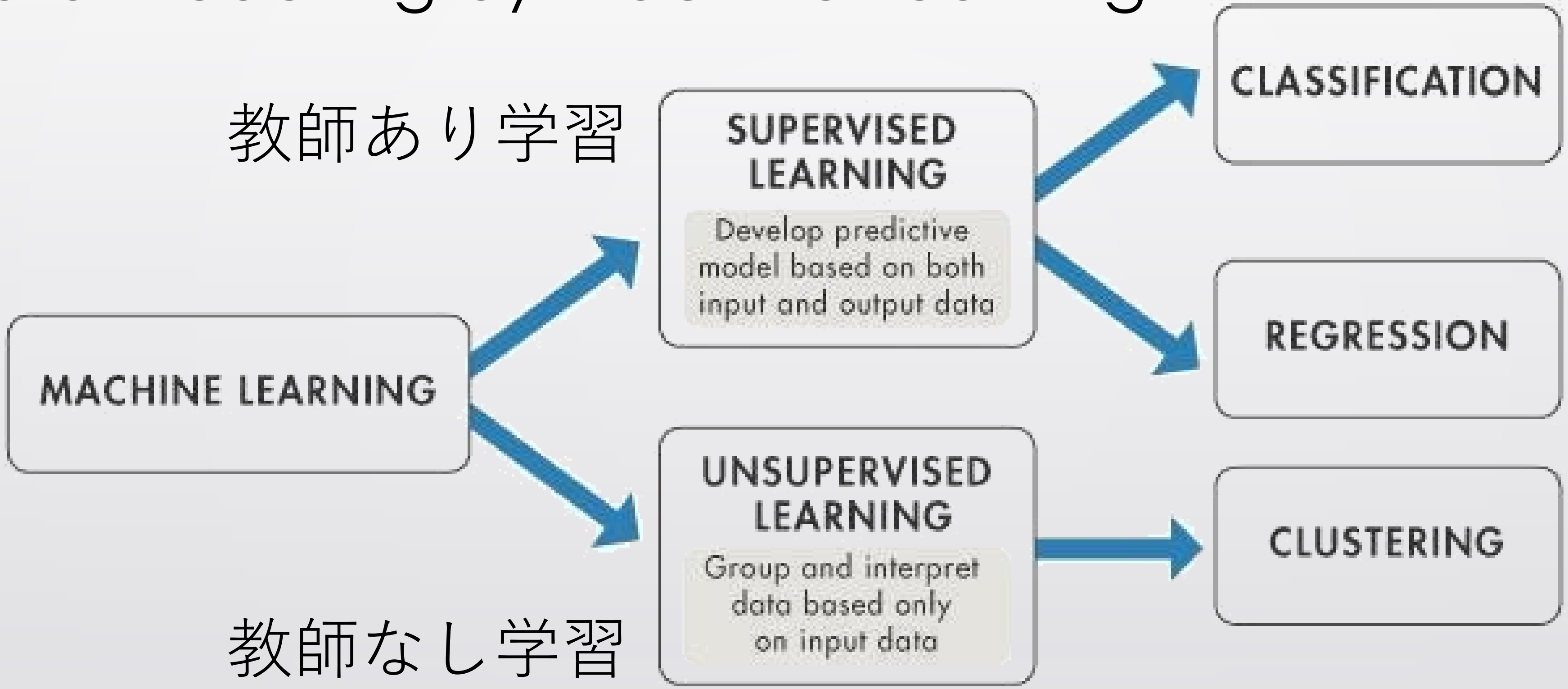
データマイニング
Data Mining

6: 分類① Classification

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

教師なし学習

分類

回帰

クラスタリング

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

変数の種類 Types of Variables

- カテゴリー変数 Categorical Variable
  - 名義尺度 Nominal Variable
  - 順序尺度 Ordinal Variable
- 量的変数 Numerical Variable
  - 間隔尺度 Interval Variable
  - 比例尺度 Ratio Variable

変数の種類 Types of Variables

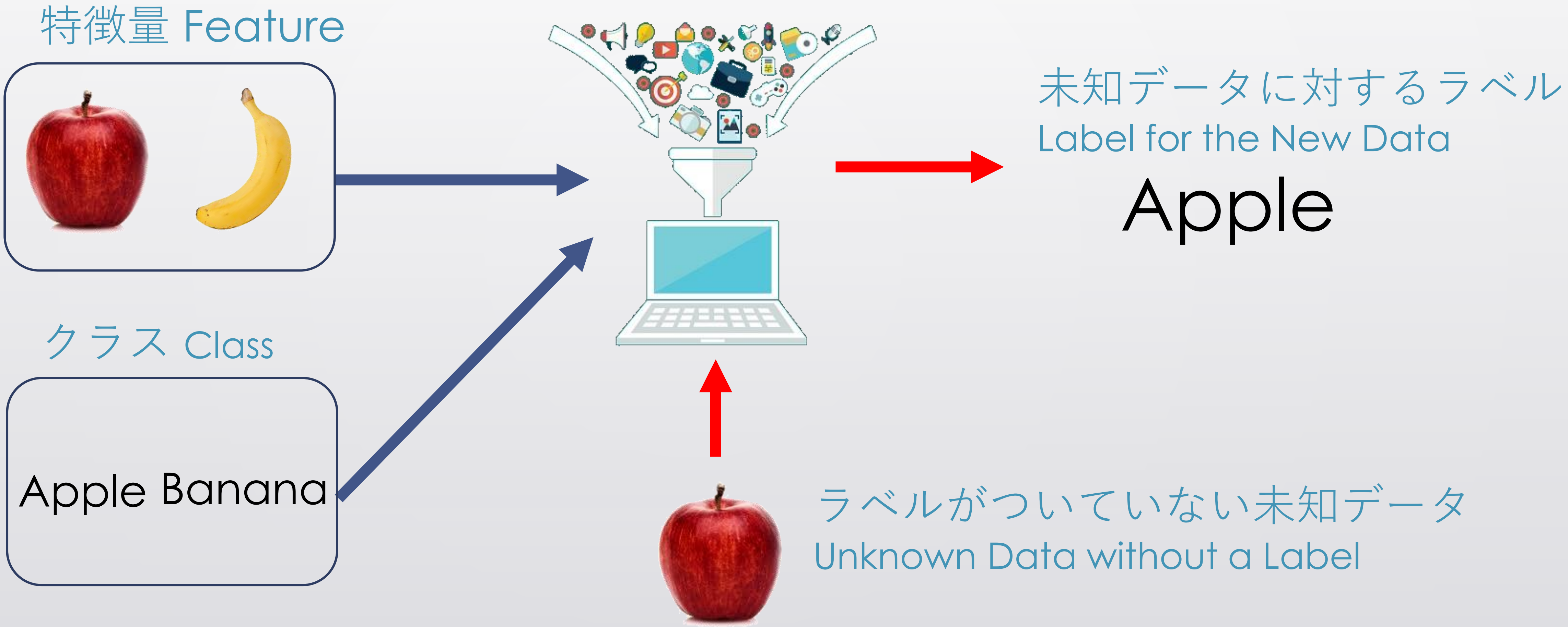
名義尺度
Nominal Variable

あるカテゴリーを、別のカテゴリーと区別するために用いられる、数値自体には意味がない変数

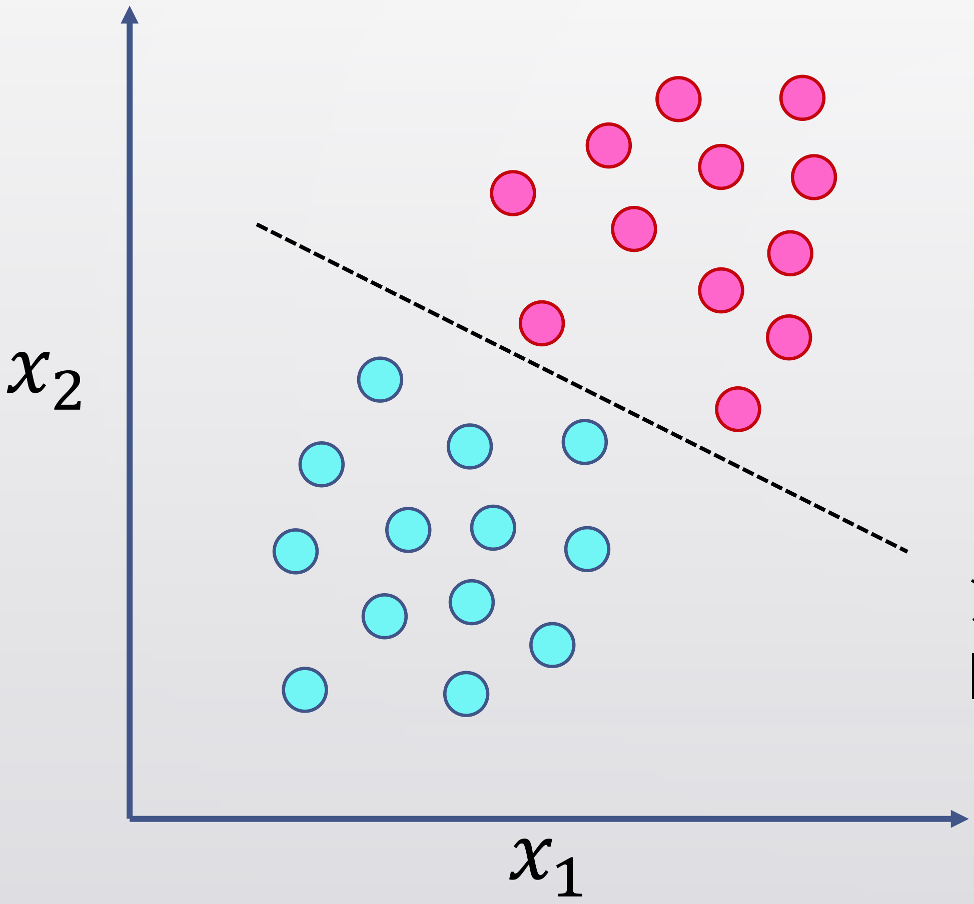
Variables, whose number has no numerical value, often used to discriminate multiple categories

ユーザーID	性別	年齢	年収
sanapon	男	26	411万円
oggi1985	女	33	536万円
murachan	女	39	681万円
shozan.s	男	23	309万円

教師あり学習 Supervised Learning



線型判別分析 Linear Discriminant Analysis (LDA)

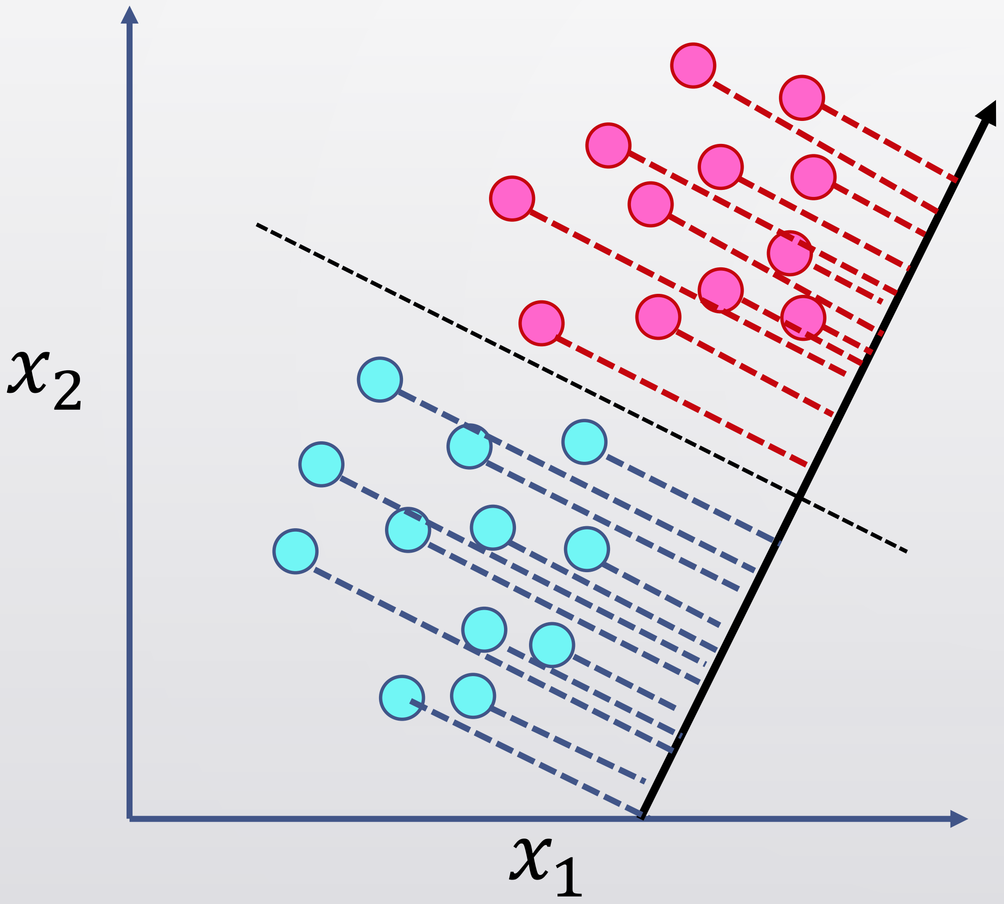


クラス1とクラス2を分離できる決定境界の引き方を見つける

Find how to draw a decision boundary that separates Class 1 and Class 2

決定境界
Decision Boundary

線型判別分析 Linear Discriminant Analysis (LDA)

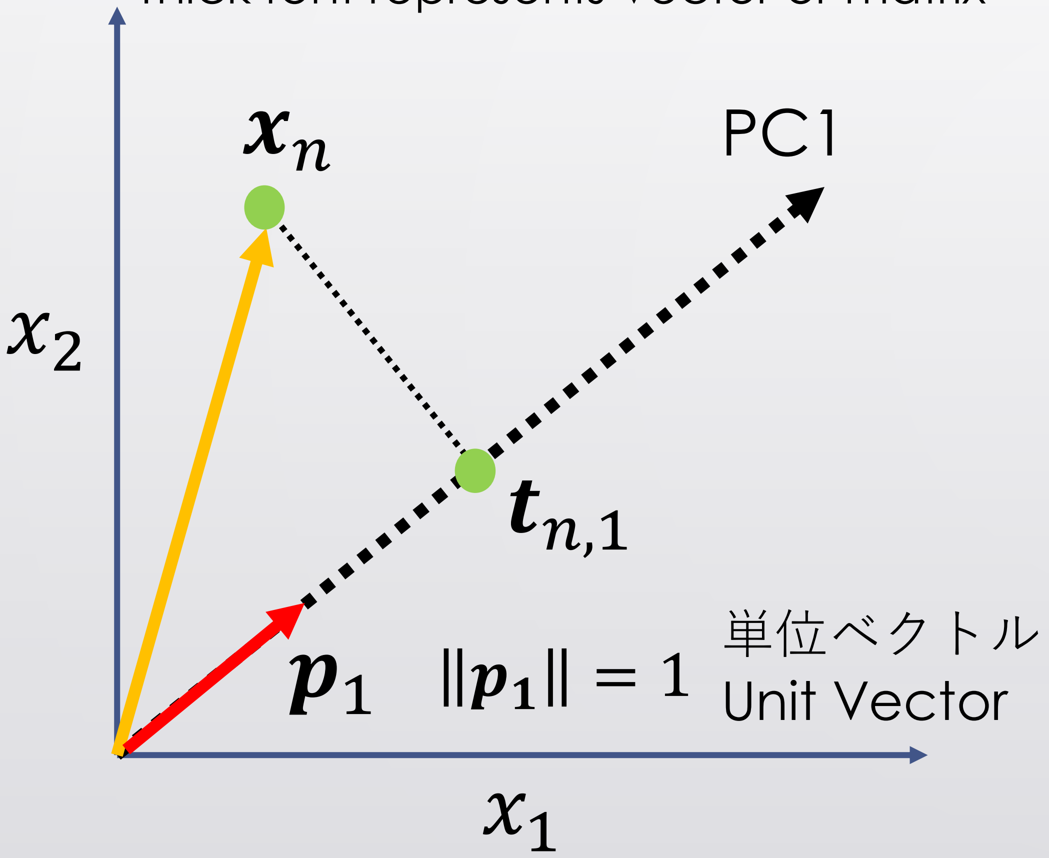


決定境界に直交する軸へのデータの射影を計算する

Consider the projection of data onto axis orthogonal to the decision boundary

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



変数を標準化しておく

Normalize the variables

観測データ$\boldsymbol{x}_n$の第1主成分軸方向への

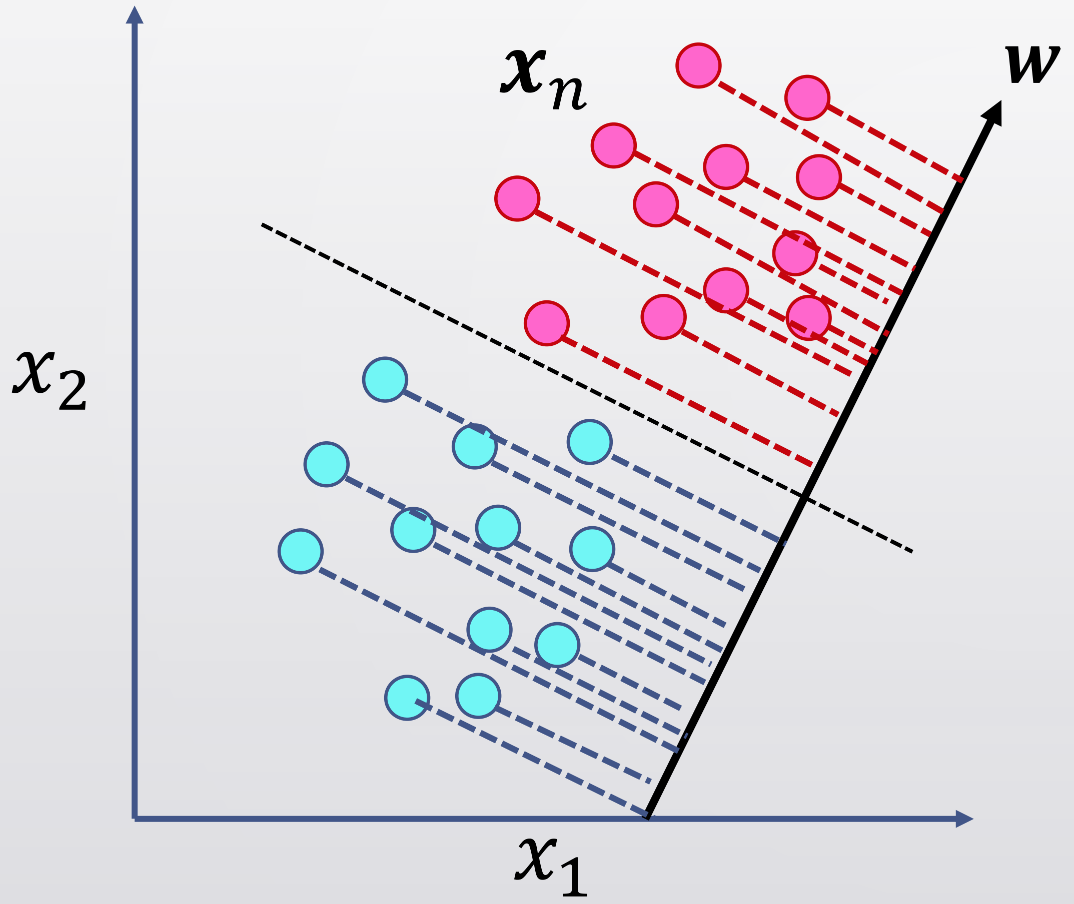
射影$\boldsymbol{t}_{n,1}$を計算する

Compute the projection $t_{n,1}$ of the observed data x_n onto the first PC axis

$t_{n,1}$はx_nとp_1の内積

$t_{n,1}$is dot(inner) product of x_n and p_1

線型判別分析 Linear Discriminant Analysis (LDA)



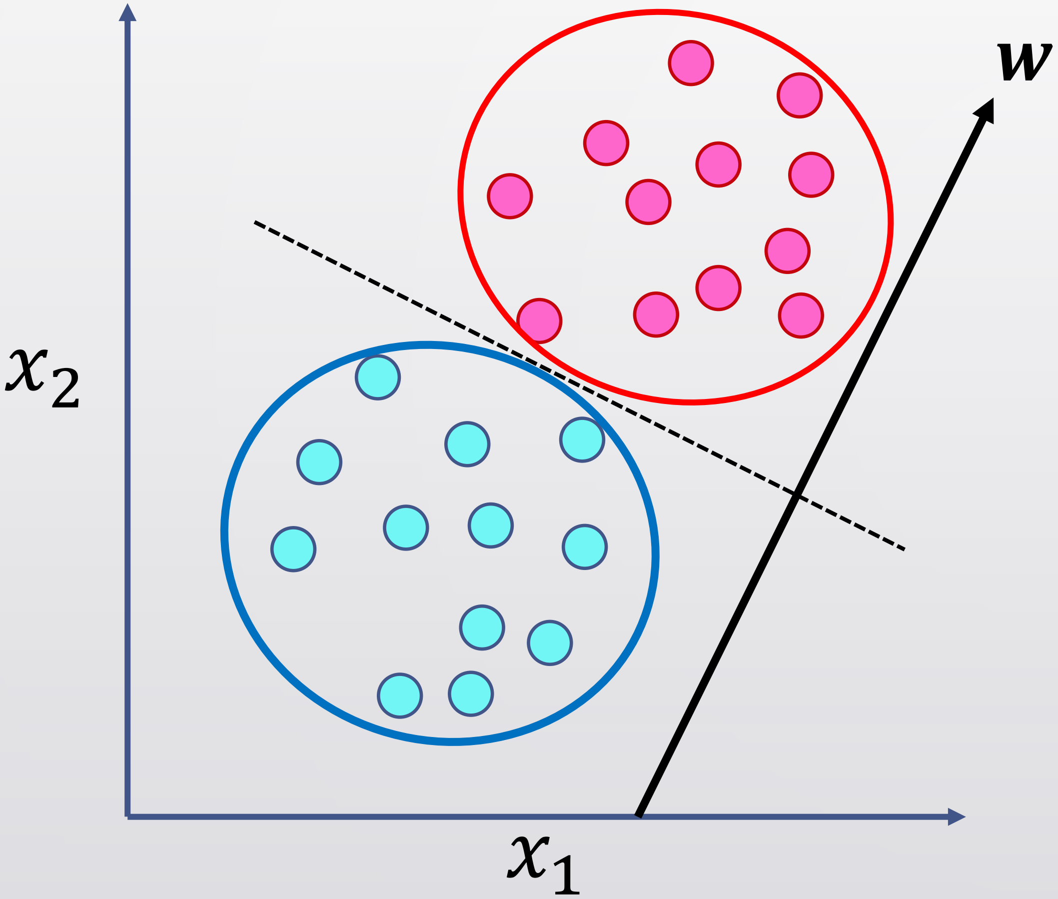
$$\boldsymbol{x}_n = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} \quad \boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_M \end{bmatrix} \quad \|\boldsymbol{w}\| = 1$$

$\boldsymbol{y}_n$は$\boldsymbol{x}_n$の軸$\boldsymbol{w}$への射影

$\boldsymbol{y}_n$ is projection of $\boldsymbol{x}_n$ onto axis $\boldsymbol{w}$

$$\boldsymbol{y}_n = \boldsymbol{w}^T \boldsymbol{x}_n$$

線型判別分析 Linear Discriminant Analysis (LDA)



よい決定境界は、下の二つの条件を満たす

A good decision boundary meets the two conditions below

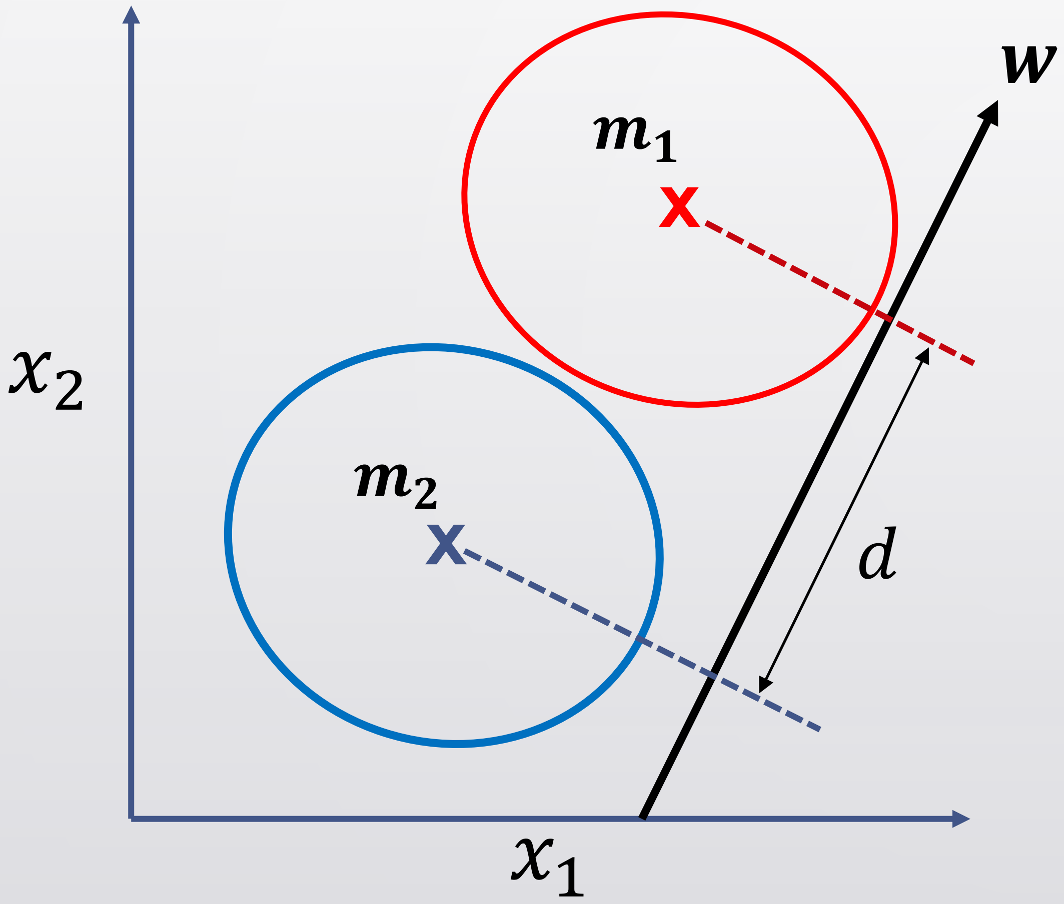
1. 2 クラスの中心が離れている

   Centers of the two classes are distant from each other

2. 各クラスのクラス内分散が小さい

   Within-class variance of each class is small

線型判別分析 Linear Discriminant Analysis (LDA)



1. 2クラスの中心が離れている

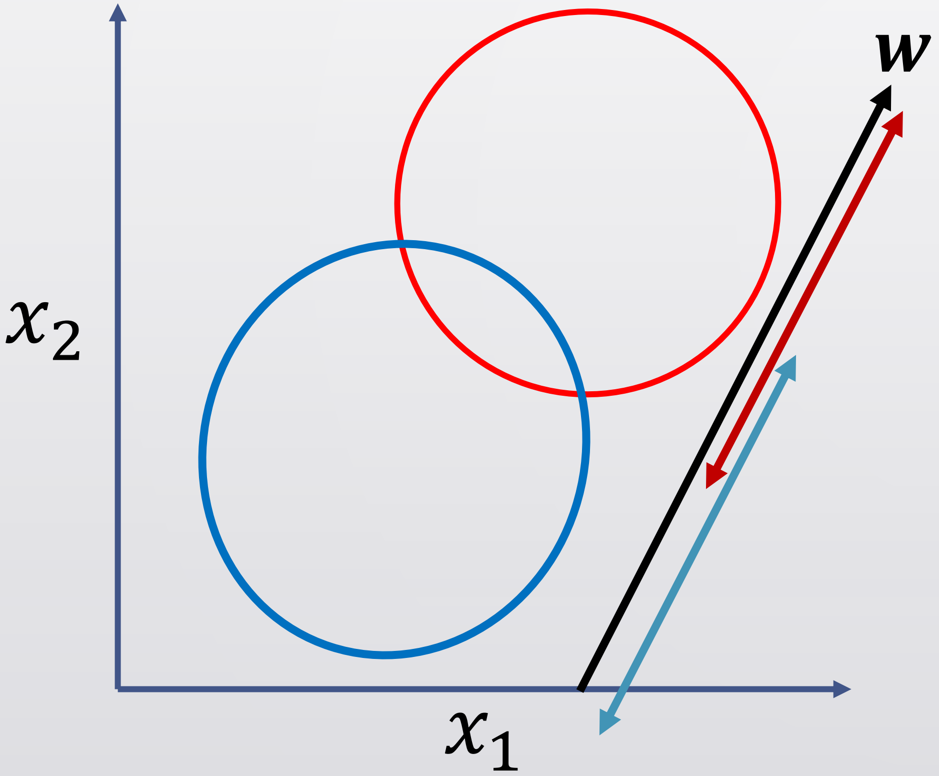
   Centers of the two classes are distant from each other

$$\boldsymbol{m_1} = \frac{1}{N_1} \sum_{\boldsymbol{x_k} \in C_1} \boldsymbol{x_k} \qquad \boldsymbol{m_2} = \frac{1}{N_2} \sum_{\boldsymbol{x_k} \in C_2} \boldsymbol{x_k}$$

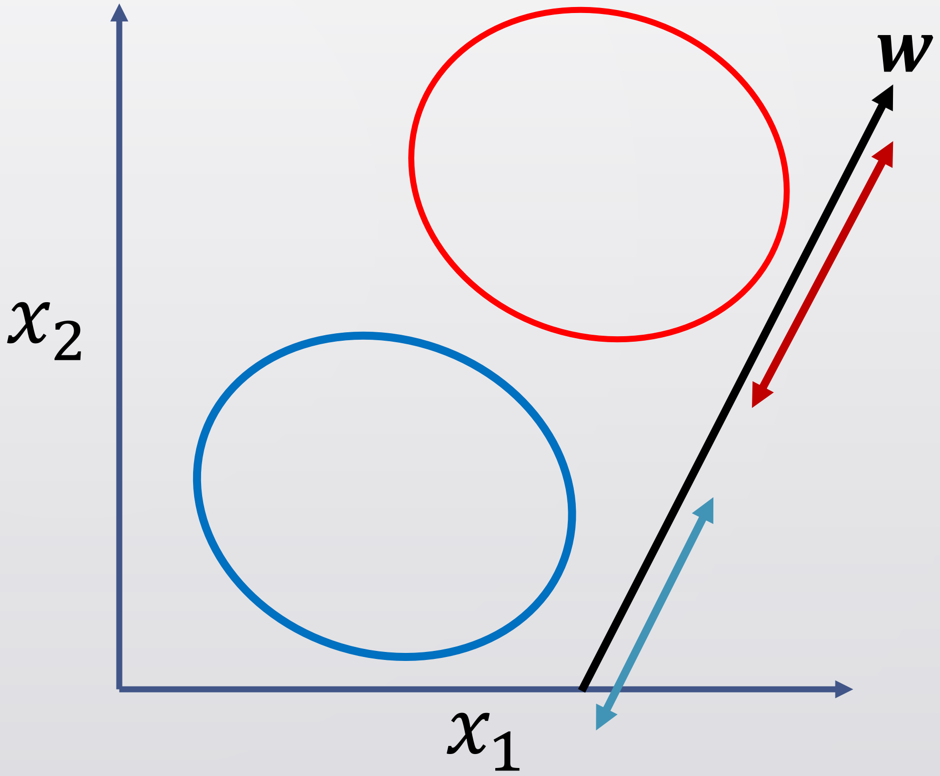
$$d = \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{m_1} - \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{m_2} = \boldsymbol{w}^{\boldsymbol{T}} (\boldsymbol{m_1} - \boldsymbol{m_2})$$

線型判別分析 Linear Discriminant Analysis (LDA)

クラス内分散 大
Large within-class variance



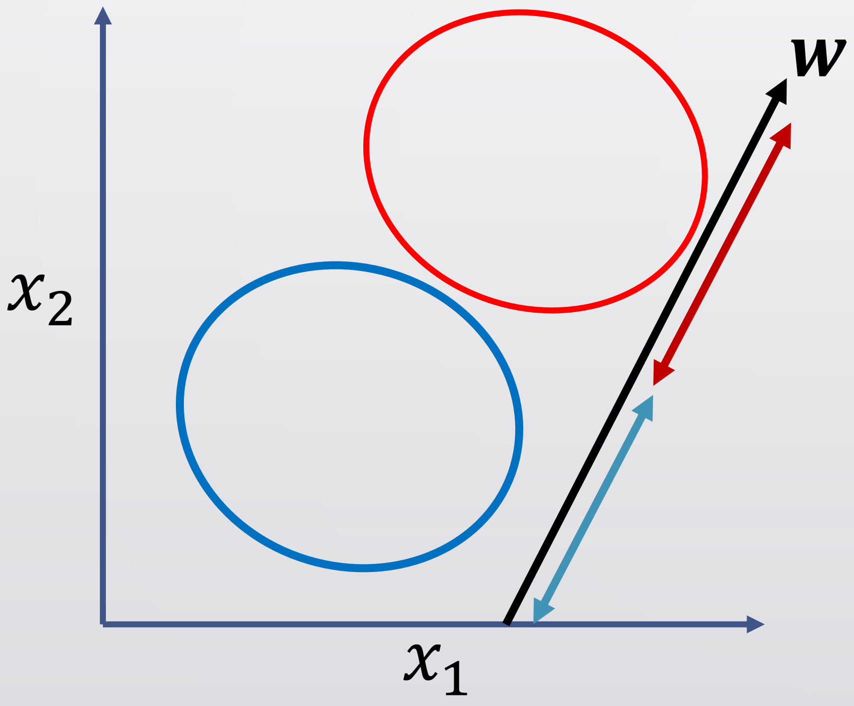
クラス内分散 小
Small within-class variance

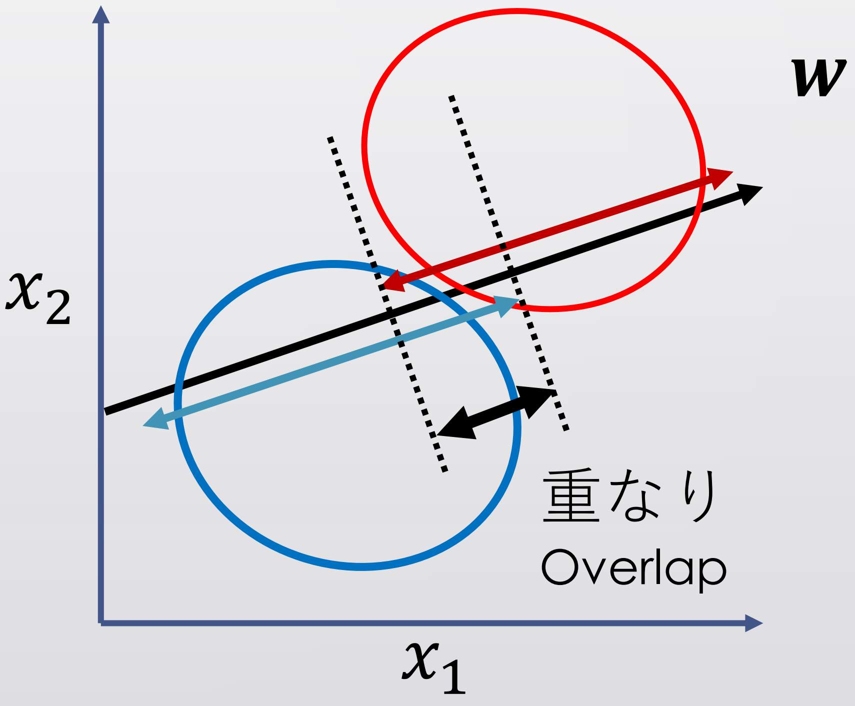


線型判別分析 Linear Discriminant Analysis (LDA)

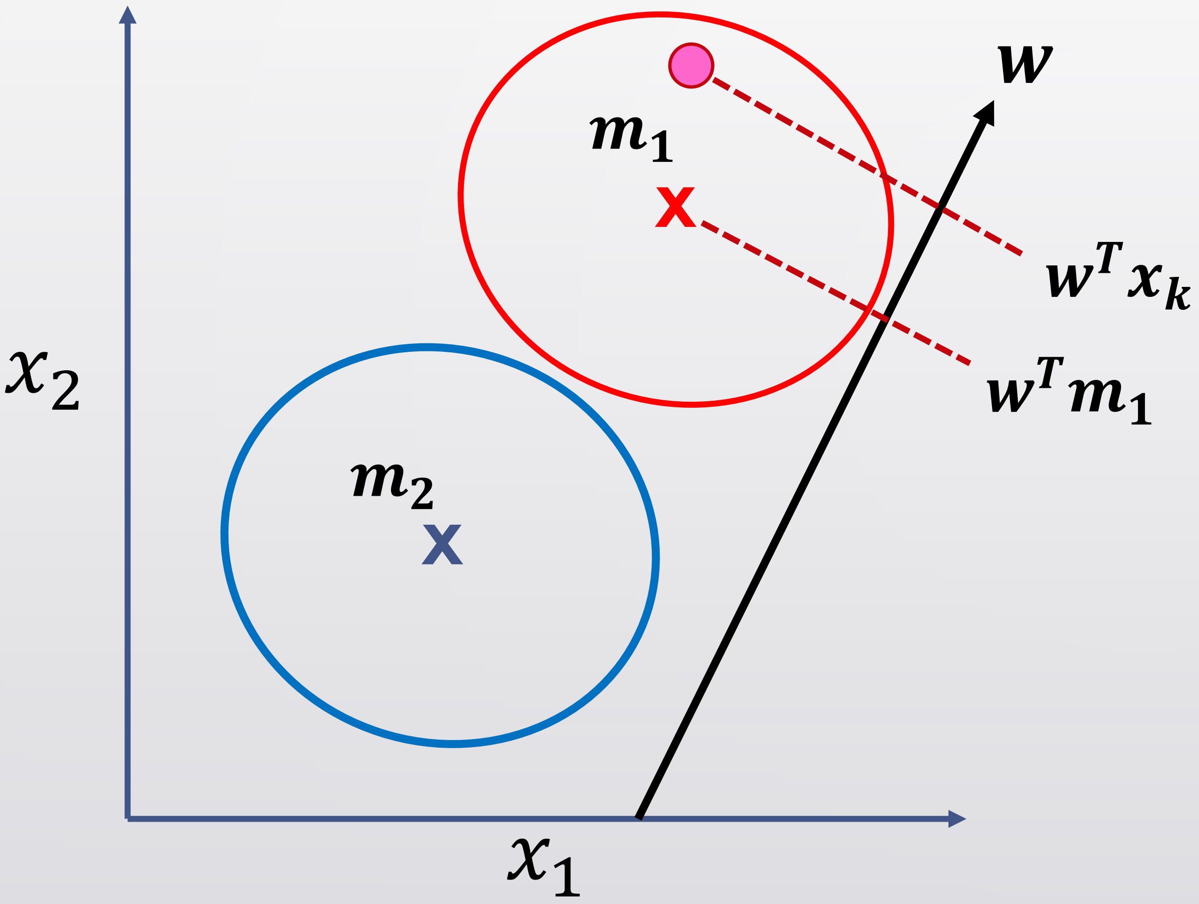
射影のクラス内分散は軸$\boldsymbol{w}$の向きにより変化する

Within-class variance of projection is dependent on the direction of axis $\boldsymbol{w}$





線型判別分析 Linear Discriminant Analysis (LDA)



$$s_1^2 = \frac{1}{N_1} \sum_{\boldsymbol{x_k} \in C_1} \left(\boldsymbol{w^T x_k} - \boldsymbol{w^T m_1}\right)^2$$

$$s_2^2 = \frac{1}{N_2} \sum_{\boldsymbol{x_k} \in C_2} \left(\boldsymbol{w^T x_k} - \boldsymbol{w^T m_2}\right)^2$$

$$s^2 = s_1^2 + s_2^2$$

線型判別分析 Linear Discriminant Analysis (LDA)

1. 2 クラスの中心が離れている Centers of the two classes are distant from each other

d を最大化する Maximize d

$$d = \boldsymbol{w}^T \boldsymbol{m}_1 - \boldsymbol{w}^T \boldsymbol{m}_2 = \boldsymbol{w}^T (\boldsymbol{m}_1 - \boldsymbol{m}_2)$$

2.各クラスのクラス内分散が小さい Within-class variance of each class is small

s^2を最小化する Minimize s^2

$$s^2 = s_1^2 + s_2^2 \quad s_j^2 = \frac{1}{N_j} \sum_{\boldsymbol{x}_k \in C_j} \left(\boldsymbol{w}^T \boldsymbol{x}_k - \boldsymbol{w}^T \boldsymbol{m}_j\right)^2$$

線型判別分析 Linear Discriminant Analysis (LDA)

1. 2 クラスの中心が離れている Centers of the two classes are distant from each other

⇒ d を最大化する ⇒ d^2 を最大化する

$$d^2 = \{\boldsymbol{w}^T(\boldsymbol{m}_1 - \boldsymbol{m}_2)\}^2 = \{\boldsymbol{w}^T(\boldsymbol{m}_1 - \boldsymbol{m}_2)\}\,\{\boldsymbol{w}^T(\boldsymbol{m}_1 - \boldsymbol{m}_2)\}^T$$

$$= \boldsymbol{w}^T(\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{S}_B\boldsymbol{w}$$

$$\boldsymbol{S}_B = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T$$

線型判別分析 Linear Discriminant Analysis (LDA)

2.各クラスのクラス内分散が小さい Within-class variance of each class is small

s^2を最小化する Minimize s^2

$$s^2 = s_1^2 + s_2^2 = \boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}$$

$$\boldsymbol{S}_w = \sum_{\boldsymbol{x}_k \in C_1} (\boldsymbol{x}_k - \boldsymbol{m}_1)(\boldsymbol{x}_k - \boldsymbol{m}_1)^T + \sum_{\boldsymbol{x}_k \in C_2} (\boldsymbol{x}_k - \boldsymbol{m}_2)(\boldsymbol{x}_k - \boldsymbol{m}_2)^T$$

線型判別分析 Linear Discriminant Analysis (LDA)

1. 2 クラスの中心が離れている Centers of the two classes are distant from each other

2. 各クラスのクラス内分散が小さい Within-class variance of each class is small

$J(w)$を最大化する Maximize $J(w)$

$$J(w) = \frac{\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{S}_{\boldsymbol{B}} \boldsymbol{w}}{\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{S}_{\boldsymbol{w}} \boldsymbol{w}}$$

線型判別分析 Linear Discriminant Analysis (LDA)

$J(w)$を最大化する Maximize $J(w)$ $\quad J(w) = \dfrac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}}$

$J(w)$を最大化する$\boldsymbol{w}$は下の固有方程式を満たす

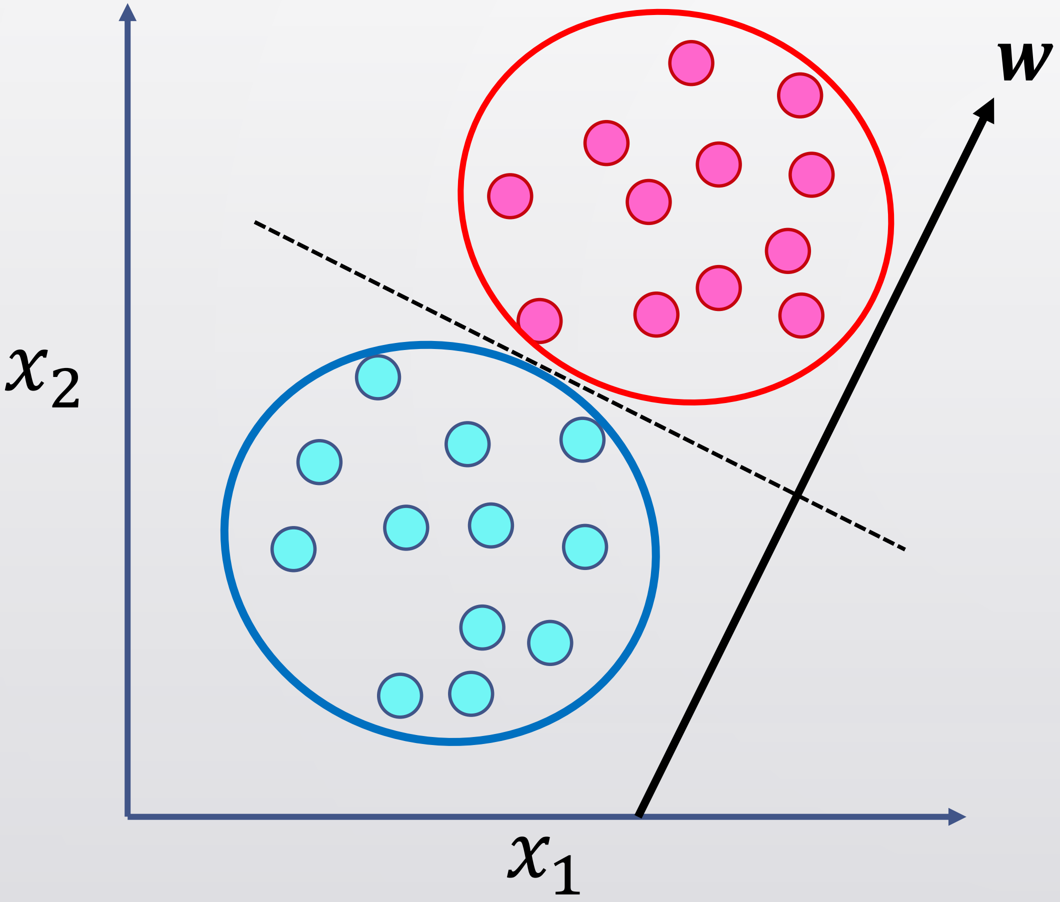
$\boldsymbol{w}$ that maximizes $J(w)$ satisfies the eigen equation below

$$\boldsymbol{S}_w^{-1} \boldsymbol{S}_B \boldsymbol{w} = \lambda \boldsymbol{w}$$

$$\boldsymbol{S}_B \boldsymbol{w} = \boldsymbol{S}_w \boldsymbol{S}_w^{-1} \boldsymbol{S}_B \boldsymbol{w} = \lambda \boldsymbol{S}_w \boldsymbol{w}$$

$$J(w) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}} = \frac{\lambda \boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}} = \lambda$$

線型判別分析 Linear Discriminant Analysis (LDA)



$$\boldsymbol{S}_{\boldsymbol{w}}^{-1}\boldsymbol{S}_{\boldsymbol{B}}\boldsymbol{w} = \lambda\boldsymbol{w}$$

$$J(w) = \frac{\boldsymbol{w}^T\boldsymbol{S}_{\boldsymbol{B}}\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_{\boldsymbol{w}}\boldsymbol{w}} = \frac{\lambda\boldsymbol{w}^T\boldsymbol{S}_{\boldsymbol{w}}\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_{\boldsymbol{w}}\boldsymbol{w}} = \lambda$$

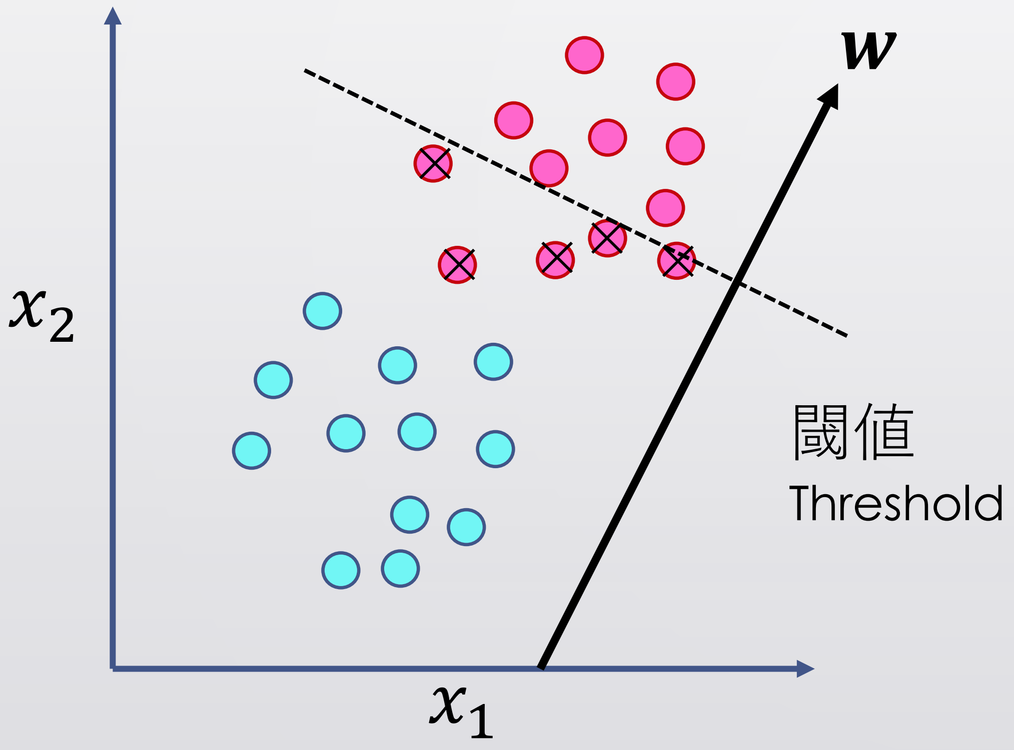
軸$\boldsymbol{w}$は最大の固有値に対応する固有ベクトルと並行

Axis $\boldsymbol{w}$ is in parallel with eigen vector corresponding to the largest eigen value

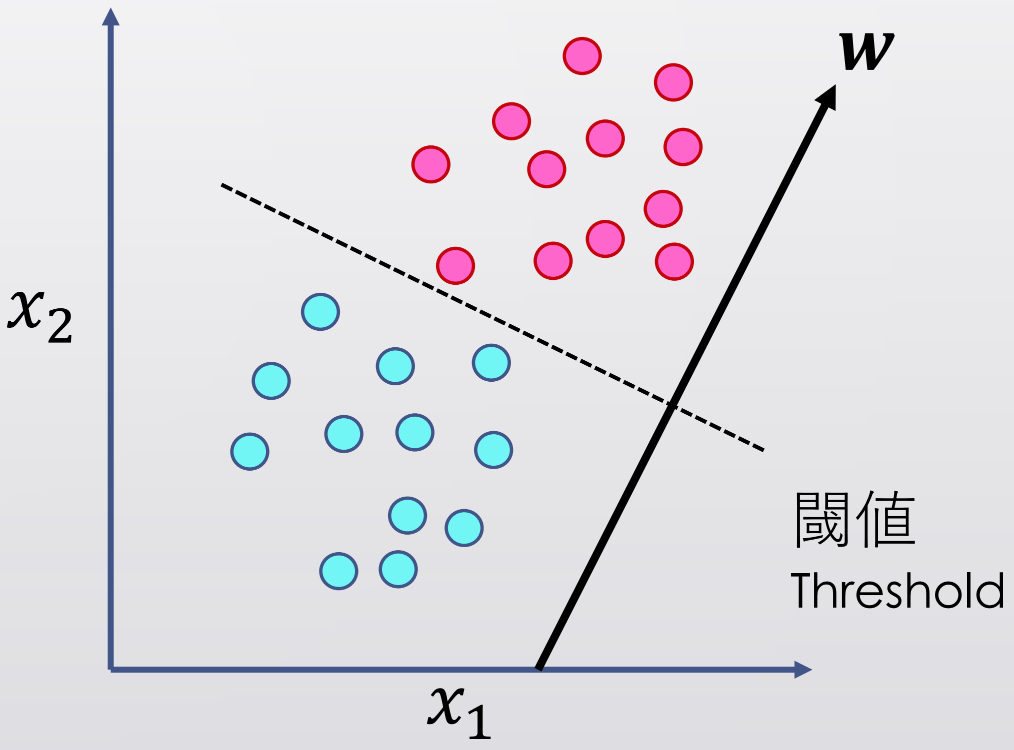
決定境界の向きが決まる
Orientation of decision boundary is determined

閾値をどう決定するか？How should we determine the threshold?

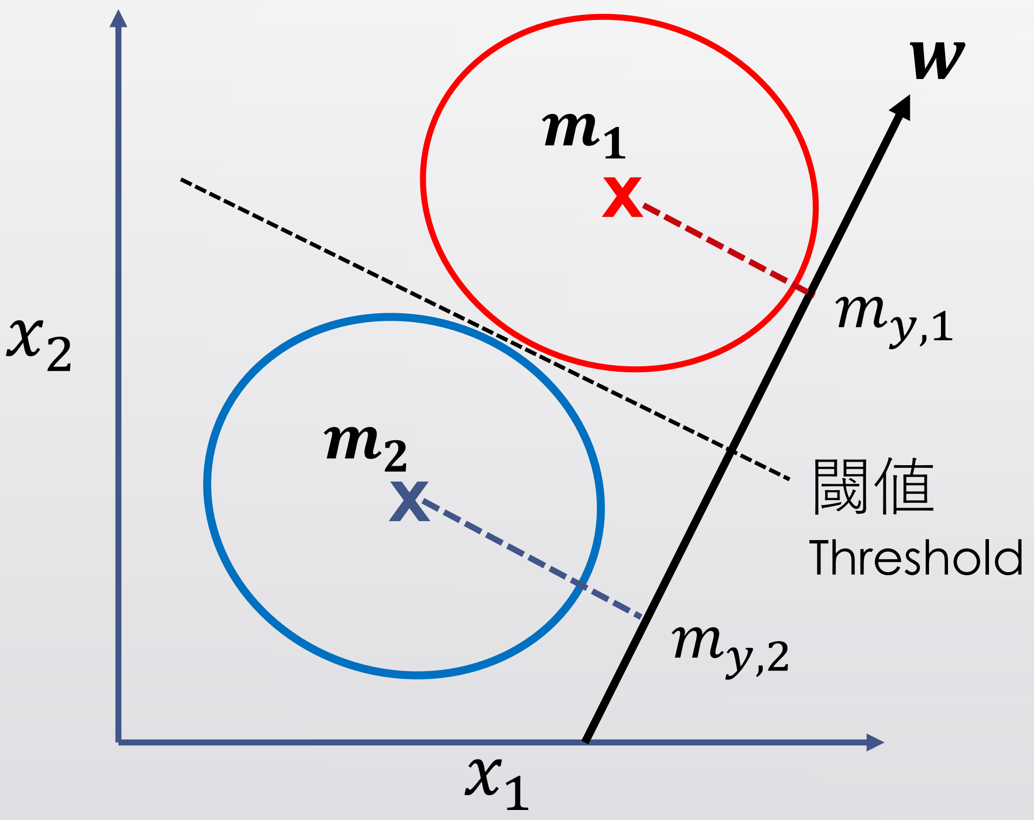
不適切な閾値 Inappropriate threshold



適切な閾値 Appropriate threshold



閾値をどう決定するか？How should we determine the threshold?



各クラスの中心の射影の加重平均を閾値にする

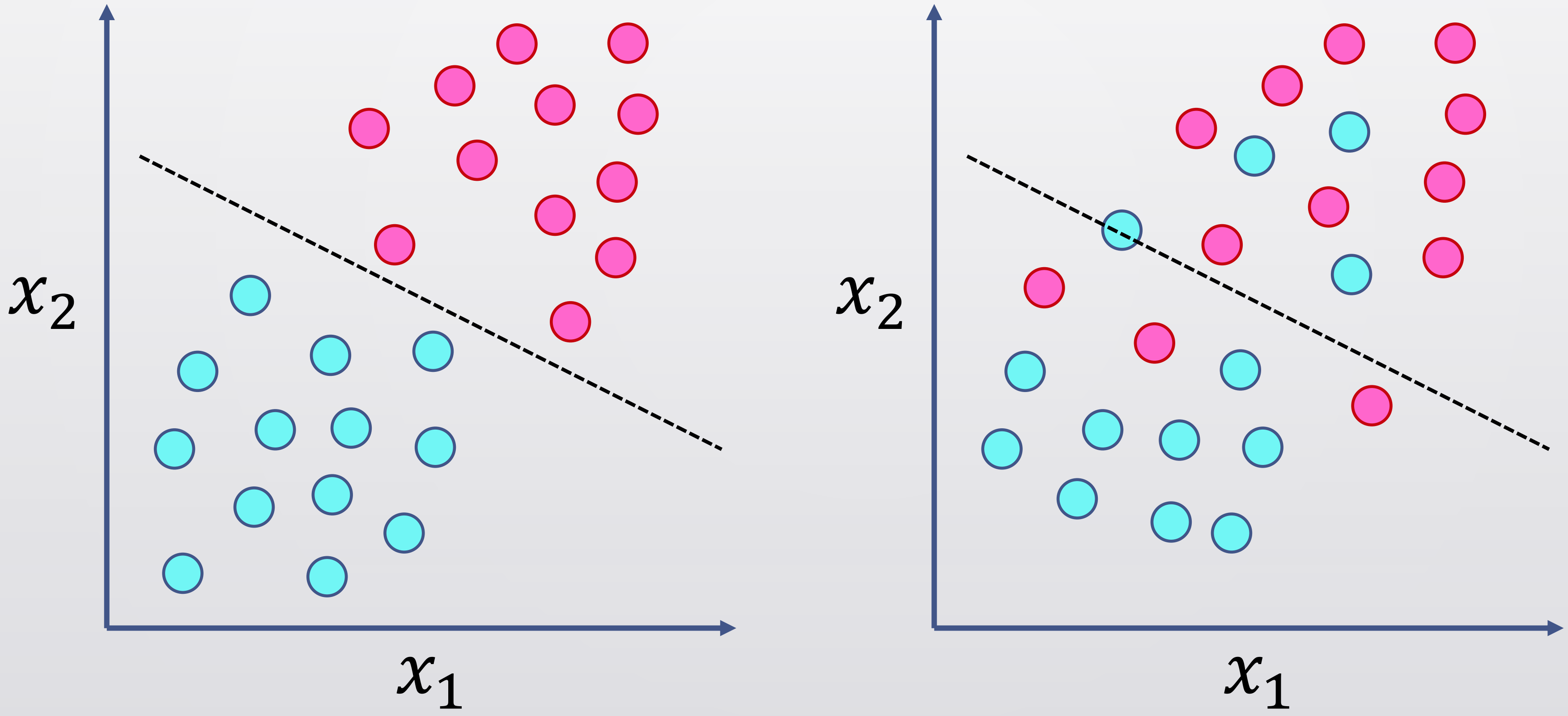
Adopt as threshold value the weighted mean of projection of center of each class

$$Threshold = \frac{N_1 s_{y,1}^2 m_{y,1} + N_2 s_{y,2}^2 m_{y,2}}{N_1 s_{y,1}^2 + N_1 s_{y,2}^2}$$

$s_{y,j}^2$: クラスjのデータの射影の分散

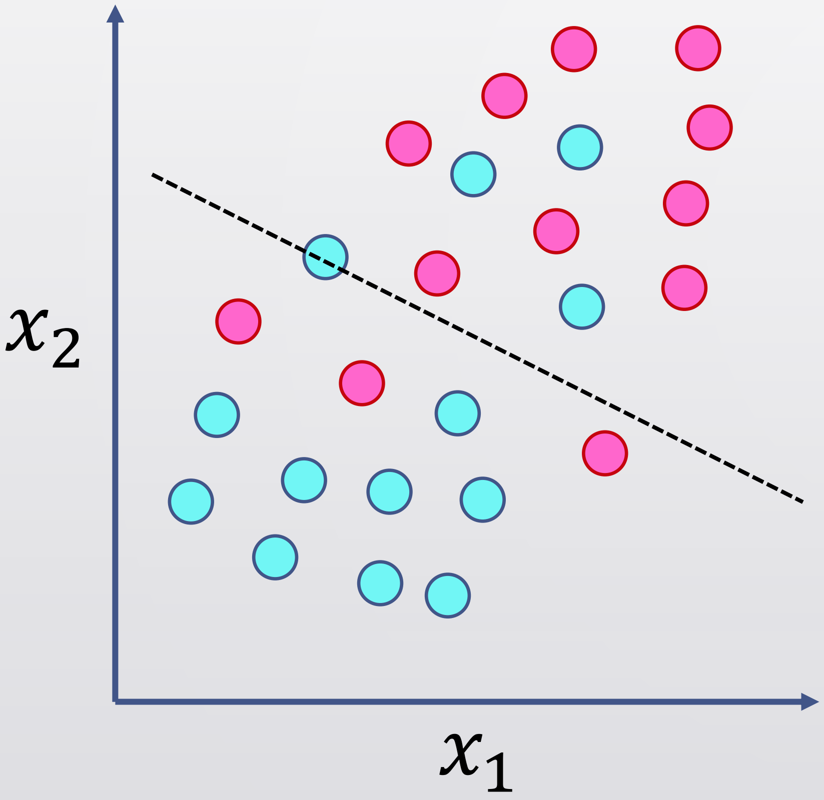
$m_{y,j}$: クラスjの重心の射影

"線型分離可能"とは？
What does "Linearly Separable" Mean?



“線型分離可能”とは？

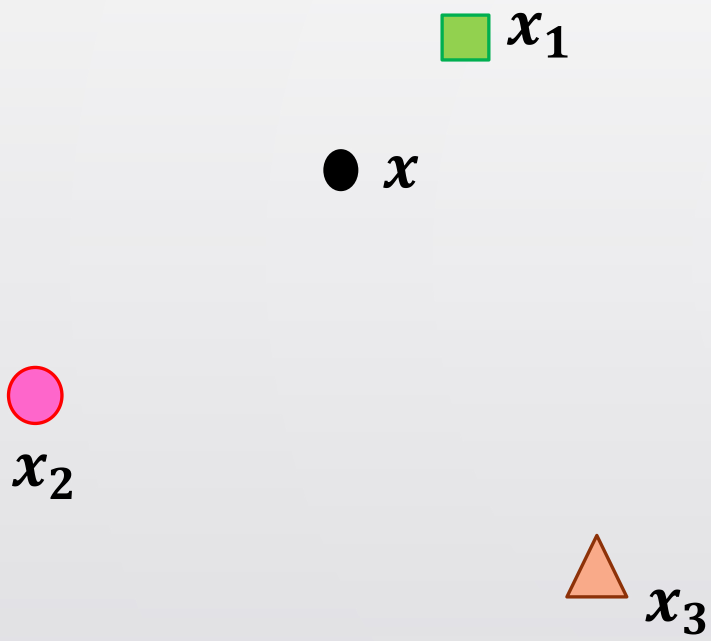
What does “Linearly Separable” Mean?



線型分離可能でない問題には、LDAがうまく機能しない

LDA does not work well for linearly inseparable problems

最近傍法 Nearest Neighbor Method



データx　は最も近くにある鋳型データx_jと同じクラスに属するとみなす

Data x is judged to belong to the same class as template data x_j

最近傍法 Nearest Neighbor Method

クラスC_iの鋳型と$\boldsymbol{x}$の最小距離

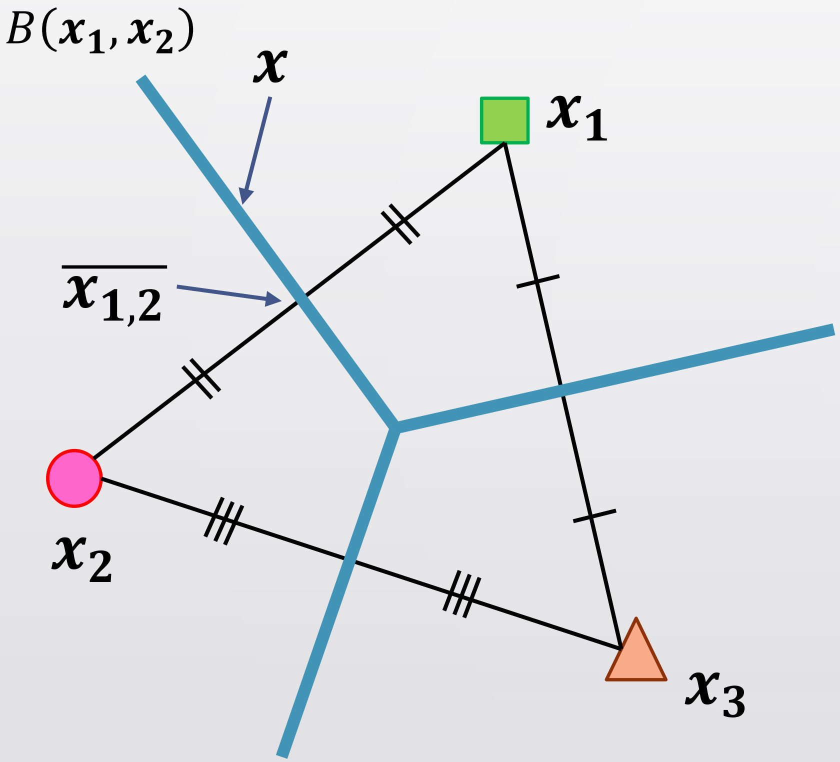
Shortest distance between $\boldsymbol{x}$ and templates of class C_i

$$\begin{cases} \underline{argmin_i \left\{ min_j \ \ d\left(\boldsymbol{x}, \boldsymbol{x}_j^i\right) \right\}} \ \ if\ min_{i,j}\ \ d\left(\boldsymbol{x}, \boldsymbol{x}_j^i\right) < t \\ \text{Reject } if\ min_{i,j}\ \ d\left(\boldsymbol{x}, \boldsymbol{x}_j^i\right) \geq t \end{cases}$$

$\boldsymbol{x}$と最も近い鋳型が属するクラスの識別番号を返す

Returns the identifier of the class to which the template closest to $\boldsymbol{x}$

ボロノイ境界 Voronoi Boundary



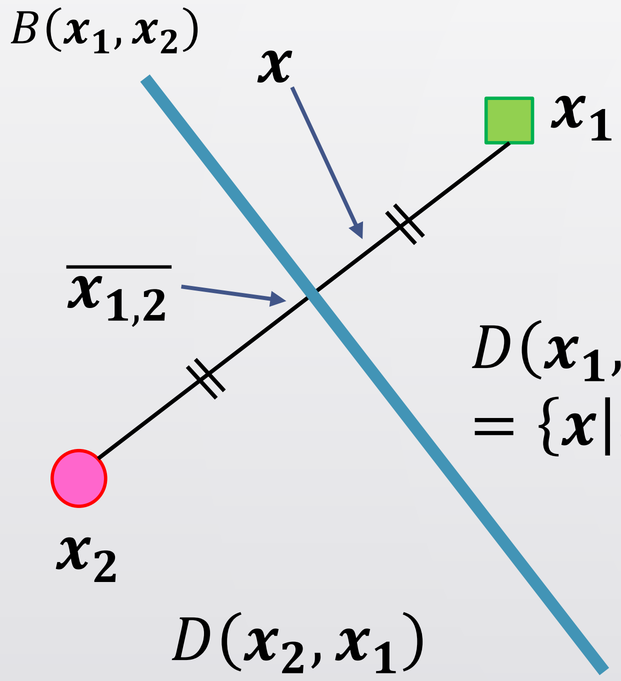
一対の鋳型から等距離にある点の集合

Set of points equidistant from a pair of templates

$$B(\boldsymbol{x}_j, \boldsymbol{x}_j) = \{\boldsymbol{x} | d(\boldsymbol{x}, \boldsymbol{x}_i) = d(\boldsymbol{x}, \boldsymbol{x}_j)\}$$

$$(\boldsymbol{x} - \overline{\boldsymbol{x}_{i,j}}) \cdot (\boldsymbol{x}_i - \boldsymbol{x}_j) = 0$$

ボロノイ領域 Voronoi Region



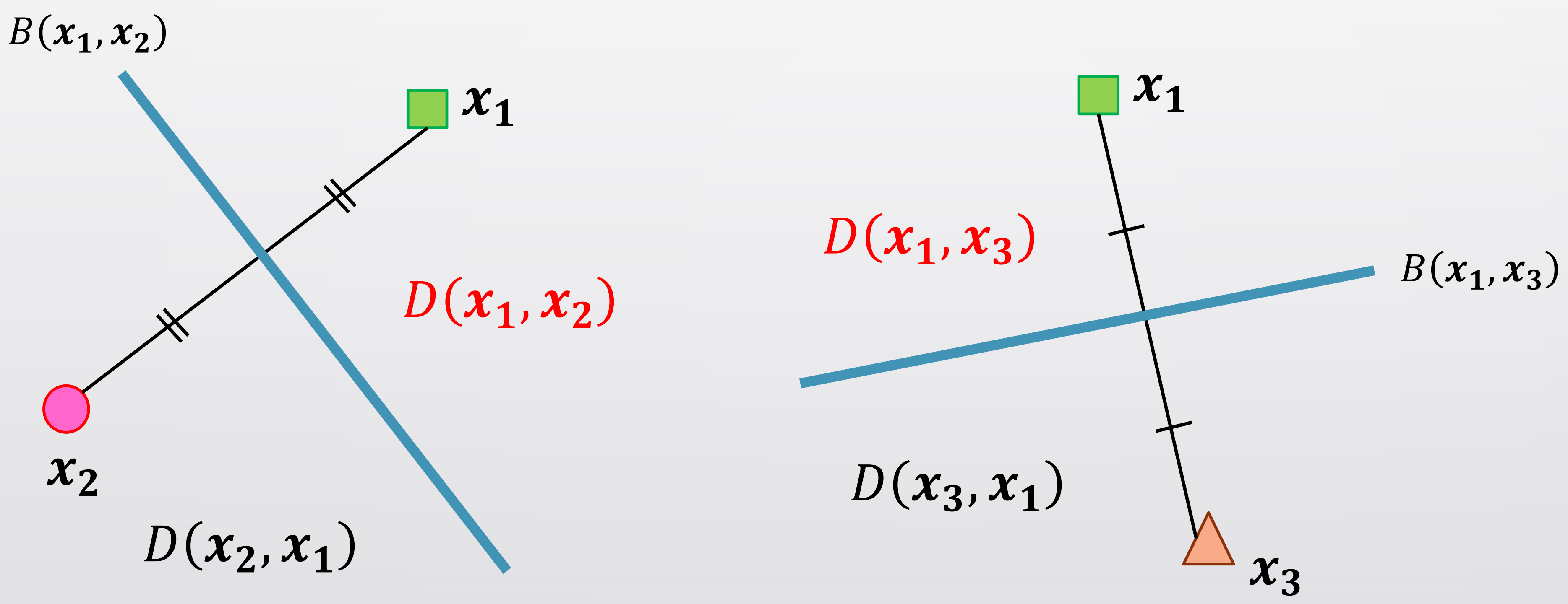
$$D(\boldsymbol{x}_1, \boldsymbol{x}_2) = \{\boldsymbol{x} | d(\boldsymbol{x}, \boldsymbol{x}_1) < d(\boldsymbol{x}, \boldsymbol{x}_2)\}$$

$D(\boldsymbol{x}_i, \boldsymbol{x}_j)$:

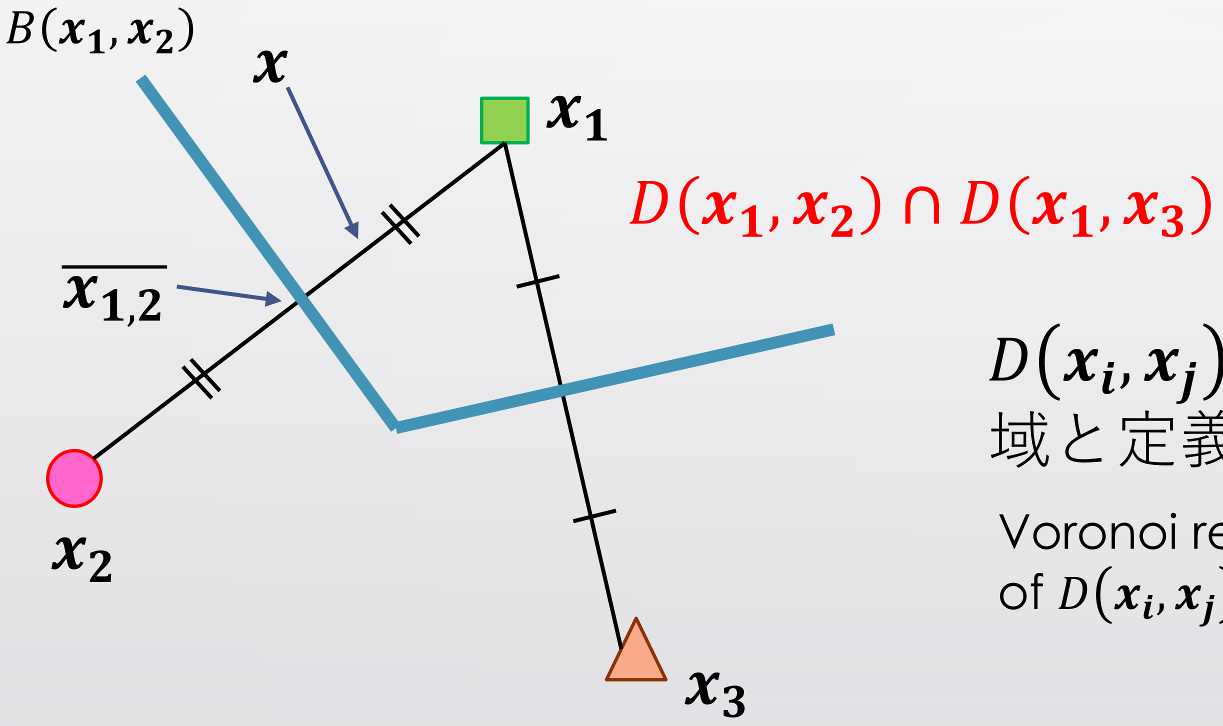
点$\boldsymbol{x}_j$より点$\boldsymbol{x}_i$に距離が近い点の集合

Set of points closer to $\boldsymbol{x}_i$ than $\boldsymbol{x}_j$

ボロノイ領域 Voronoi Region



ボロノイ領域 Voronoi Region



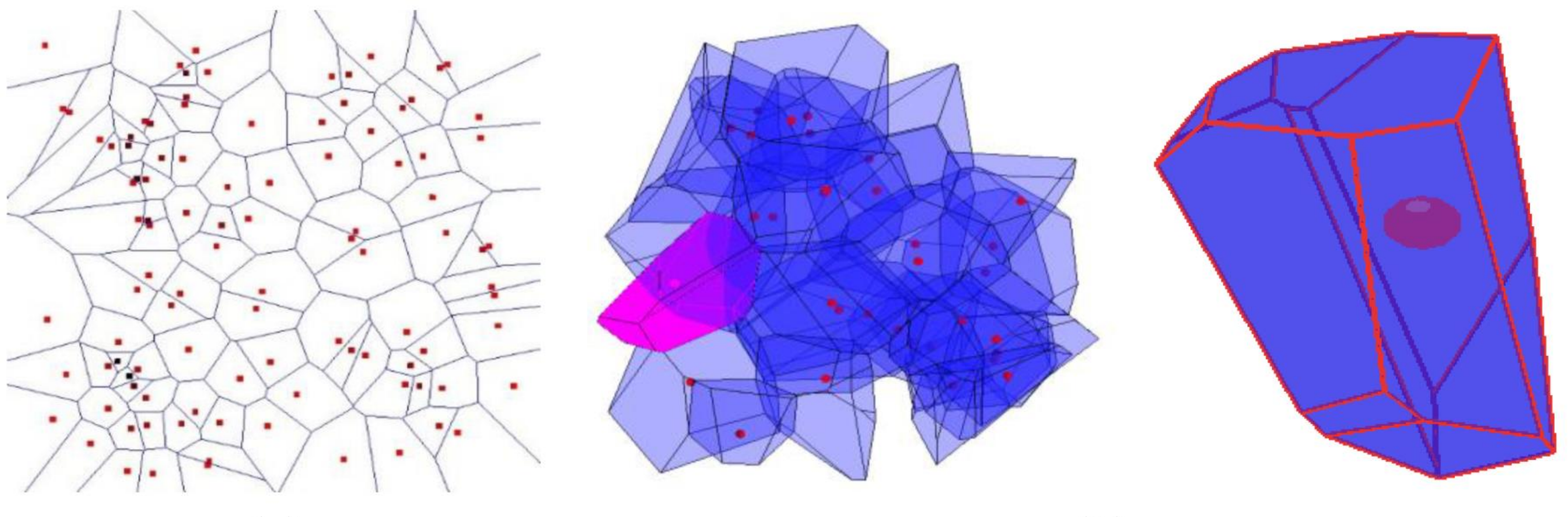
$D(\boldsymbol{x}_i, \boldsymbol{x}_j)(j \neq i)$の積集合を$\boldsymbol{x}_i$のボロノイ領域と定義する

Voronoi region of $\boldsymbol{x}_i$ is defined as set intersection of $D(\boldsymbol{x}_i, \boldsymbol{x}_j)(j \neq i)$

ボロノイ図 Voronoi Diagram

最近傍法の決定境界はボロノイ図を描く

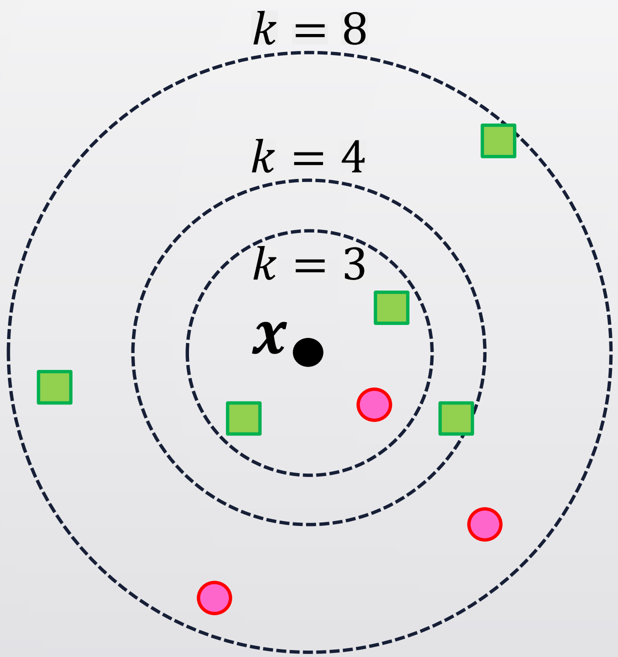
Voronoi diagram shows configuration of decision boundaries in nearest neighbor method



(a) (b)

https://www.mdpi.com/2220-9964/4/3/1480

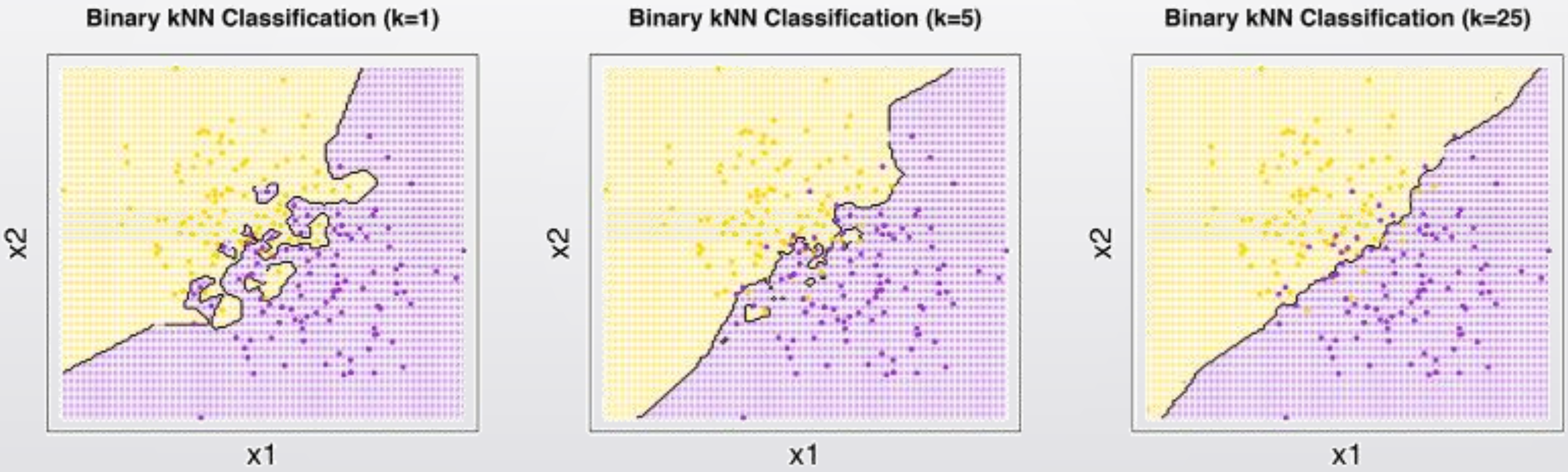
k最近傍法 k Nearest Neighbor Method



データ$\boldsymbol{x}$のクラスを最近傍にある$\boldsymbol{k}$個のデータの多数決投票により決定する

Class of data $\boldsymbol{x}$ is determined by majority voting of k data points closest to $\boldsymbol{x}$

*k*最近傍法 *k* Nearest Neighbor Method



https://elvyna.github.io/2019/knn-explained/