機械学習 Machine Learning

線形識別モデル Linear Models for Classification

福嶋 誠 Makoto Fukushima

情報科学部
School of Informatics and Data Science

分類の目的 The goal in classification

入力ベクトル $\mathbf{x} \in \mathbb{R}^D$ を K 個の離散クラス $\mathcal{C}_k$ の1つに割り当てること.
ただし, $k = 1, \ldots, K$.
Assign an input vector $\mathbf{x} \in \mathbb{R}^D$ to one of K discrete classes $\mathcal{C}_k$,
where $k = 1, \ldots, K$.

入力空間は決定領域に分離され，この領域の境界を決定境界または決定面と呼ぶ.
The input space is divided into decision regions whose boundaries are called decision boundaries or decision surfaces.

線形識別モデル Linear models for classification

決定面が入力ベクトル $\mathbf{x}$ の線形関数であり，D 次元入力空間に対して,
その決定面が $(D-1)$ 次元の超平面で定義される.
The decision surfaces are linear functions of the input vector $\mathbf{x}$ and are defined by $(D-1)$-dimensional hyperplanes within the D-dimensional input space.

分類に対する3つの異なるアプローチ
Three distinct approaches to solving classification problems

1. 識別関数（判別関数） Discriminant functions

入力ベクトル $\mathbf{x}$ から直接クラスを推定する.
Directly assign each input vector $\mathbf{x}$ to a specific class.

2. 確率的生成モデル Probabilistic generative models 生成分類器 Generative classifiers
3. 確率的識別モデル Probabilistic discriminative models 識別分類器 Discriminative classifiers

推論の段階で $p(\mathcal{C}_k|\mathbf{x})$ をモデル化し，これを利用して決定の段階で最適決定を得る.
Model $p(\mathcal{C}_k|\mathbf{x})$ in an inference stage and use it to make optimal decisions.

3.では $p(\mathcal{C}_k|\mathbf{x})$ を直接モデル化. Model $p(\mathcal{C}_k|\mathbf{x})$ directly in 3.

2.では $p(\mathbf{x}|\mathcal{C}_k)$ と $p(\mathcal{C}_k)$ をモデル化し，ベイズの定理を用いて $p(\mathcal{C}_k|\mathbf{x})$ を計算.
Model $p(\mathbf{x}|\mathcal{C}_k)$ and $p(\mathcal{C}_k)$, and compute $p(\mathcal{C}_k|\mathbf{x})$ using Bayes' theorem in 2.

識別関数（判別関数） Discriminant functions
2クラス Two classes

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$$

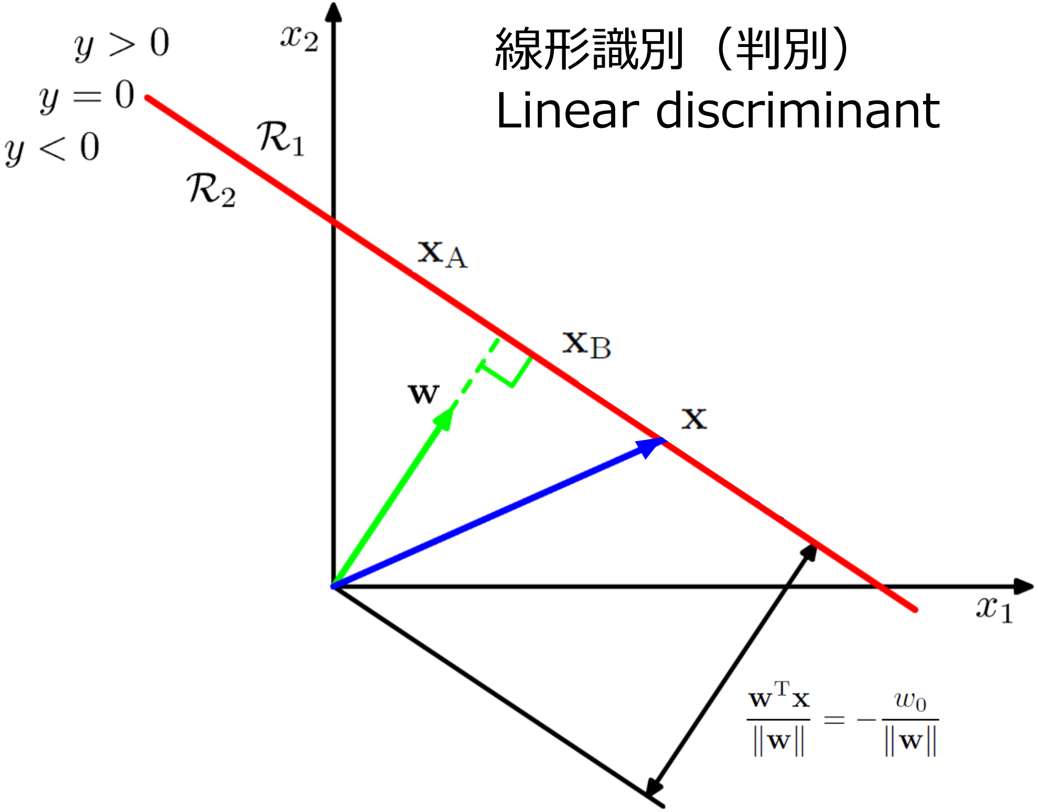
$$\mathbf{x} \begin{cases} \mathcal{C}_1 \text{ if } y(\mathbf{x}) \geqslant 0 \\ \mathcal{C}_2 \text{ otherwise} \end{cases}$$

決定境界 Decision boundary

$$y(\mathbf{x}) = 0$$

$$y(\mathbf{x}_A) = y(\mathbf{x}_B) = 0$$

$$\Rightarrow \quad \mathbf{w}^{\mathrm{T}}(\mathbf{x}_A - \mathbf{x}_B) = 0$$



線形識別（判別）
Linear discriminant

識別関数（判別関数） Discriminant functions
2クラス Two classes

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$$

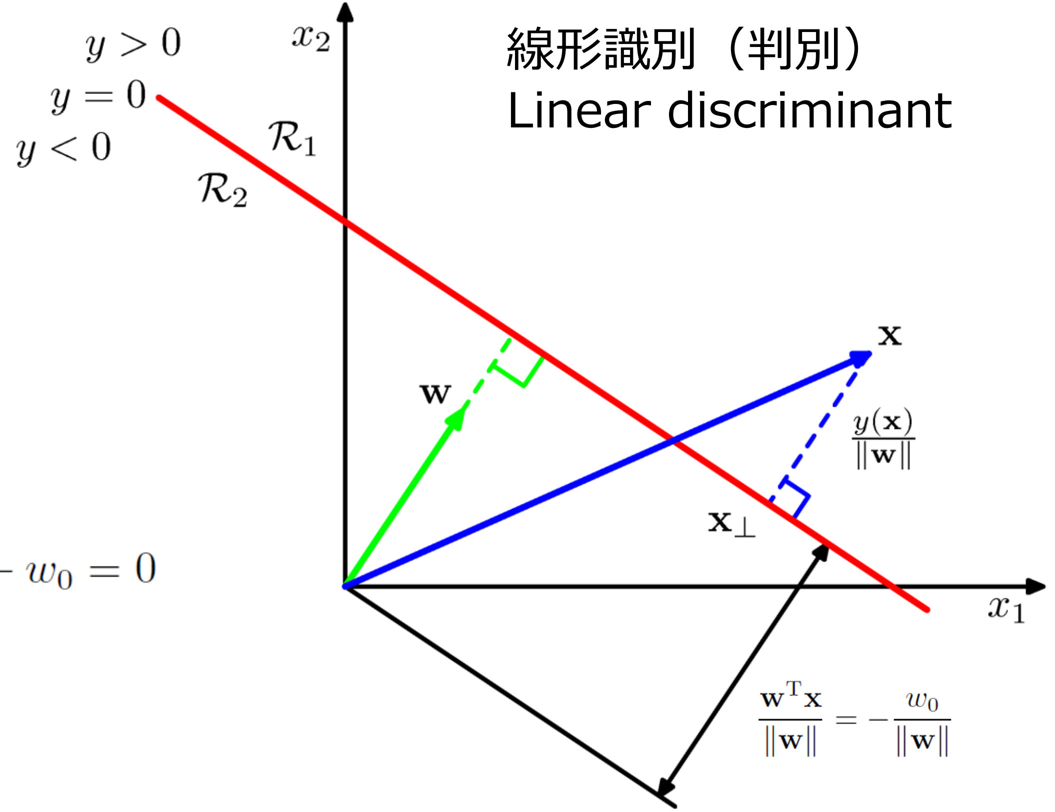
$$\mathbf{x} \begin{cases} \mathcal{C}_1 \text{ if } y(\mathbf{x}) \geqslant 0 \\ \mathcal{C}_2 \text{ otherwise} \end{cases}$$

決定境界 Decision boundary

$$y(\mathbf{x}) = 0$$

$$\mathbf{x} = \mathbf{x}_\perp + r\frac{\mathbf{w}}{\|\mathbf{w}\|} \qquad y(\mathbf{x}_\perp) = \mathbf{w}^{\mathrm{T}}\mathbf{x}_\perp + w_0 = 0$$

$$\Rightarrow \quad r = \frac{y(\mathbf{x})}{\|\mathbf{w}\|}$$



線形識別（判別）
Linear discriminant

識別関数（判別関数） Discriminant functions
2クラス Two classes

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$$

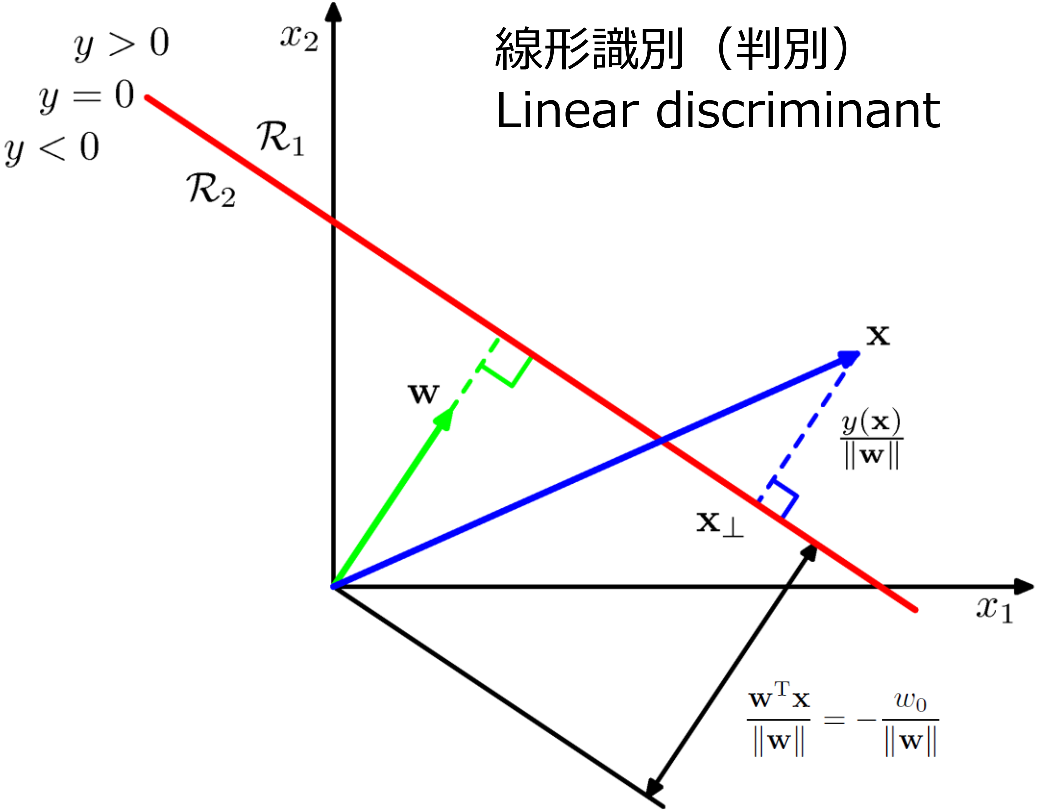
$$\mathbf{x} \quad \begin{cases} \mathcal{C}_1 \text{ if } y(\mathbf{x}) \geqslant 0 \\ \mathcal{C}_2 \text{ otherwise} \end{cases}$$

決定境界 Decision boundary

$$y(\mathbf{x}) = 0$$

$$\widetilde{\mathbf{w}} = (w_0, \mathbf{w}) \quad \widetilde{\mathbf{x}} = (x_0, \mathbf{x}) \quad x_0 = 1$$

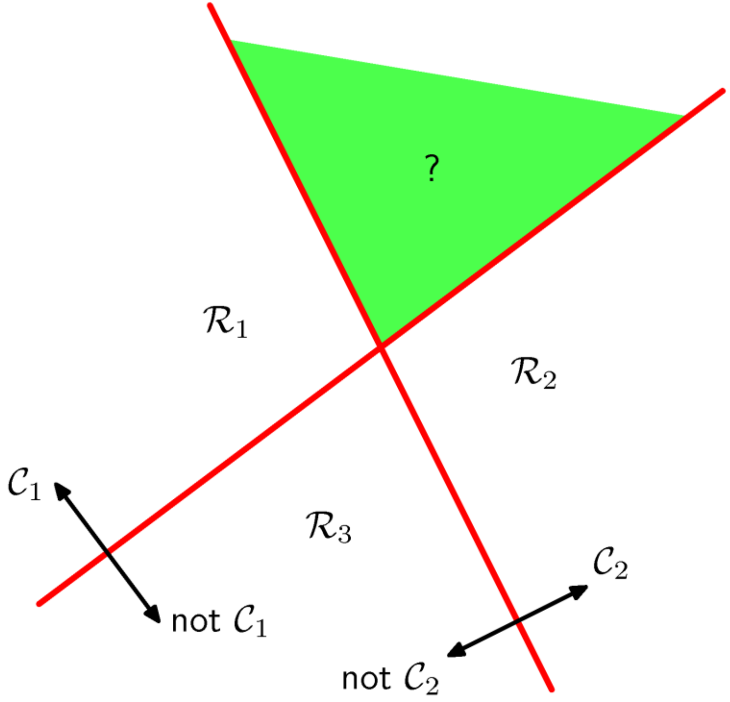
$$\Rightarrow \quad y(\mathbf{x}) = \widetilde{\mathbf{w}}^{\mathrm{T}}\widetilde{\mathbf{x}}$$



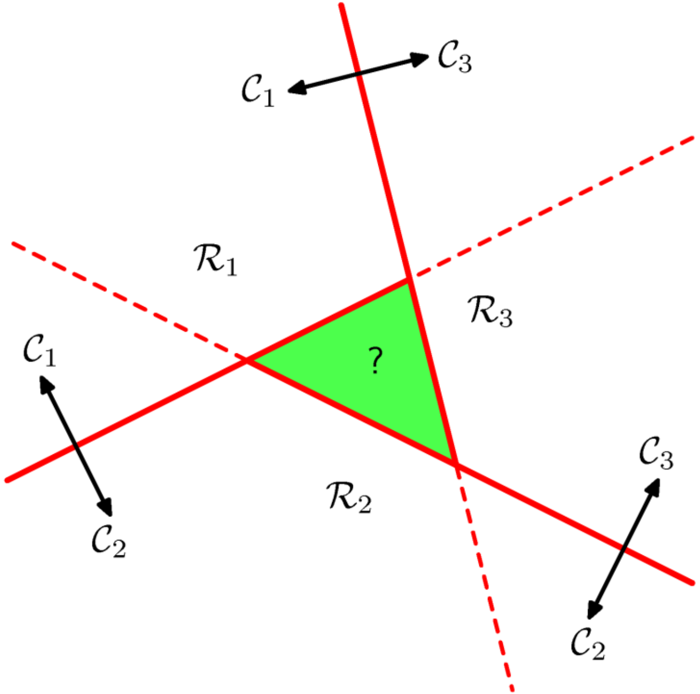
線形識別（判別）
Linear discriminant

識別関数（判別関数） Discriminant functions
多クラス Multiple classes

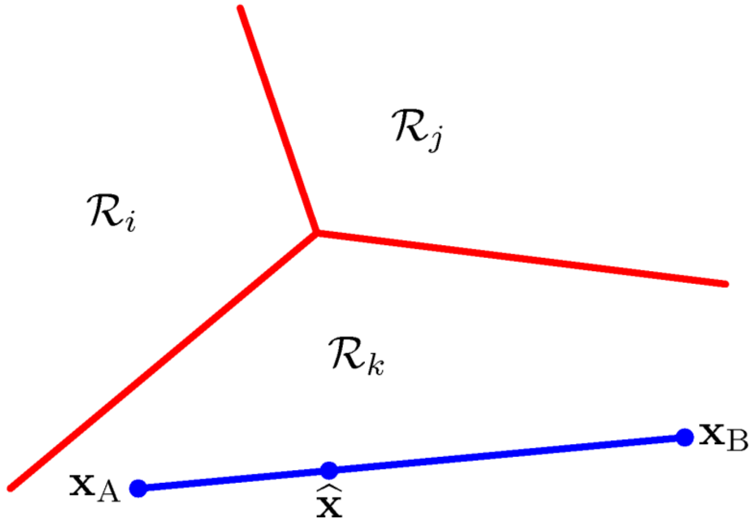
1対他分類器
One-versus-the-rest classifier



1対１分類器
One-versus-one classifier



識別関数（判別関数） Discriminant functions
多クラス Multiple classes



$$y_k(\mathbf{x}) = \mathbf{w}_k^\mathrm{T}\mathbf{x} + w_{k0}$$

$$\mathbf{x} \longrightarrow \mathcal{C}_k \text{ if } y_k(\mathbf{x}) > y_j(\mathbf{x}) \text{ for all } j \neq k$$

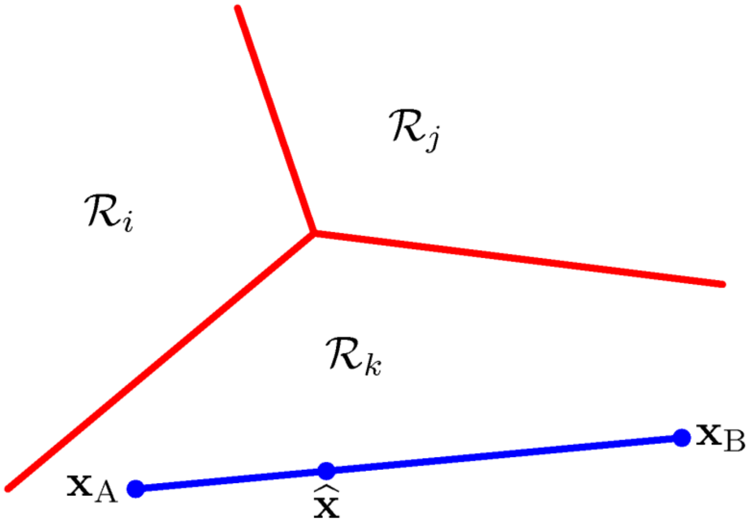
決定境界 Decision boundary

$$y_k(\mathbf{x}) = y_j(\mathbf{x})$$

$$\Rightarrow (\mathbf{w}_k - \mathbf{w}_j)^\mathrm{T}\mathbf{x} + (w_{k0} - w_{j0}) = 0$$

このような識別器の決定領域は常に単一接続していて，凸領域となっている．
The decision regions of such a discriminant are always singly connected and convex.

識別関数（判別関数） Discriminant functions
多クラス Multiple classes



$$y_k(\mathbf{x}) = \mathbf{w}_k^\mathrm{T}\mathbf{x} + w_{k0}$$

$$\mathbf{x} \longrightarrow \mathcal{C}_k \text{ if } y_k(\mathbf{x}) > y_j(\mathbf{x}) \text{ for all } j \neq k$$

$$\widehat{\mathbf{x}} = \lambda\mathbf{x}_\mathrm{A} + (1-\lambda)\mathbf{x}_\mathrm{B} \qquad 0 \leqslant \lambda \leqslant 1$$

$$y_k(\widehat{\mathbf{x}}) = \lambda y_k(\mathbf{x}_\mathrm{A}) + (1-\lambda)y_k(\mathbf{x}_\mathrm{B})$$

If $\begin{array}{l} y_k(\mathbf{x}_\mathrm{A}) > y_j(\mathbf{x}_\mathrm{A}) \\ y_k(\mathbf{x}_\mathrm{B}) > y_j(\mathbf{x}_\mathrm{B}) \end{array}$ for all $j \neq k$.

$\Rightarrow$ $y_k(\widehat{\mathbf{x}}) > y_j(\widehat{\mathbf{x}})$

識別関数（判別関数） Discriminant functions
1-of-K表記 1-of-K coding

For two-classes: $t \in \{0, 1\}$

$t = 1$ はクラス $\mathcal{C}_1$ を表す． $t = 1$ represents class $\mathcal{C}_1$.
$t = 0$ はクラス $\mathcal{C}_2$ を表す． $t = 0$ represents class $\mathcal{C}_2$.

For $K > 2$ classes: $\mathbf{t}$ は長さKのベクトル $\mathbf{t}$ is a vector of length K

クラス $\mathcal{C}_j$ であるとき，要素 t_j を除く $\mathbf{t}$ のすべての要素 t_k は0となり，要素 t_j の値は1となる．
If the class is $\mathcal{C}_j$, then all elements t_k of $\mathbf{t}$ are zero except element t_j , which takes the value 1.

例 Example: $K = 5$ の場合で，クラス2からのデータ点
A data point from class 2 when $K = 5$ ⇨ $\mathbf{t} = (0, 1, 0, 0, 0)^{\mathrm{T}}$

識別関数（判別関数） Discriminant functions
分類における最小二乗 Least squares for classification

$$y_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$$

$$\Rightarrow \quad \mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{x}} \qquad \widetilde{\mathbf{w}}_k = (w_{k0}, \mathbf{w}_k^{\mathrm{T}})^{\mathrm{T}} \qquad \widetilde{\mathbf{x}} = (1, \mathbf{x}^{\mathrm{T}})^{\mathrm{T}}$$

新たな入力 $\mathbf{x}$ は出力 $y_k = \widetilde{\mathbf{w}}_k^{\mathrm{T}}\widetilde{\mathbf{x}}$ が最大となるクラスに割り当てられる.
A new input $\mathbf{x}$ is assigned to the class for which the output $y_k = \widetilde{\mathbf{w}}_k^{\mathrm{T}}\widetilde{\mathbf{x}}$ is the largest.

二乗和誤差関数を最小化して，パラメータ行列 $\widetilde{\mathbf{W}}$ を決定する.
We determine the parameter matrix $\widetilde{\mathbf{W}}$ by minimizing a sum-of-squares error function.

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\left\{(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^{\mathrm{T}}(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})\right\}$$

$\mathbf{t}_n^{\mathrm{T}}$: nth row (行) of $\mathbf{T}$

訓練データ集合 Training data set $\{\mathbf{x}_n, \mathbf{t}_n\} \quad n = 1, \ldots, N$

$\widetilde{\mathbf{x}}_n^{\mathrm{T}}$: nth row (行) of $\widetilde{\mathbf{X}}$

識別関数（判別関数） Discriminant functions
分類における最小二乗 Least squares for classification

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\left\{(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^{\mathrm{T}}(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})\right\}$$

$\widetilde{\mathbf{W}}$ に関する導関数を0とおき，整理すると
Setting the derivative with respect to $\widetilde{\mathbf{W}}$ to zero and rearranging, we obtain

$$\widetilde{\mathbf{W}} = (\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^{\mathrm{T}}\mathbf{T} = \widetilde{\mathbf{X}}^{\dagger}\mathbf{T}$$

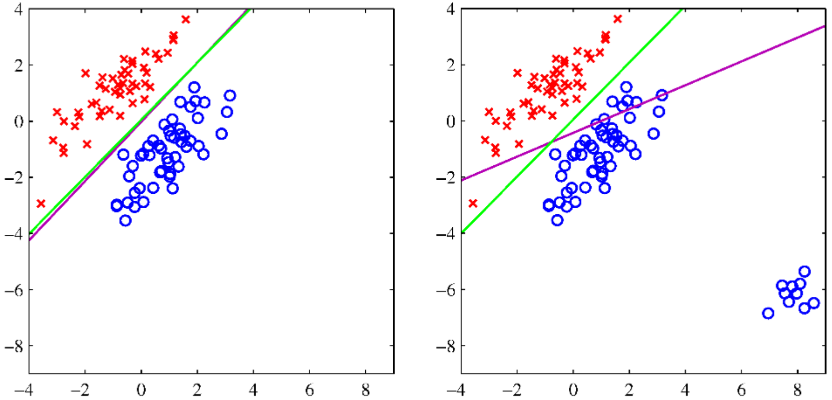
このとき以下の識別関数が得られる.
We then obtain the discriminant function in the form

$$\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{x}} = \mathbf{T}^{\mathrm{T}}\left(\widetilde{\mathbf{X}}^{\dagger}\right)^{\mathrm{T}}\widetilde{\mathbf{x}}$$

識別関数（判別関数） Discriminant functions
分類における最小二乗 Least squares for classification

問題点 Problems

外れ値に対して過敏
Sensitive to the presence of outliers



$\mathbf{y}(\mathbf{x})$ の各要素値が $(0,1)$ の範囲を超えてしまうことがある.
Elements of $\mathbf{y}(\mathbf{x})$ can have values outsider the range $(0,1)$.

決定理論 Decision theory

推論段階 Inference stage

$p(t|\mathbf{x})$ または $p(\mathbf{x}, t)$ を決める.
Determine either $p(t|\mathbf{x})$ or $p(\mathbf{x}, t)$.

決定段階 Decision stage

ある $\mathbf{x}$ に対して，最適な t を決める.
For given $\mathbf{x}$, determine optimal t.

$\mathbf{x}$ を誤ったクラスに分類する可能性を最小にしたければ，直感的には高い事後確率をもつクラスを選べばよい.
If our aim is to minimize the chance of assigning $\mathbf{x}$ to the wrong class, then intuitively we would choose the class having the higher posterior probability.

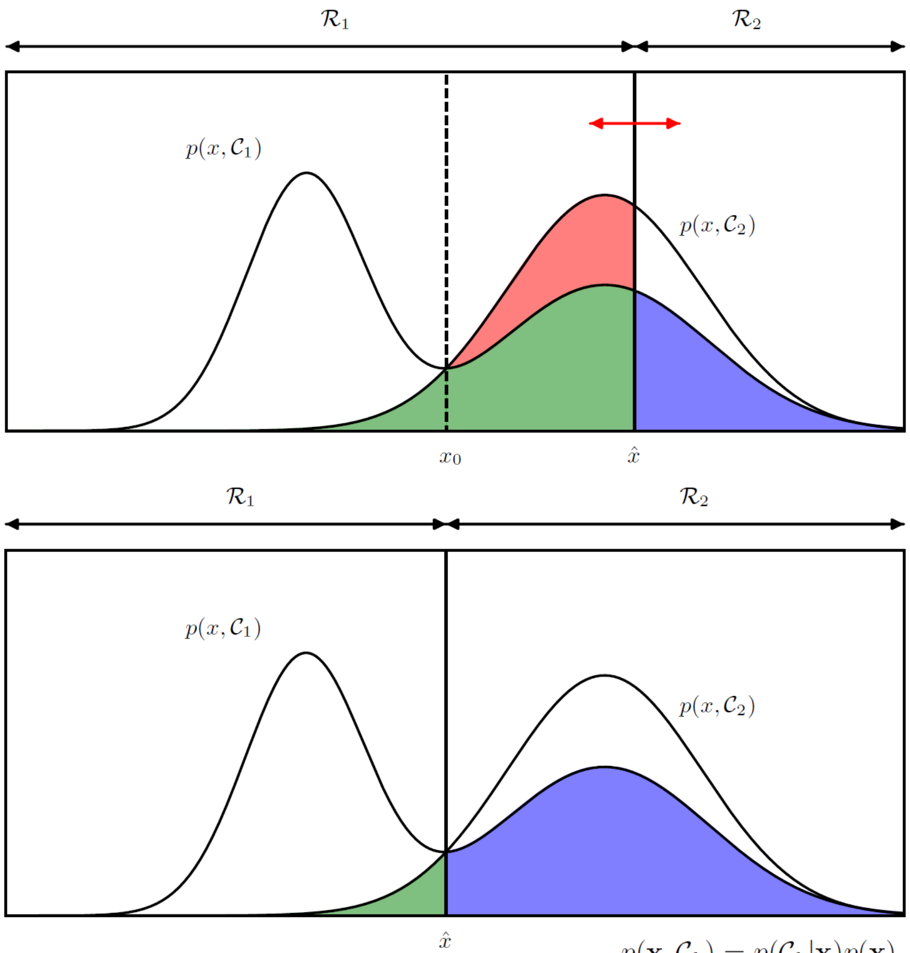
$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

決定理論 Decision theory
誤識別率 Misclassification rate

誤りの確率
The probability of taking a mistake

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}. \end{aligned}$$

ある $\mathbf{x}$ に対し, $p(\mathbf{x}, \mathcal{C}_1) > p(\mathbf{x}, \mathcal{C}_2)$ なら $\mathbf{x}$ にはクラス $\mathcal{C}_1$ を割り当てるべき.
If $p(\mathbf{x}, \mathcal{C}_1) > p(\mathbf{x}, \mathcal{C}_2)$ for given $\mathbf{x}$, then we should assign that $\mathbf{x}$ to class $\mathcal{C}_1$.

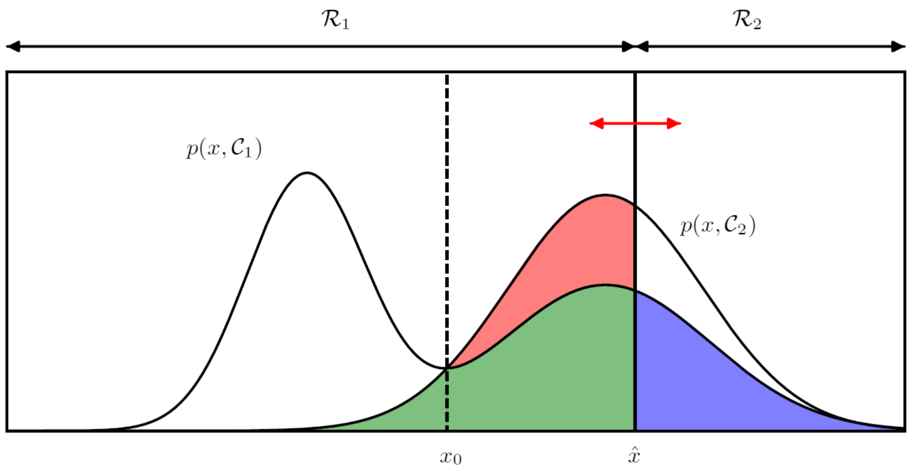


$p(\mathbf{x}, \mathcal{C}_k) = p(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})$

決定理論 Decision theory
誤識別率 Misclassification rate

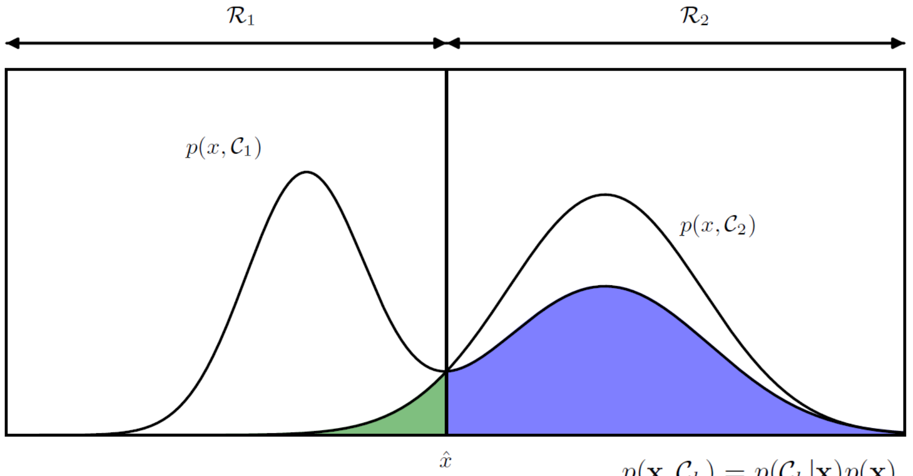
誤りの確率
The probability of taking a mistake

$$
\begin{aligned}
p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\
&= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}.
\end{aligned}
$$



正解の確率
The probability of being correct

$$
\begin{aligned}
p(\text{correct}) &= \sum_{k=1}^{K} p(\mathbf{x} \in \mathcal{R}_k, \mathcal{C}_k) \\
&= \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, \mathcal{C}_k)\,\mathrm{d}\mathbf{x}
\end{aligned}
$$



$$p(\mathbf{x}, \mathcal{C}_k) = p(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})$$

決定理論 Decision theory
期待損失 Expected loss

損失行列 Loss matrix

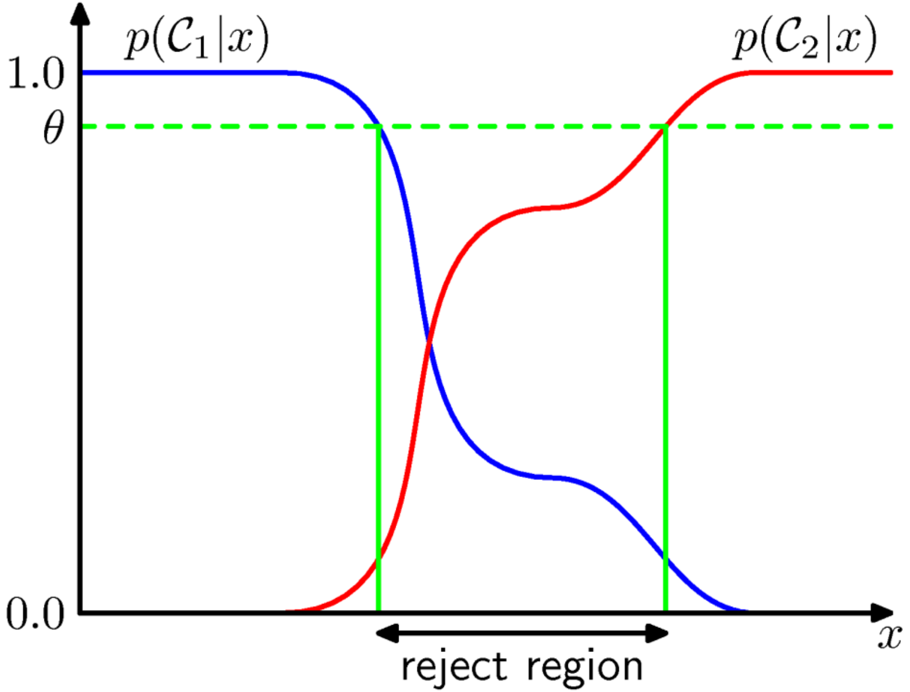
$$
\begin{array}{ccc|cc}
 & & & \text{決定 Decision} & \\
 & & & \text{正常 normal} & \text{癌 cancer} \\
\text{真実 Truth} & \text{正常} & \text{normal} & 0 & 1 \\
 & \text{癌} & \text{cancer} & 100 & 0
\end{array}
$$

以下を最小化する決定領域 $\mathcal{R}_j$ を選ぶ.
Decision regions $\mathcal{R}_j$ are chosen to minimize

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k)\, \mathrm{d}\mathbf{x}$$

⇨ $\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$ を最小化すればよい. Minimize $\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$.

決定理論 Decision theory
棄却オプション The reject option



棄却領域

決定理論 Decision theory
推論と決定 Inference and decision

3つの異なるアプローチ Three distinct approaches

(a) $p(\mathbf{x}|\mathcal{C}_k)$ を決める推論問題を解き，また $p(\mathcal{C}_k)$ を求め，ベイズの定理より $p(\mathcal{C}_k|\mathbf{x})$ を求める．決定理論を用いて新たな入力 $\mathbf{x}$ をクラスの1つに割り当てる．

Solve the inference problem of determining $p(\mathbf{x}|\mathcal{C}_k)$, infer $p(\mathcal{C}_k)$, and use Bayes' theorem to find $p(\mathcal{C}_k|\mathbf{x})$. Use decision theory to assign each new input $\mathbf{x}$ to one of the classes.

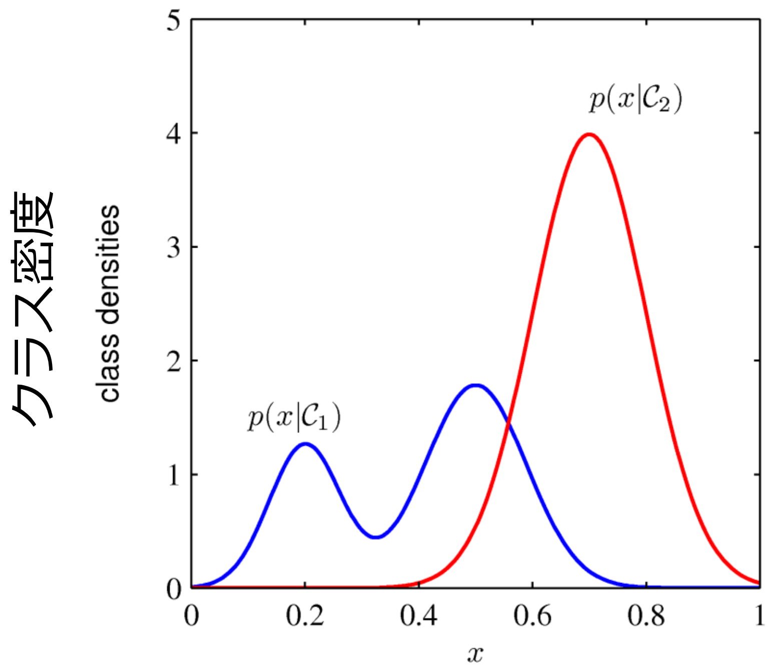
(b) $p(\mathcal{C}_k|\mathbf{x})$ を決める推論問題を解き，決定理論を用いて新たな入力 $\mathbf{x}$ をクラスの1つに割り当てる．

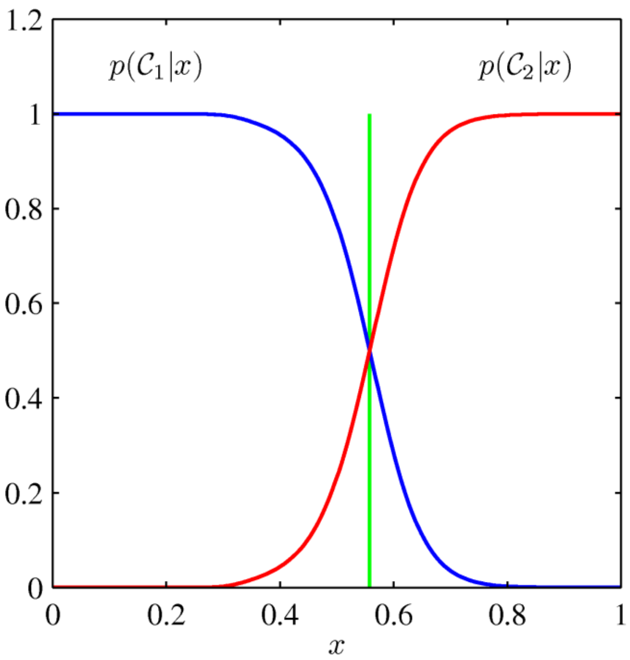
Solve the inference problem of determining $p(\mathcal{C}_k|\mathbf{x})$ and use decision theory to assign each new input $\mathbf{x}$ to one of the classes.

(c) 入力 $\mathbf{x}$ から直接クラスラベルに写像する関数 $f(\mathbf{x})$ を見つける．

Find a function $f(\mathbf{x})$ that maps a input $\mathbf{x}$ directly onto a class label.

決定理論 Decision theory
推論と決定 Inference and decision





$p(\mathbf{x}|\mathcal{C}_k)$ は $p(\mathcal{C}_k|\mathbf{x})$ にあまり影響を及ぼさない複雑な構造を取りうる.

$p(\mathbf{x}|\mathcal{C}_k)$ may contain a complex structure that has little effect on $p(\mathcal{C}_k|\mathbf{x})$.

決定理論 Decision theory
推論と決定 Inference and decision

事後確率 $p(\mathcal{C}_k|\mathbf{x})$ を計算する理由
Reasons to compute the posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$

・リスク最小化 Minimizing risk

・棄却オプション Reject option

・クラス事前確率の補正 Compensating for class priors

・モデルの結合 Combining models

決定理論 Decision theory
分類器の精度 Classifier accuracy

決定 Decision

$$\begin{array}{cc|cc} & & \text{正常 normal} & \text{癌 cancer} \\ \text{真実 Truth} & \text{正常 normal} & \multicolumn{2}{c}{\multirow{2}{*}{$\begin{pmatrix} N_{\mathrm{TN}} & N_{\mathrm{FP}} \\ N_{\mathrm{FN}} & N_{\mathrm{TP}} \end{pmatrix}$}} \\ & \text{癌 cancer} & & \end{array}$$

$$N = N_{\mathrm{TP}} + N_{\mathrm{FP}} + N_{\mathrm{TN}} + N_{\mathrm{FN}}$$

True positive: 真陽性
False positive: 偽陽性
True negative: 真陰性
False negative: 偽陰性

正解率

$$\text{Accuracy} = \frac{N_{\mathrm{TP}} + N_{\mathrm{TN}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}} + N_{\mathrm{TN}} + N_{\mathrm{FN}}}$$

適合率 $\text{Precision} = \dfrac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}}$

再現率 $\text{Recall} = \dfrac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}}$

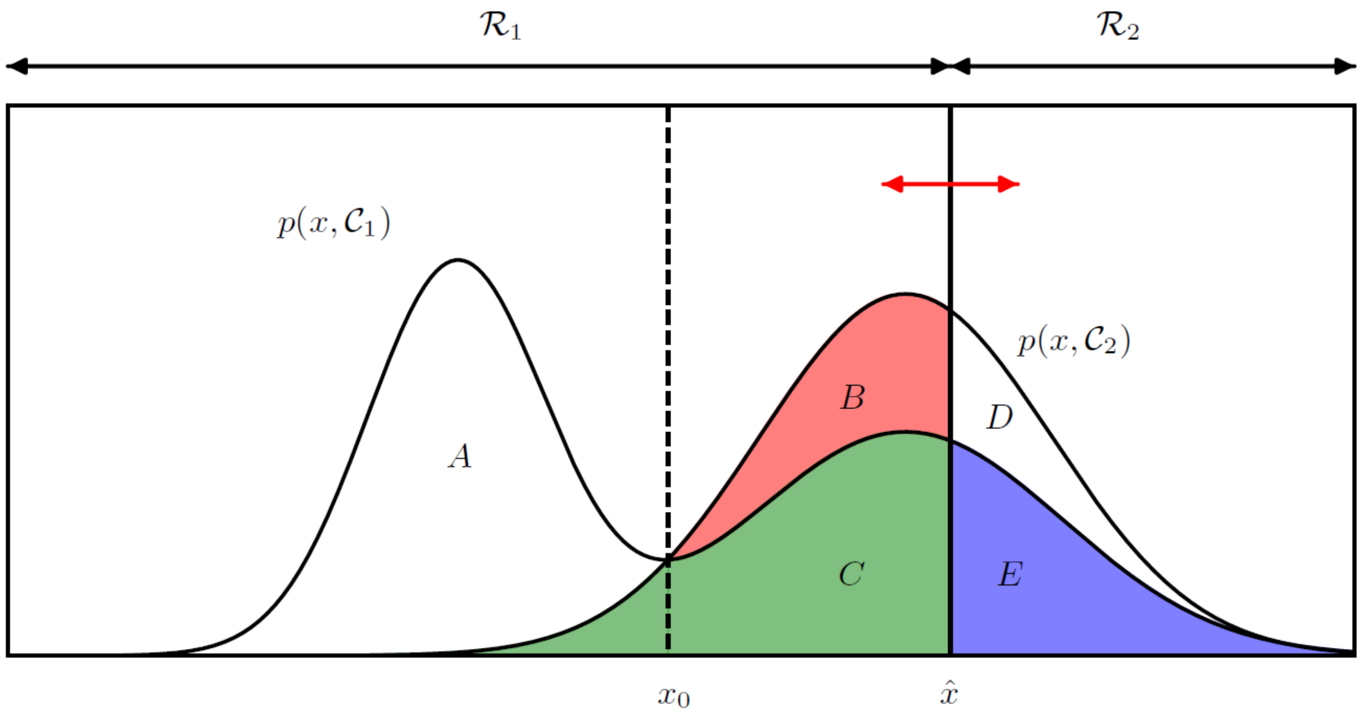
偽陽性率 $\text{False positive rate} = \dfrac{N_{\mathrm{FP}}}{N_{\mathrm{FP}} + N_{\mathrm{TN}}}$

偽発見率 $\text{False discovery rate} = \dfrac{N_{\mathrm{FP}}}{N_{\mathrm{FP}} + N_{\mathrm{TP}}}$

Fスコア
F-score

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2N_{\mathrm{TP}}}{2N_{\mathrm{TP}} + N_{\mathrm{FP}} + N_{\mathrm{FN}}}.$$

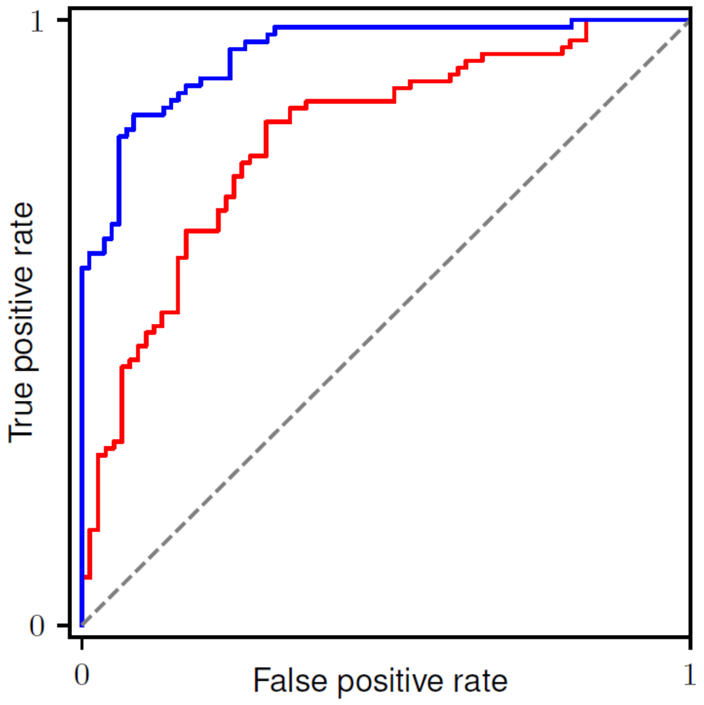
決定理論 Decision theory
分類器の精度 Classifier accuracy



$$N_{\mathrm{FP}}/N = E$$
$$N_{\mathrm{TP}}/N = D + E$$
$$N_{\mathrm{FN}}/N = B + C$$
$$N_{\mathrm{TN}}/N = A + C$$

決定理論 Decision theory
ROC曲線 ROC curve

ROC曲線 ROC curve



閾値を変化させたときの真陽性率と偽陽性率をプロット.
Plot the true positive rate and the false positive rate during changing the threshold.

よりよい識別器においては，曲線より下の面積（AUC）がより大きい.
The are under the curve (AUC) is greater for better classifier.

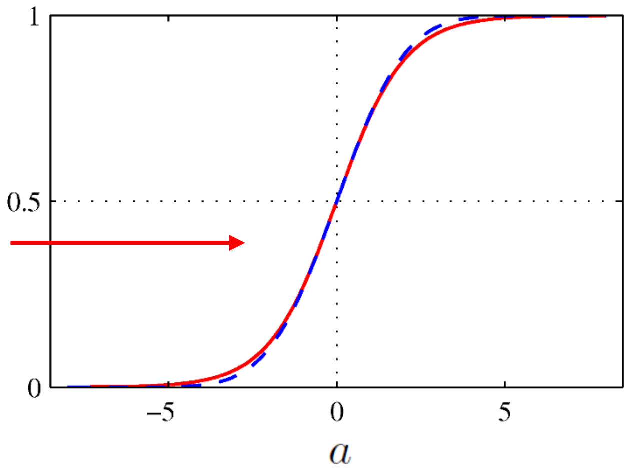
生成分類器 Generative classifiers

$p(\mathbf{x}|\mathcal{C}_k)$ と $p(\mathcal{C}_k)$ をモデル化し，$p(\mathcal{C}_k|\mathbf{x})$ をベイズの定理を用いて求める．
Model $p(\mathbf{x}|\mathcal{C}_k)$ and $p(\mathcal{C}_k)$, and compute $p(\mathcal{C}_k|\mathbf{x})$ through Bayes' theorem.

<u>2クラスの場合 If there are two class</u>

$$\begin{aligned} p(\mathcal{C}_1|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)+p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \\ &= \frac{1}{1+\exp(-a)} = \sigma(a) \end{aligned}$$

ロジスティックシグモイド関数
Logistic sigmoid function



ただし
where

$$a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

性質
Properties

$$\sigma(-a) = 1 - \sigma(a) \qquad a = \ln\left(\frac{\sigma}{1-\sigma}\right) = \ln\left[p(\mathcal{C}_1|\mathbf{x})/p(\mathcal{C}_2|\mathbf{x})\right]$$

生成分類器 Generative classifiers

$K>2$ クラスの場合 If there are $K>2$ classes

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)}$$

$$= \frac{\exp(a_k)}{\sum_j \exp(a_j)},$$

← ソフトマックス関数
Softmax function

ただし
where

$$a_k = \ln\left(p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)\right)$$

if $a_k \gg a_j$ for all $j \neq k$

$$p(\mathcal{C}_k|\mathbf{x}) \simeq 1$$
$$p(\mathcal{C}_j|\mathbf{x}) \simeq 0$$

生成分類器 Generative classifiers
連続値入力 Continuous inputs

すべてのクラスで同じ共分散行列を共有.
All classes share the same covariance matrix.

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}$$

2クラスの場合
If there are two class

$$\begin{aligned} p(\mathcal{C}_1|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)+p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \\ &= \frac{1}{1+\exp(-a)} = \sigma(a) \qquad a = \ln\frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \end{aligned}$$

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}+w_0)$$

$\mathbf{x}$ の2次の項がキャンセルされる.
The quadratic terms of $\mathbf{x}$ have cancelled.

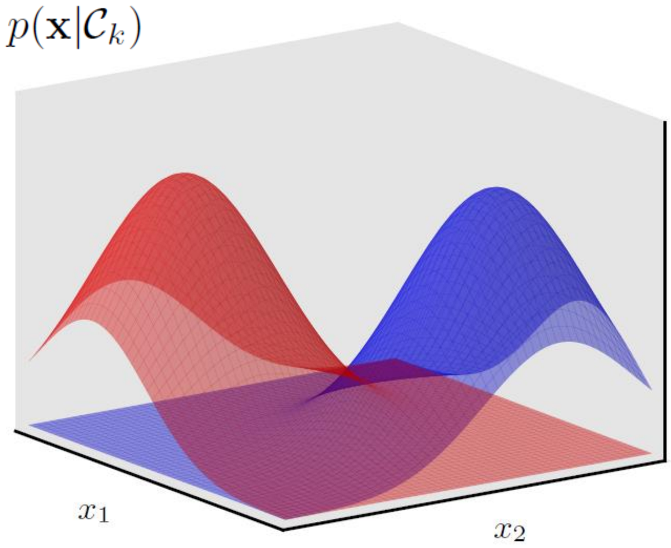
ただし
where

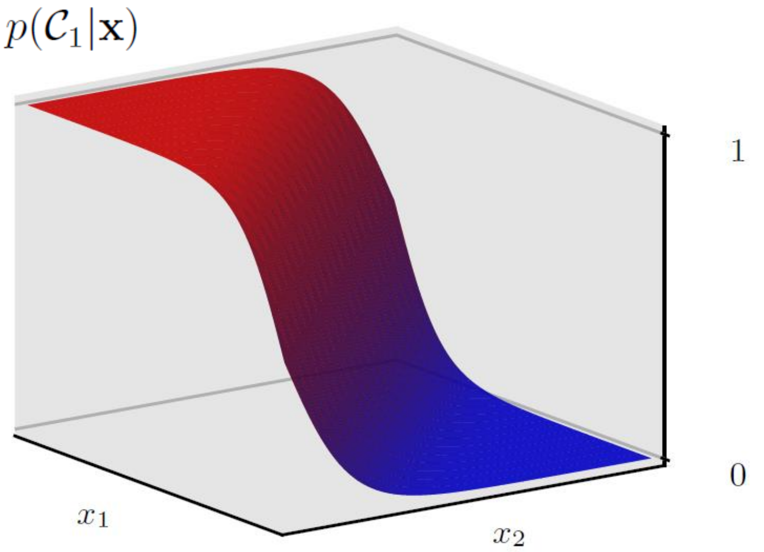
$$\begin{aligned} \mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \\ w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \end{aligned}$$

生成分類器 Generative classifiers
連続値入力 Continuous inputs

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0)$$

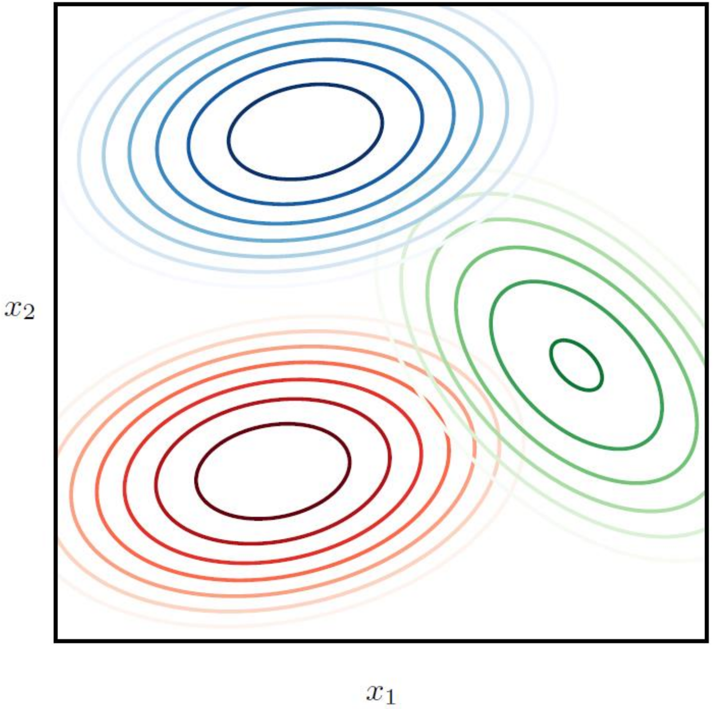
$$\begin{aligned}\mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\\ w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\end{aligned}$$

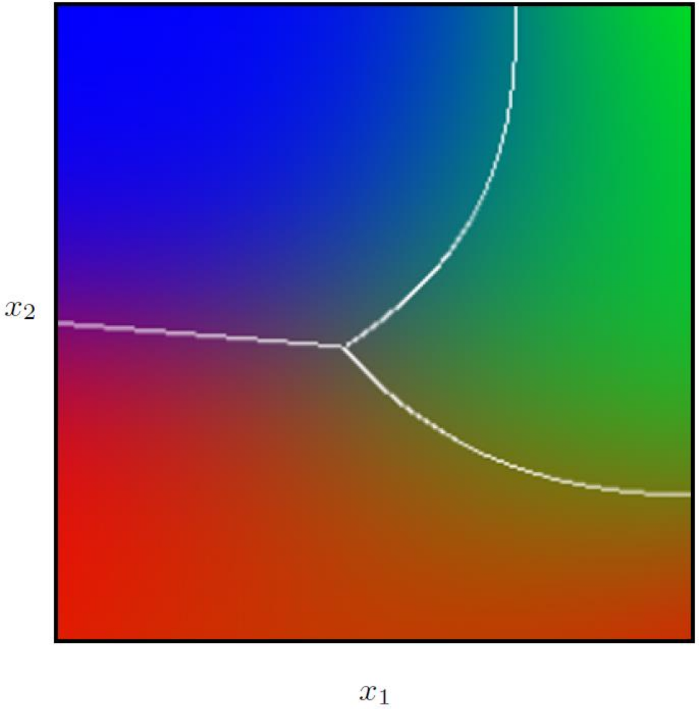




連続値入力 Continuous inputs

$K > 2$ クラスの場合 If there are $K > 2$ classes





生成分類器 Generative classifiers
最尤解 Maximum likelihood solution

2クラスの場合
If there are two class

訓練データ集合 Training data set

$$\{\mathbf{x}_n, t_n\} \quad n = 1, \ldots, N$$
$$\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$$

$t_n = 1$ denotes class $\mathcal{C}_1$ クラス $\mathcal{C}_1$を表す

$t_n = 0$ denotes class $\mathcal{C}_2$ クラス $\mathcal{C}_2$を表す

クラス事前確率 Class prior probability $p(\mathcal{C}_1) = \pi \quad p(\mathcal{C}_2) = 1 - \pi$

同時確率 Joint probability

$$p(\mathbf{x}_n, \mathcal{C}_1) = p(\mathcal{C}_1)p(\mathbf{x}_n|\mathcal{C}_1) = \pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$
$$p(\mathbf{x}_n, \mathcal{C}_2) = p(\mathcal{C}_2)p(\mathbf{x}_n|\mathcal{C}_2) = (1 - \pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

尤度関数 Likelihood function

$$p(\mathbf{t}, \mathbf{X}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} [\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n} [(1 - \pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1-t_n}$$

生成分類器 Generative classifiers
最尤解 Maximum likelihood solution

2クラスの場合
If there are two class

尤度関数
Likelihood function
$$p(\mathbf{t}, \mathbf{X}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} [\pi \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n} [(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1-t_n}$$

π に依存する対数尤度関数の項
The terms in the log likelihood that depend on π

$$\sum_{n=1}^{N} \{t_n \ln \pi + (1-t_n) \ln(1-\pi)\}$$

π に関する導関数を0として整理すると,
Setting the derivative with respect to π equal to zero and rearranging, we obtain

$$\pi = \frac{1}{N} \sum_{n=1}^{N} t_n = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$$

生成分類器 Generative classifiers
最尤解 Maximum likelihood solution

2クラスの場合
If there are two class

尤度関数
Likelihood function

$$p(\mathbf{t}, \mathbf{X}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} [\pi \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n} [(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1-t_n}$$

$\boldsymbol{\mu}_1$ に依存する対数尤度関数の項
The terms in the log likelihood that depend on $\boldsymbol{\mu}_1$

$$-\frac{1}{2}\sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) + \text{const.}$$

$\boldsymbol{\mu}_1$ に関する導関数を0として整理すると,
Setting the derivative with respect to $\boldsymbol{\mu}_1$ equal to zero and rearranging, we obtain

$$\boldsymbol{\mu}_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n \mathbf{x}_n$$

$\boldsymbol{\mu}_2$ についても同様
Similarly for $\boldsymbol{\mu}_2$

$$\boldsymbol{\mu}_2 = \frac{1}{N_2}\sum_{n=1}^{N} (1-t_n) \mathbf{x}_n$$

生成分類器 Generative classifiers
最尤解 Maximum likelihood solution

2クラスの場合
If there are two class

尤度関数
Likelihood function

$$p(\mathbf{t}, \mathbf{X}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} [\pi \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n} [(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1-t_n}$$

$\boldsymbol{\Sigma}$ に依存する対数尤度関数の項
The terms in the log likelihood that depend on $\boldsymbol{\Sigma}$

$$\begin{aligned}
&-\frac{1}{2}\sum_{n=1}^{N} t_n \ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) \\
&-\frac{1}{2}\sum_{n=1}^{N} (1-t_n) \ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} (1-t_n)(\mathbf{x}_n - \boldsymbol{\mu}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_2) \\
&= -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{N}{2}\mathrm{Tr}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{S}\right\}
\end{aligned}$$

ただし where

$$\begin{aligned}
\mathbf{S} &= \frac{N_1}{N}\mathbf{S}_1 + \frac{N_2}{N}\mathbf{S}_2 \\
\mathbf{S}_1 &= \frac{1}{N_1}\sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_1)^{\mathrm{T}} \\
\mathbf{S}_2 &= \frac{1}{N_2}\sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_2)^{\mathrm{T}}
\end{aligned}$$

$\boldsymbol{\Sigma}$ に関する導関数を0とすると,
By setting the derivative with respect to $\boldsymbol{\Sigma}$ equal to zero, ⇨ $\boldsymbol{\Sigma} = \mathbf{S}$

生成分類器 Generative classifiers
離散特徴 Discrete features

二値の離散特徴 Binary discrete feature $x_i \in \{0, 1\}$

$\mathbf{x}$ がクラス $\mathcal{C}_k$ に対して条件付き独立となるナイーブベイズを仮定したとき
When we make the naive Bayes assumption in which $\mathbf{x}$ is treated as independent and conditioned on the class $\mathcal{C}_k$.

$$p(\mathbf{x}|\mathcal{C}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1 - \mu_{ki})^{1-x_i}$$

$a_k = \ln\left(p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)\right)$ は次のように計算される.
is computed as

$$a_k(\mathbf{x}) = \sum_{i=1}^{D} \{x_i \ln \mu_{ki} + (1 - x_i)\ln(1 - \mu_{ki})\} + \ln p(\mathcal{C}_k)$$

生成分類器 Generative classifiers
指数型分布族 Exponential family

2クラスの場合
If there are two class

$p(\mathbf{x}|\mathcal{C}_k)$ が以下のような指数型分布族のサブセットの一つであるとき
When $p(\mathbf{x}|\mathcal{C}_k)$ is a member of the subset of the exponential family of distributions given by

$$p(\mathbf{x}|\boldsymbol{\lambda}_k, s) = \frac{1}{s} h\left(\frac{1}{s}\mathbf{x}\right) g(\boldsymbol{\lambda}_k) \exp\left\{\frac{1}{s}\boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{x}\right\}$$

$$a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

は次のように計算される.
is computed as

$$a(\mathbf{x}) = (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)^{\mathrm{T}}\mathbf{x} + \ln g(\boldsymbol{\lambda}_1) - \ln g(\boldsymbol{\lambda}_2) + \ln p(\mathcal{C}_1) - \ln p(\mathcal{C}_2)$$

識別分類器 Discriminative classifiers

生成分類器 Generative classifiers

$p(\mathbf{x}|\mathcal{C}_k)$ と $p(\mathcal{C}_k)$ をモデル化し，$p(\mathcal{C}_k|\mathbf{x})$ をベイズの定理を用いて求める.
Model $p(\mathbf{x}|\mathcal{C}_k)$ and $p(\mathcal{C}_k)$, and compute $p(\mathcal{C}_k|\mathbf{x})$ through Bayes' theorem.

識別分類器 Discriminative classifiers

$p(\mathcal{C}_k|\mathbf{x})$を直接モデル化する.
Directly model $p(\mathcal{C}_k|\mathbf{x})$.

利点: 学習して決めるパラメータが少ない.
Advantage: Fewer learnable parameters to be determined

識別分類器 Discriminative classifiers
活性化関数 Activation functions

	モデルによる予測 Model prediction	値域 Range
回帰 Regression	$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$	$(-\infty, \infty)$
分類 Classification	$y(\mathbf{x}, \mathbf{w}) = f\left(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0\right)$	$(0, 1)$

$f(\cdot)$: 活性化関数 Activation function

活性化関数の逆関数は，連結関数と呼ばれる.
The inverse of an activation function is called a link function

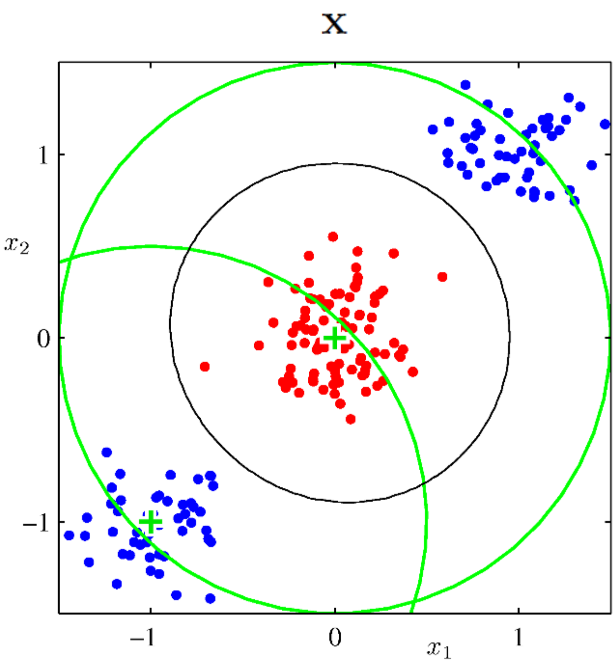
決定面は左記のときに対応
The decision surfaces correspond to

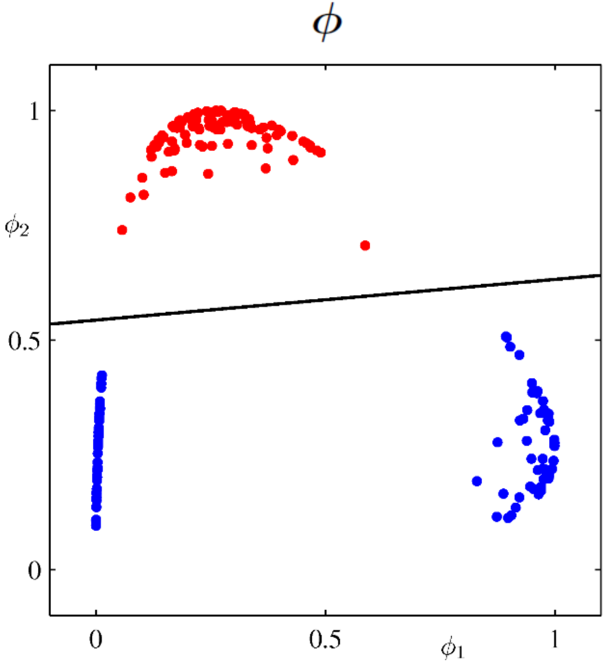
$$y(\mathbf{x}) = \text{constant}$$

$$\Rightarrow \mathbf{w}^{\mathrm{T}}\mathbf{x} = \text{constant}$$

識別分類器 Discriminative classifiers
固定基底関数 Fixed basis functions

固定基底関数を用いた入力ベクトルの非線形変換 $\phi(\mathbf{x})$
Nonlinear transformation of an input vector using fixed basis functions $\phi(\mathbf{x})$





バイアス項
The bias term
$\phi_0(\mathbf{x}) = 1$

識別分類器 Discriminative classifiers
ロジスティック回帰 Logistic regression

ロジスティック回帰モデル Logistic regression model

（2クラス分類問題で使用 Used for the problem of two-class classification）

$$p(\mathcal{C}_1|\phi) = y(\phi) = \sigma\left(\mathbf{w}^{\mathrm{T}}\phi\right) \qquad \sigma(a) = \frac{1}{1+\exp(-a)}$$

$$p(\mathcal{C}_2|\phi) = 1 - p(\mathcal{C}_1|\phi)$$

学習して決めるパラメータの数は M .
The number of learnable parameters to be determined is M .

識別分類器 Discriminative classifiers

$p(\mathbf{x}|\mathcal{C}_k)$ と $p(\mathcal{C}_k)$ を最尤法を用いて学習する場合は,
If we learn $p(\mathbf{x}|\mathcal{C}_k)$ and $p(\mathcal{C}_k)$ using maximum likelihood,

生成分類器 Generative classifiers

$$2M + M(M+1)/2 + 1 = M(M+5)/2+1$$

識別分類器 Discriminative classifiers
ロジスティック回帰 Logistic regression

ロジスティック回帰モデル Logistic regression model

（2クラス分類問題で使用 Used for the problem of two-class classification）

$$p(\mathcal{C}_1|\phi) = y(\phi) = \sigma\left(\mathbf{w}^{\mathrm{T}}\phi\right) \qquad \sigma(a) = \frac{1}{1+\exp(-a)}$$

$$p(\mathcal{C}_2|\phi) = 1 - p(\mathcal{C}_1|\phi)$$

$$t_n \in \{0,1\}$$

$$\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$$

訓練データ集合 Training data set $\{\phi_n, t_n\} \quad n = 1, \ldots, N \qquad \phi_n = \phi(\mathbf{x}_n)$

尤度関数 Likelihood function

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}\left\{1 - y_n\right\}^{1-t_n} \qquad y_n = p(\mathcal{C}_1|\phi_n)$$

交差エントロピー誤差関数
Cross-entropy error function

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}\left\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\right\}$$

識別分類器 Discriminative classifiers
ロジスティック回帰 Logistic regression

誤差関数の $\mathbf{w}$ についての勾配を求める.
Derive the gradient of the error function with respect to $\mathbf{w}$.

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$$

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \frac{\partial E}{\partial y_n}\frac{\partial y_n}{\partial a_n}\nabla a_n = \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\phi}_n$$

$$\frac{\mathrm{d}\sigma}{\mathrm{d}a} = \sigma(1-\sigma)$$

$$y_n = \sigma(a_n)$$

$$a_n = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n$$

$$\begin{aligned}\frac{\partial E}{\partial y_n} &= \frac{1-t_n}{1-y_n} - \frac{t_n}{y_n} \\ &= \frac{y_n(1-t_n) - t_n(1-y_n)}{y_n(1-y_n)} \\ &= \frac{y_n - y_n t_n - t_n + y_n t_n}{y_n(1-y_n)} \\ &= \frac{y_n - t_n}{y_n(1-y_n)}.\end{aligned}$$

$$\begin{aligned}\frac{\partial y_n}{\partial a_n} &= \frac{\partial \sigma(a_n)}{\partial a_n} \\ &= \sigma(a_n)\left(1-\sigma(a_n)\right) \\ &= y_n(1-y_n)\end{aligned}$$

$$\nabla a_n = \boldsymbol{\phi}_n$$

識別分類器 Discriminative classifiers
ロジスティック回帰 Logistic regression

ロジスティック回帰における誤差関数の勾配
The gradient of the error function in logistic regression

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\phi_n$$

線形回帰における対数尤度関数の勾配
The gradient of the likelihood function in linear regression

$$\nabla \ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^{N} \left\{ t_n - \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_n) \right\} \phi(\mathbf{x}_n)^{\mathrm{T}}$$

識別分類器 Discriminative classifiers
ロジスティック回帰 Logistic regression

ロジスティック回帰における誤差関数の勾配
The gradient of the error function in logistic regression

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\phi_n$$

最尤解は $\nabla E(\mathbf{w}) = 0$ を計算することにより得られる.
The maximum likelihood solution is obtained by computing $\nabla E(\mathbf{w}) = 0$.

$y(\cdot)$ の非線形性のため，最尤解は閉形式の形では求まらない.
The maximum likelihood solution is not a close-form solution due to the nonlinearity in $y(\cdot)$.

⇨ 一般には，確率的勾配降下法を用いて解く.
In general, solve it using stochastic gradient descent.

識別分類器 Discriminative classifiers
多クラスロジスティック回帰 Multi-class logistic regression

多クラスロジスティック回帰モデル Multi-class logistic regression model

（多クラス分類問題で使用 Used for the problem of multi-class classification）

$$p(\mathcal{C}_k|\boldsymbol{\phi}) = y_k(\boldsymbol{\phi}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \qquad a_k = \mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi}$$

尤度関数 Likelihood function

$$p(\mathbf{T}|\mathbf{w}_1,\ldots,\mathbf{w}_K) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(\mathcal{C}_k|\boldsymbol{\phi}_n)^{t_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

$y_{nk} = y_k(\boldsymbol{\phi}_n)$

$\mathbf{T}$: A matrix with elements t_{nk}

t_{nk} を要素とする行列

多クラス交差エントロピー誤差関数 Multi-class cross-entropy error function

$$E(\mathbf{w}_1,\ldots,\mathbf{w}_K) = -\ln p(\mathbf{T}|\mathbf{w}_1,\ldots,\mathbf{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ln y_{nk}$$

識別分類器 Discriminative classifiers
多クラスロジスティック回帰 Multi-class logistic regression

誤差関数の $\mathbf{w}_j$ についての勾配を求める.
Derive the gradient of the error function with respect to $\mathbf{w}_j$.

$$E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\ln p(\mathbf{T}|\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}$$

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \sum_{n=1}^{N} (y_{nj} - t_{nj})\, \boldsymbol{\phi}_n$$

$$\begin{aligned}
\frac{\partial E}{\partial a_{nj}} &= \sum_{k=1}^{K} \frac{\partial E}{\partial y_{nk}} \frac{\partial y_{nk}}{\partial a_{nj}} \\
&= -\sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} y_{nk} (I_{kj} - y_{nj}) \\
&= y_{nj} - t_{nj},
\end{aligned}$$

$$\nabla a_n = \boldsymbol{\phi}_n$$

$$\frac{\partial E}{\partial y_{nk}} = -\frac{t_{nk}}{y_{nk}}$$

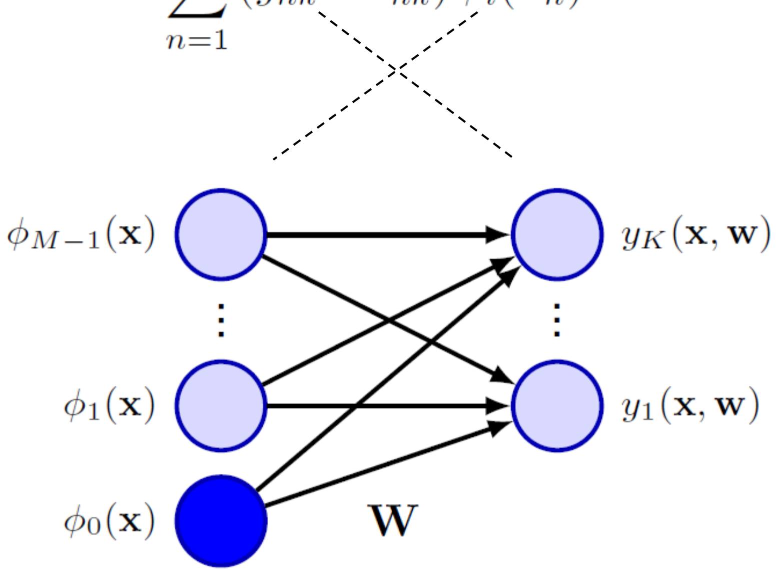
$$\frac{\partial y_k}{\partial a_j} = y_k (I_{kj} - y_j)$$

I_{kj} : An element in the identity matrix
単位行列の要素

$\sum_k t_{nk} = 1$

識別分類器 Discriminative classifiers
多クラスロジスティック回帰 Multi-class logistic regression

$$\frac{\partial E(\mathbf{w}_1, \ldots, \mathbf{w}_K)}{\partial w_{ik}} = \sum_{n=1}^{N} (y_{nk} - t_{nk})\, \phi_i(\mathbf{x}_n)$$



識別分類器 Discriminative classifiers
正準連結関数 Canonical link functions

$$y = f(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})$$

$f^{-1}(\cdot)$ として正準連結関数を選び，指数型分布族の中から目的変数に対する条件付き分布を選ぶと，誤算関数が $y_n - t_n$ と $\boldsymbol{\phi}_n$ の積の形になる．
If we choose $f^{-1}(\cdot)$ as the canonical link function and assume a conditional distribution for the target variable from the exponential family distribution, the error function takes the form of $y_n - t_n$ times $\boldsymbol{\phi}_n$.

指数型分布族の仮定 The assumption of exponential family distribution

$$p(t|\eta, s) = \frac{1}{s} h\left(\frac{t}{s}\right) g(\eta) \exp\left\{\frac{\eta t}{s}\right\}$$

$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$ を導出したときと同じ議論によって，
Using the same line of argument as led to the derivation of $-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$

$$y \equiv \mathbb{E}[t|\eta] = -s \frac{d}{d\eta} \ln g(\eta)$$

$\eta = \psi(y)$ とおく．
We denote $\eta = \psi(y)$.

識別分類器 Discriminative classifiers
正準連結関数 Canonical link functions

対数尤度関数 Log likelihood function

$$\ln p(\mathbf{t}|\eta, s) = \sum_{n=1}^{N} \ln p(t_n|\eta, s) = \sum_{n=1}^{N} \left\{ \ln g(\eta_n) + \frac{\eta_n t_n}{s} \right\} + \text{const}$$

対数尤度関数の勾配 The gradient of the log likelihood function

$$\begin{aligned} \nabla_{\mathbf{w}} \ln p(\mathbf{t}|\eta, s) &= \sum_{n=1}^{N} \left\{ \frac{d}{d\eta_n} \ln g(\eta_n) + \frac{t_n}{s} \right\} \frac{d\eta_n}{dy_n} \frac{dy_n}{da_n} \nabla_{\mathbf{w}} a_n \\ &= \sum_{n=1}^{N} \frac{1}{s} \{t_n - y_n\} \psi'(y_n) f'(a_n) \boldsymbol{\phi}_n \end{aligned}$$

$$y \equiv \mathbb{E}[t|\eta] = -s \frac{d}{d\eta} \ln g(\eta)$$

$$\eta = \psi(y)$$

$$y_n = f(a_n)$$

$$a_n = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}_n$$

識別分類器 Discriminative classifiers
正準連結関数 Canonical link functions

$f^{-1}(y) = \psi(y)$ （正準連結関数）とした場合,

If we choose $f^{-1}(y) = \psi(y)$ (canonical link function),

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\eta, s) = \sum_{n=1}^{N} \frac{1}{s}\{t_n - y_n\} \underbrace{\psi'(y_n) f'(a_n)}_{=1} \phi_n$$

⇨ $$\nabla \ln E(\mathbf{w}) = \frac{1}{s} \sum_{n=1}^{N} \{y_n - t_n\} \phi_n$$

$$f(\psi(y)) = y \qquad f'(\psi)\psi'(y) = 1$$

$$a = f^{-1}(y) \qquad a = \psi \qquad f'(a)\psi'(y) = 1$$

課題

提出課題 III：線形識別モデル
Assignment III: Linear Models for Classification

提出期限：**11月5日（火曜日）23:59:00 [日本標準時]**
Submission deadline: **November 5 (Tuesday) 23:59:00 [Japan Standard Time]**

提出課題は「一般」チャネルの「ファイル」にアップロードされます.
同チャネルに出現する通知のリンク先から解答を送信（提出）してください.
Assignments will be uploaded to "File" in the "General" channel. Send (submit) your answers via the link that will appear in the same channel.

- 全6回の課題への解答をもとに成績評価を行います.
  Your grades will be based on your answers to all six assignments.
- 解答送信後の解答再送信はできません.
  Once submitted, answers cannot be resubmitted.
- 提出期限は一切延長しません.
  The submission deadline will never be extended.