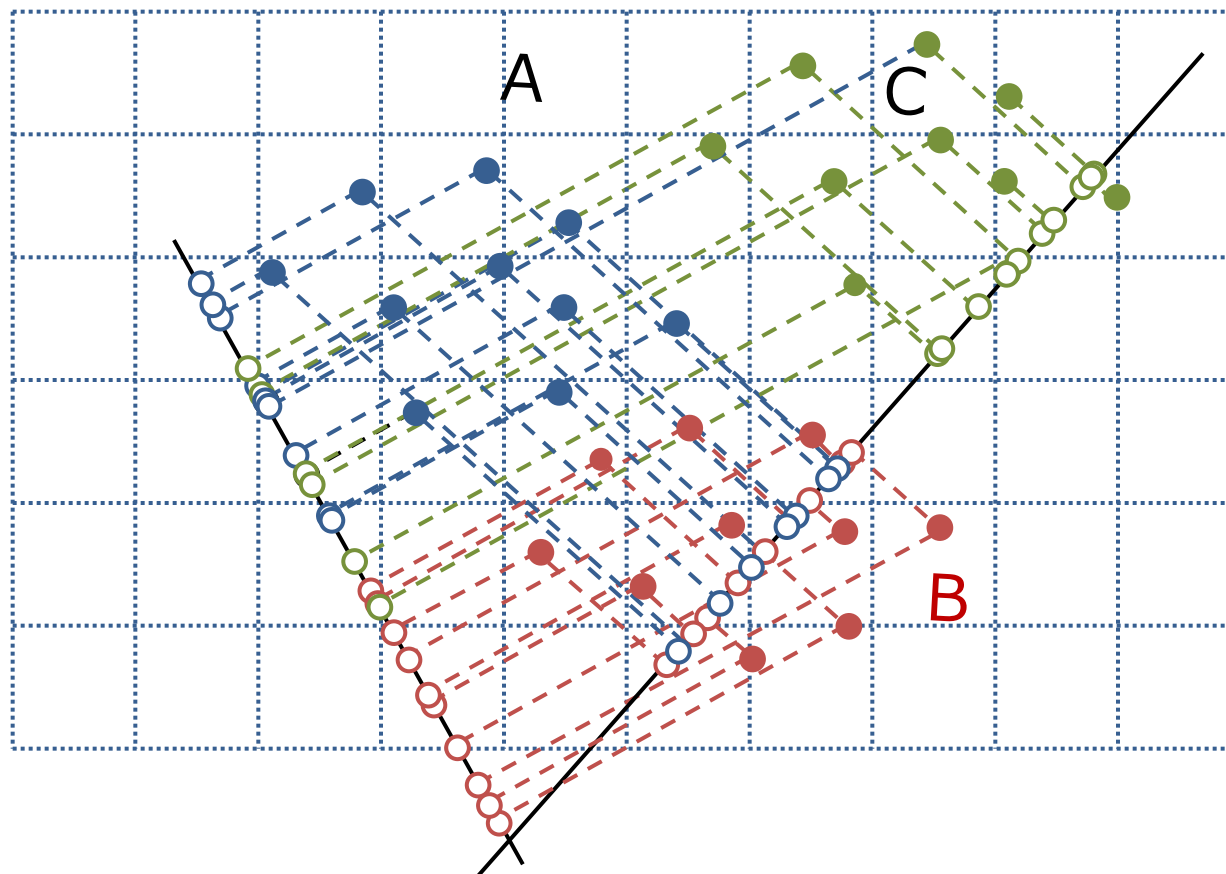
を得る。

よって, 問題は $S_w^{-1} S_B$ の固有値を求める問題に帰着する。

即ち, $S_w^{-1} S_B$ の最大固有値を λ_1 とすれば, J の最大値を与える \mathbf{w} は λ_1 に対する固有ベクトルとして求めることができる。



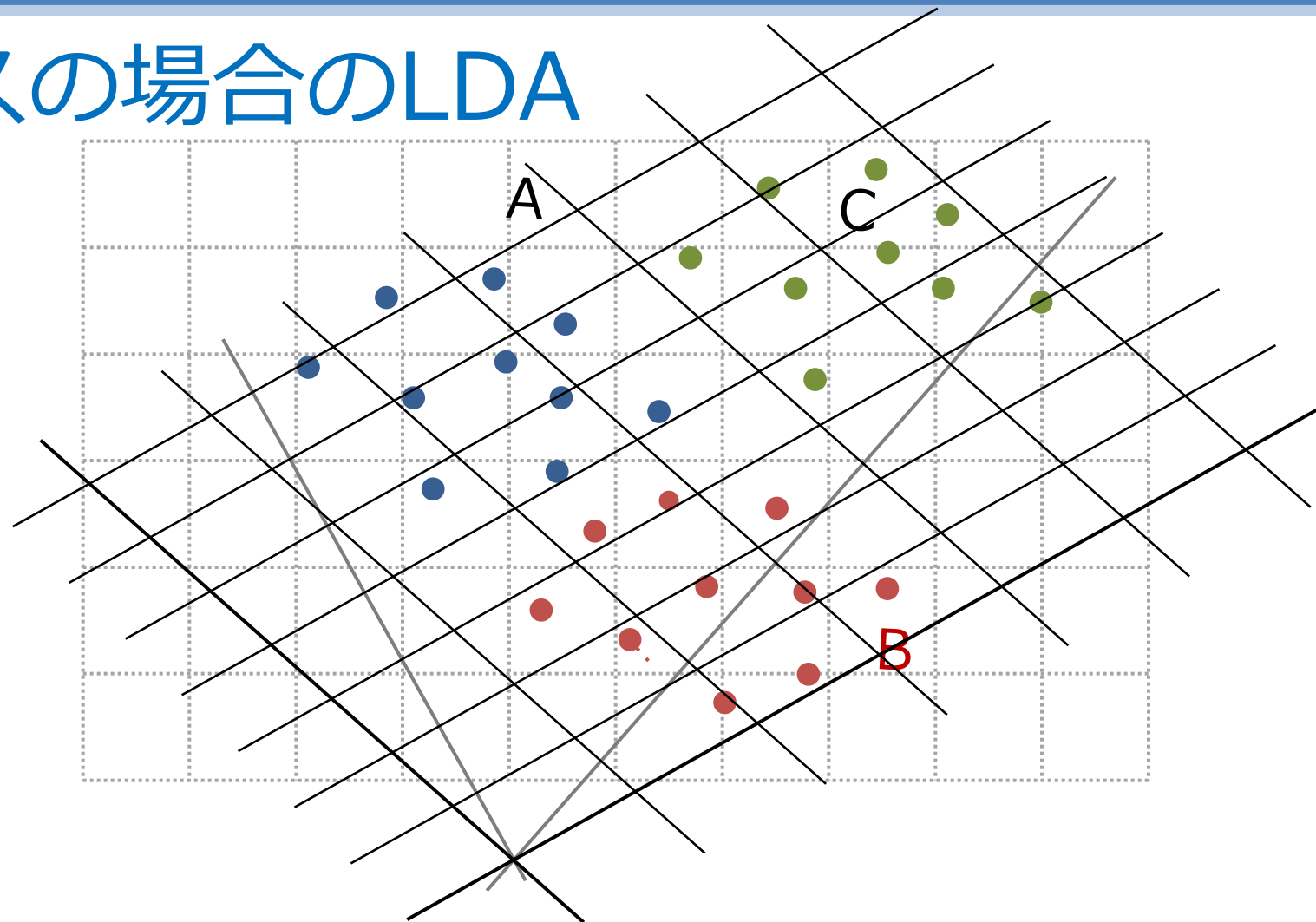
多クラスの場合のLDA



Kクラスの問題については、 $K-1$ 軸に射影することを考える。



多クラスの場合のLDA



斜交の座標軸が構成される。



ベクトル \mathbf{w}_i ($1, 2, \dots, K - 1$) に \mathbf{x} を射影して作った $K - 1$ 次元の空間において,

\mathbf{y} : 射影して作られた \mathbf{x} の像

$$\mathbf{y} = W^T \mathbf{x}, \quad W = (\mathbf{w}_1 \mathbf{w}_2 \cdots \mathbf{w}_{N-1})$$

$\tilde{\mu}_i$: 射影先の空間におけるクラス i のデータの平均値

Y_i : 射影先の空間におけるクラス i のデータの集合

とすれば, 射影先の空間におけるクラス i のデータの共分散行列 \tilde{C}_i , 変動行列 \tilde{S}_i は, 次式で与えられる。

$$\tilde{C}_i = \frac{1}{N_i} \sum_{\mathbf{x} \in Y_i} (\mathbf{y} - \tilde{\mu}_i)(\mathbf{y} - \tilde{\mu}_i)^T = W^T C_i W \quad (1)$$

$$\tilde{S}_i = N_i \tilde{C}_i = N_i W^T C_i W = W^T S_i W \quad (2)$$



2クラスのと看と同様に,

$$S_W = \sum_{i=1}^K S_i = \sum_{i=1}^K \sum_{x \in X_i} (x - \mu_i)(x - \mu_i)^T \quad (3)$$

$$S_B = \sum_{i=1}^K N_i (\mu_i - \mu)(\mu_i - \mu)^T \quad (4)$$

また,

$$\tilde{S}_W = W^T S_W W \quad (5)$$

$$\tilde{S}_B = W^T S_B W \quad (6)$$

\tilde{S}_B と \tilde{S}_W の比が大きくなるように W を設定すればよい。



2クラスのとときと異なり, \tilde{S}_B , \tilde{S}_W は行列であるから, 単純な比を取る代わりに, 次の J を評価関数としてとる。

$$J = \frac{\text{tr } \tilde{S}_B}{\text{tr } \tilde{S}_W} = \frac{\text{tr } W^T S_B W}{\text{tr } W^T S_W W} \quad (6)$$

この関数の W による最大化は,

$$W^T S_W W = I \quad (7)$$

なる条件の下で, 分子を最大化することと等価であって, 次の評価関数の最大化問題となる。

$$E = \sum_{k=1}^{K-1} \mathbf{w}_i^T S_B \mathbf{w}_i - \sum_{k=1}^{K-1} \lambda_i (\mathbf{w}_i^T S_W \mathbf{w}_i - 1) \quad (8)$$



w_i で偏微分して 零 と置くと,

$$S_B \mathbf{w}_i - \lambda_i S_W \mathbf{w}_i = 0 \quad (9)$$

を得る。

よって, 問題は, $S_B^{-1} S_W$ の固有値を求める問題に帰着する。

即ち, $S_B^{-1} S_W$ の固有値の大きいほうから $K - 1$ 個とって, これらに対応する固有ベクトルを求めれば, これが W の各列ベクトル w_i となる。

(以上, 変動行列 S_B, S_W の代わりに, 共分散行列 C_B, C_W で用いて解く場合もある)



§ 重み付きPCA,LDA

- PCAや, LDAを Pairwise の距離を導入して表現することができる。
- Pairwise の距離を導入すると, データ間の局所的な関係に応じて重みをかけることができるので, 重み付きPCA, 重み付きLDAのような形で, 様々な特徴を持つ線形座標変換を実現できる。



データ間の距離(Pair-wise 距離)と分散の関係

$$\frac{1}{N^2} \sum_{i=1}^N \sum_{j=i+1}^N (x_i - x_j)^2 = \frac{1}{2N^2} \sum_{i=1}^N \sum_{j=1}^N (x_i - x_j)^2 \quad (1)$$

$$= \frac{1}{2N^2} \left(2N \sum_{i=1}^N x_i^2 - 2 \sum_{i=1}^N x_i \sum_{j=1}^N x_j \right)$$

$$= \frac{1}{N} \sum_{i=1}^N x_i^2 - \frac{1}{N} \sum_{i=1}^N x_i \frac{1}{N} \sum_{j=1}^N x_j$$

$$= \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

データ間の距離の二乗を
あらゆる組み合わせに対
し和をとったものは分散



PCA の Pair-wise 距離表現

$$z_i = \mathbf{x}_i^T \mathbf{w}$$

$$E = \sum_{i=1}^N \sum_{j=i+1}^N (z_i - z_j)^2 = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \mathbf{w}^T (\mathbf{x}_i - \mathbf{x}_j) (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{w} \quad (2)$$

$$= \mathbf{w}^T \left(N \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{w} - \mathbf{w}^T \left(\sum_{i=1}^N \sum_{j=1}^N \mathbf{x}_j \mathbf{x}_i^T \right) \mathbf{w}$$

$$= \mathbf{w}^T X B X^T \mathbf{w} - \mathbf{w}^T X \mathbf{1} X^T \mathbf{w} = \mathbf{w}^T X C X^T \mathbf{w} \quad (3)$$

$$B = N I, \quad C = B - \mathbf{1} \quad (4)$$

- 主成分分析は、上記 $X C X^T$ の最大固有値に対する、固有ベクトルを求める問題となる。



重みつき分散の導入

□ ここで, i, j で決まる重み $a_{ij} (= a_{ji})$ を導入する。

$$z_i = \mathbf{x}_i^T \mathbf{w}$$

$$E = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_{ij} (z_i - z_j)^2 = N \sum_{i=1}^N \bar{a}_i z_i^2 + \sum_{i=1}^N \sum_{j=1}^N a_{ij} z_i z_j \quad (5)$$

$$= \mathbf{w}^T \left(N \sum_{i=1}^N \bar{a}_i \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{w} - \mathbf{w}^T \left(\sum_{i=1}^N \sum_{j=1}^N a_{ij} \mathbf{x}_j \mathbf{x}_i^T \right) \mathbf{w}$$

$$= \mathbf{w}^T X D X^T \mathbf{w} - \mathbf{w}^T X A X^T \mathbf{w} = \mathbf{w}^T X L X^T \mathbf{w} \quad (6)$$

$$D = (d_{ij}), d_{ii} = \bar{a}_i = \frac{1}{N} \sum_{j=1}^N a_{ij}, d_{ij(\neq i)} = 0, A = (a_{ij}), L = D - A \quad (7)$$

□ 重み付き分散を最大化するベクトルは, $X L X^T$ の最大固有値に対する固有ベクトルとして求められる。



重みつき分散の導入

□ 例えば,

$$a_{ij} = \exp - \frac{\|x_i - x_j\|^2}{t} \quad (8)$$

のように, a_{ij} を x_i と x_j との距離が大きい時小さく, 距離が小さいとき1となるようにとって, 前ページ E を最小化すれば, 変換前の座標系で近くに分布しているデータ同士の写像後の距離関係を保存することができる。

⇒ Locality Preserving Projection



重みつき分散比の導入

$$z_i = \mathbf{x}_i^T \mathbf{w}$$

$$E = \frac{\sum_{i=1}^N \sum_{j=1}^N a_{ij} (z_i - z_j)^2}{\sum_{i=1}^N \sum_{j=1}^N b_{ij} (z_i - z_j)^2} \quad (9)$$

$$= \frac{\mathbf{w}^T \left(N \sum_{i=1}^N \bar{a}_i \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{w} - \mathbf{w}^T \left(\sum_{i=1}^N \sum_{j=1}^N a_{ij} \mathbf{x}_j \mathbf{x}_i^T \right) \mathbf{w}}{\mathbf{w}^T \left(N \sum_{i=1}^N \bar{b}_i \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{w} - \mathbf{w}^T \left(\sum_{i=1}^N \sum_{j=1}^N b_{ij} \mathbf{x}_j \mathbf{x}_i^T \right) \mathbf{w}}$$

$$= \frac{\mathbf{w}^T X D X^T \mathbf{w} - \mathbf{w}^T X A X^T \mathbf{w}}{\mathbf{w}^T X G X^T \mathbf{w} - \mathbf{w}^T X B X^T \mathbf{w}} = \frac{\mathbf{w}^T X L X^T \mathbf{w}}{\mathbf{w}^T X M X^T \mathbf{w}} \quad (10)$$

$$G = (g_{ij}), g_{ii} = \bar{b}_i = \frac{1}{N} \sum_{j=1}^N b_{ij}, g_{ij(\neq i)} = 0, B = (b_{ij}), M = G - B \quad (11)$$

LDA の Pair-wise 距離表現

$$a_{ij} = \begin{cases} 1 & \dots \mathbf{x}_i \text{ と } \mathbf{x}_j \text{ が違うクラス} \\ 0 & \dots \mathbf{x}_i \text{ と } \mathbf{x}_j \text{ が同じクラス} \end{cases}$$

$$b_{ij} = \begin{cases} 0 & \dots \mathbf{x}_i \text{ と } \mathbf{x}_j \text{ が違うクラス} \\ 1 & \dots \mathbf{x}_i \text{ と } \mathbf{x}_j \text{ が同じクラス} \end{cases}$$

として, 前ページ E を最大化すれば,
すなわち, $(XMX^T)^{-1}XLX^T$ の固有値問題を解けば,
LDAと等価。

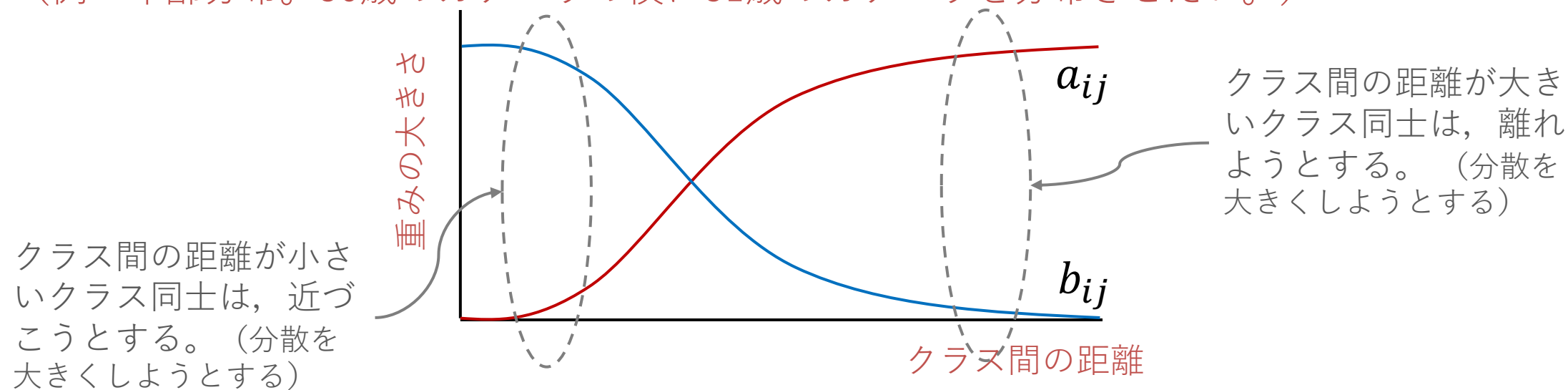


重みつき分散比とCDDA(クラス間距離を考慮した判別分析)

幾つかの多クラスの問題では、

クラス間の距離に応じて、データを分布させたいことがある。

(例：年齢分布。30歳のカテゴリの横に31歳のカテゴリを分布させたい。)

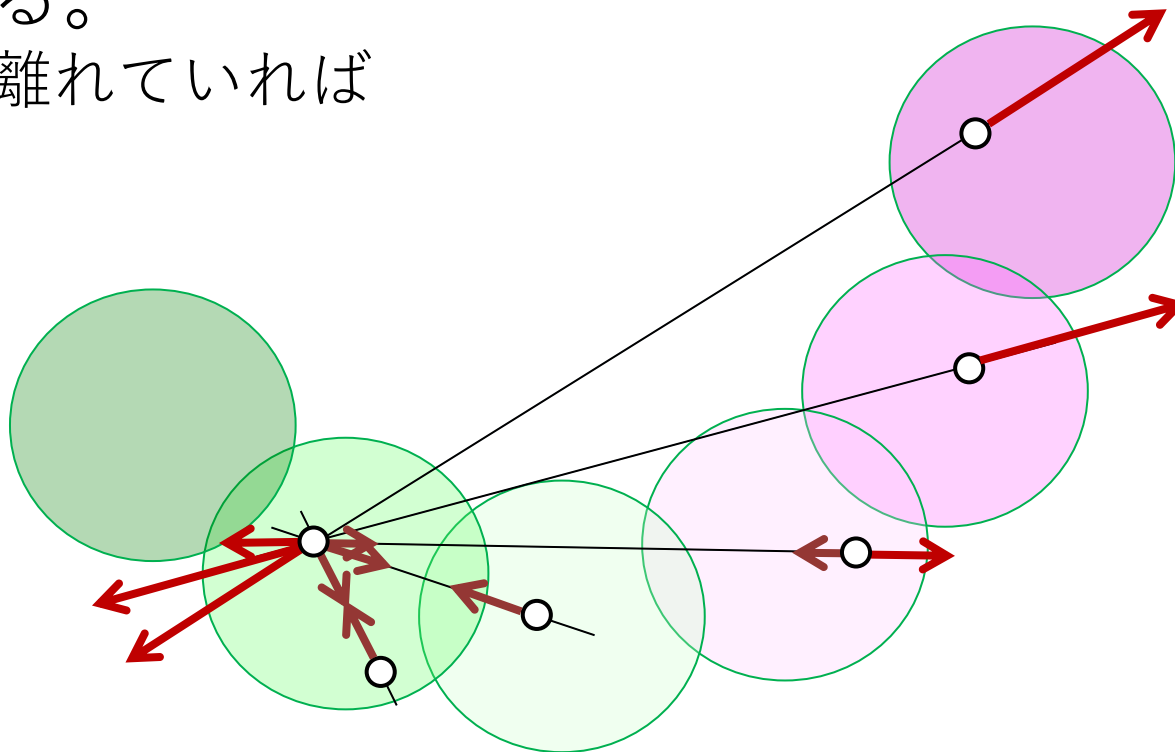


重みを上の図のようにクラス間の距離に応じて定め、
前出の E を最大化すれば、
クラス間の距離の小さいものが近くに並ぶ。



CDDA : Class Distance-based Discriminant Analysis

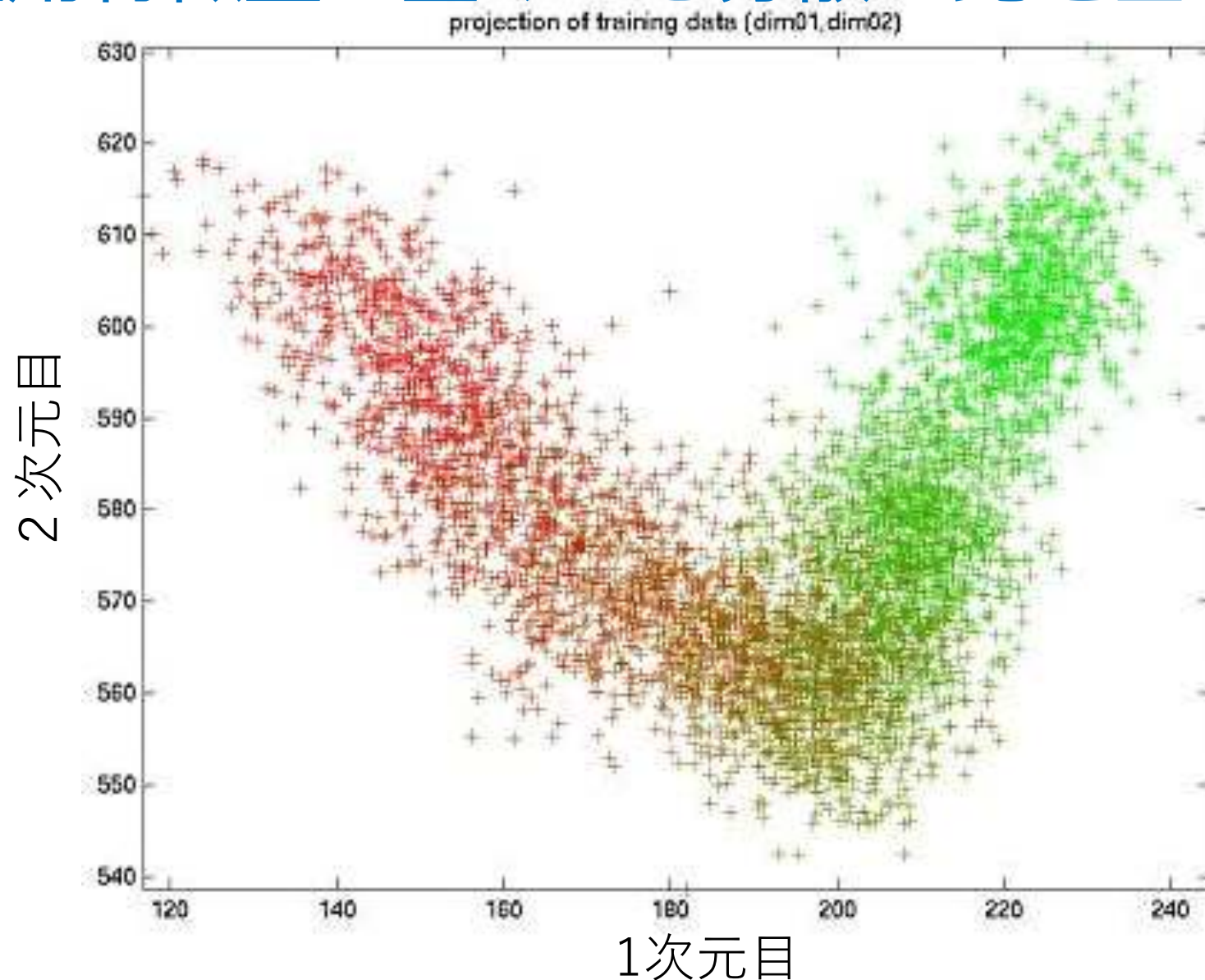
クラスが遠いデータは
離れようとする。
ただし、十分に離れていれば
差はつけない。



クラスが近いデータは、近づこうとする
ただし、十分に近いデータには、差をつけない



年齢推定用特徴量：重みつき分散の比を基に次元圧縮



まとめ

- 一般に、高次元のデータをそのまま扱うことが困難な場合、（パターン認識における次元の呪いなど）、扱うデータの次元を予め落とすため、**情報圧縮（次元削減）**が行われる。
- 情報のロスを抑えて（データの分散をできるだけ広くとって）次元を削減する線形写像を、**主成分分析（PCA）**と呼ぶ。
- クラス間の分離度（級間分散と級内分散の比）を最大化する基準で次元を削減する線形写像を、**線形判別分析（LDA）**と呼ぶ。
- Pairwise の距離を導入することで、様々な特徴を持つ線形座標変換を実現できる。



§

演習問題

1. 次のような5つのデータが与えられたとする。

$(-3, -2), (1, -1), (0, 0), (3, 2), (-1, 1)$

- 1) 平均値ベクトル μ , および共分散行列 C を求めよ。(ただし, 共分散行列の算出において, 偏差の二乗平均を求める際, データ数-1でなく, データ数で割ることにする。)
- 2) C の最大固有値と, その固有値に対する固有ベクトルを求めよ。
- 3) データの散布図を描け。また, この散布図に重ねて 2) で求めた固有ベクトルを描け。

演習問題

2. クラス α のデータとして, 4つのデータ,

$$\mathbf{x}_1^\alpha = \begin{pmatrix} -5 \\ 0 \end{pmatrix}, \mathbf{x}_2^\alpha = \begin{pmatrix} 1 \\ 4 \end{pmatrix}, \mathbf{x}_3^\alpha = \begin{pmatrix} -2 \\ 1 \end{pmatrix}, \mathbf{x}_4^\alpha = \begin{pmatrix} -2 \\ 3 \end{pmatrix}$$

が与えられ, クラス β のデータとして, 4つのデータ,

$$\mathbf{x}_1^\beta = \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \mathbf{x}_2^\beta = \begin{pmatrix} 2 \\ -3 \end{pmatrix}, \mathbf{x}_3^\beta = \begin{pmatrix} 5 \\ 0 \end{pmatrix}, \mathbf{x}_4^\beta = \begin{pmatrix} -1 \\ -4 \end{pmatrix}$$

が与えられている。

- 1) クラス毎に, 平均ベクトル μ_α, μ_β , 共分散行列 C_α, C_β を求めよ。
- 2) 級内分散共列 C_W を, 2つのクラスの共分散行列の平均として定義するとき, C_W を求めよ。

- 3) 級間分散行列 C_B を, それぞれのクラスの平均ベクトルの共分散行列として定義するとき, C_B を求めよ。
- 4) ベクトル $\mathbf{w} = (w_1 \ w_2)^T$ を用いて, $z_i = \mathbf{x}_i^T \mathbf{w}$ と座標変換するとき, 変換後の座標空間における級間分散 σ_B^2 と級内分散 σ_W^2 を, C_B, C_W, \mathbf{w} , の式で表せ。
- 5) 級間分散 σ_B^2 と級内分散 σ_W^2 の比 σ_B^2 / σ_W^2 を最大化するように \mathbf{w} を定めるとき, 変換後の座標空間において, 同じクラスのデータが纏まり, 違うクラスのデータが離れる分布を作るため, 座標空間は識別に適したものとなる。このような空間を作る, C_B, C_W, \mathbf{w} , の関係を導け。
- 6) 5)の条件を満たす \mathbf{w} を実際に求めよ。
- 7) データの散布図を描け。この散布図に重ねて \mathbf{w} を描け。