機械学習(5)
決定的識別モデル

情報科学類　佐久間　淳

見通し

	教師あり	教師なし
予測対象が連続	回帰、推薦	PCA
予測対象が離散	分類(識別)	クラスタリング

Reuterデータでニュースカテゴリ分類

- ニュース配信会社Reutersのニュースデータ
- データ:テキスト
- 特徴ベクトル:bag-of-words(21,989次元)
- 予測:カテゴリ(90種類)
- 学習データ:9,603文章
- テストデータ:3,299文章

https://unsplash.com/photos/Mwuod2cm8g4
Mr Cup / Fabien Barral

データ例（カテゴリ：acq）

- 資金調達(acquire)に関するカテゴリ
  - McDowell Enterprises Inc saidit has signed a definitive agreement to acquire an 80 pct interest …
  - Regie Nationale des Usines Renault<RENA.PA> said it and Chrysler Corp <C> have signed a letter of intent…
  - Dudley Taft and Narragansett Capital Inc said it was prepared to raise its bid to acquire…

データ例（カテゴリ：crude）

- 原油に関するカテゴリ
  - State-run oil company Ecopetrol said Colombia's main oil pipeline was bombed again…
  - Iraq said its forces sank three Iranian boats which tried to approach its disused deep water oil terminal…
  - Brazil will import 30,000 barrels per day of crude oil from Kuwait…

データ例（カテゴリ：grain）

- 穀物に関するカテゴリ
  - The U.S. corn and soybean crops are in mostly good condition and have not suffered any yield deterioration from recent hot …
  - Egypt for the third time submitted a bid of 89 dlrs per tonne in its tender for 200,000 tonnes of soft red or white wheat …
  - Japan has no plans to liberalise its farm markets, but will try to narrow the gap between …

復習

文書を特徴ベクトルに

文書における特徴ベクトルの要素は文書中の単語です

- 文書 名詞,一般,*,*,*,*,文書,ブンショ,ブンショ
- における 助詞,格助詞,連語,*,*,*,における,ニオケル
- 特徴 名詞,一般,*,*,*,*,特徴,トクチョウ,トクチョー
- ベクトル 名詞,固有名詞,一般,*,*,*,*,
- の 助詞,連体化,*,*,*,*,の,ノ,ノ
- 要素 名詞,一般,*,*,*,*,要素,ヨウソ,ヨーソ
- は 助詞,係助詞,*,*,*,*,は,ハ,ワ
- 文書 名詞,一般,*,*,*,*,文書,ブンショ,ブンショ
- 中 名詞,接尾,副詞可能,*,*,*,中,チュウ,チュー
- の 助詞,連体化,*,*,*,*,の,ノ,ノ
- 単語 名詞,一般,*,*,*,*,単語,タンゴ,タンゴ
- です 助動詞,*,*,*,特殊・デス,基本形,です,デス,デス

形態素解析

頻度表

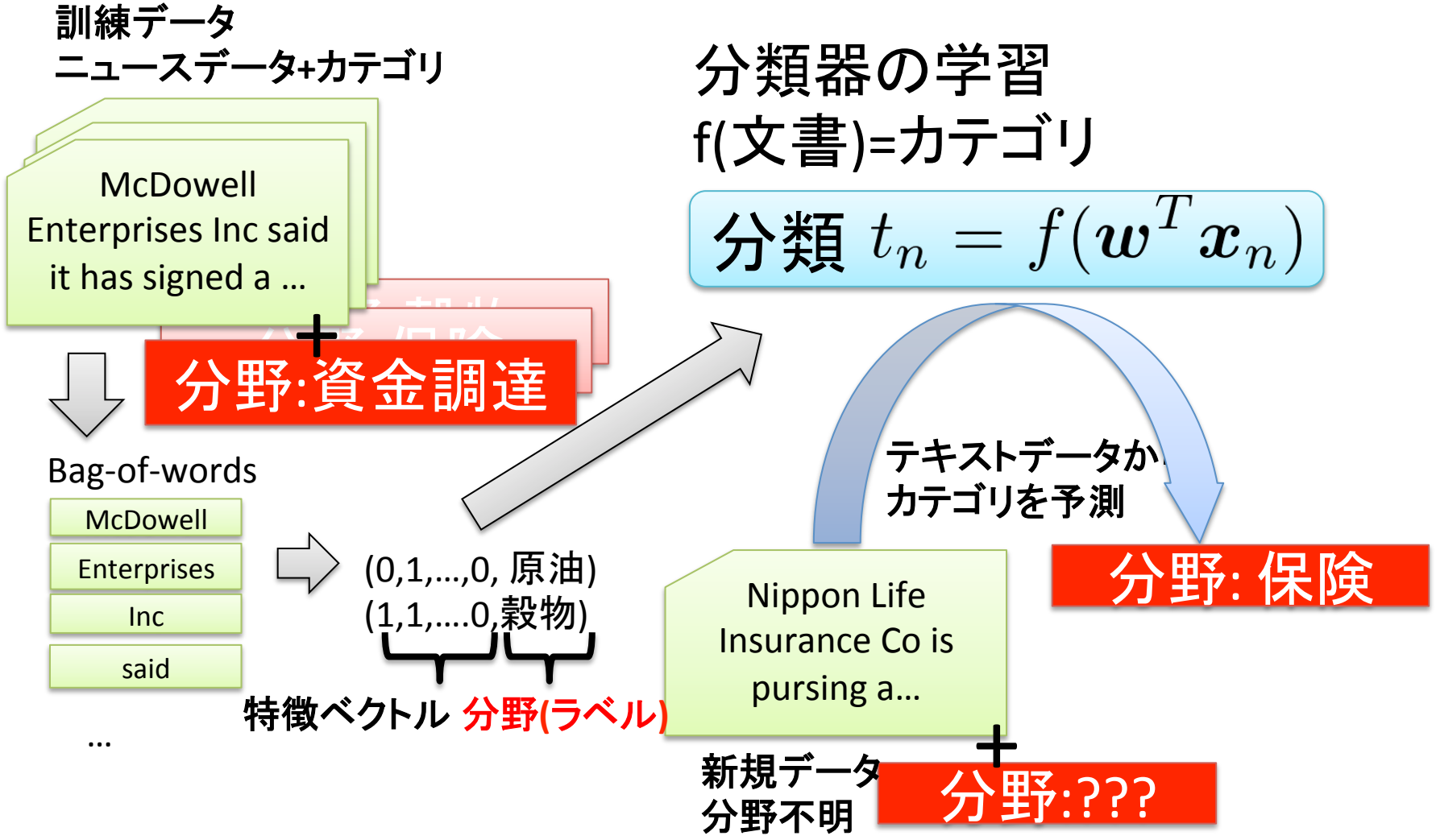
単語	出現頻度
文書	2
における	1
特徴	1
ベクトル	1
の	2
要素	1
は	1
中	1
単語	1
です	1

Stop word除去

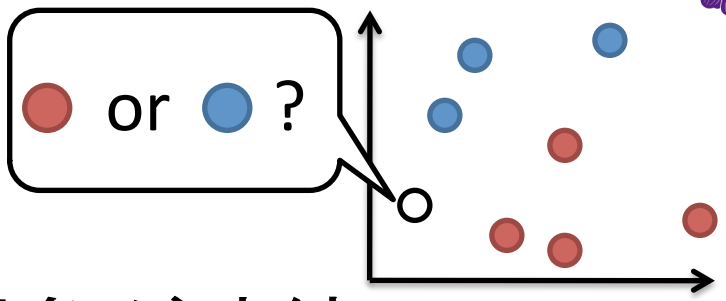
単語	頻度
文書	2
特徴	1
ベクトル	1
要素	1
単語	1

http://www.nal.ie.u-ryukyu.ac.jp/wiki/index.php?yono%2FPython%A4%C7%BC%AB%C1%B3%B8%C0%B8%EC%BD%E8%CD%FD

ニュースカテゴリ分類(シナリオ)



分類



- 回帰　$f : \mathbb{R}^D \rightarrow \mathbb{R}$ (予測対象が連続)
- 分類　$f : \mathbb{R}^D \rightarrow \{\mathcal{C}_1, \mathcal{C}_2\}$　(予測対象が離散)
  - (気温,湿度)→{晴,曇,雨}
  - 文書ベクトル→{政治,経済,スポーツ,…}
- 教師あり学習
  - 事例 (x_i, t_i)　$x_i \in \mathbb{R}^D$　$t_i \in \{\mathcal{C}_1, \mathcal{C}_2\}$
  - 目標: 未知事例 $\boldsymbol{x} \in \mathbb{R}^D$　のラベル $\in \{\mathcal{C}_1, \mathcal{C}_2\}$ を当てる
  - 慣例として、$t_i \in \{-1, 1\}$ (決定的識別モデル)
    $t_i \in \{0, 1\}$　(確率的識別モデル)

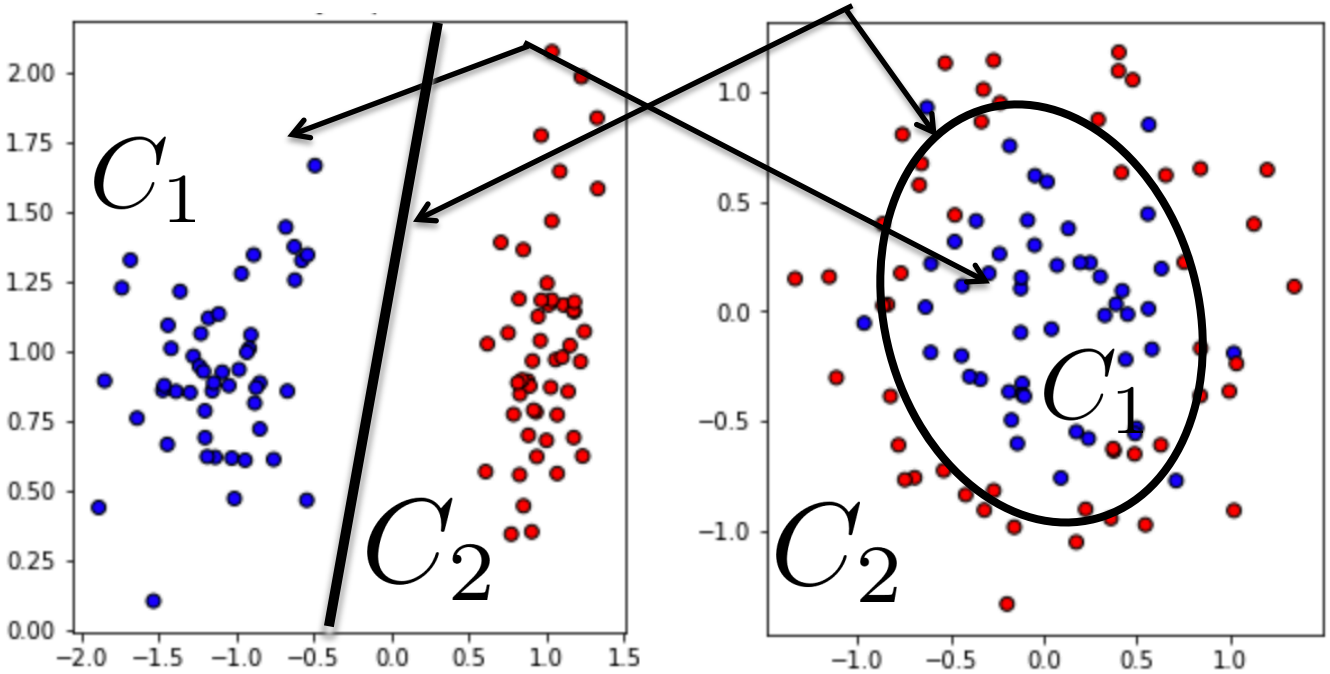
分類の例

- 2クラス分類　はじめは簡単のため2クラスだけ扱う
  - 購買履歴から、ある人がある商品を買うかどうか予測
  - 取引履歴から、ある会社が将来破たんするかどうか予測
  - 検査データから、ある人が病気xにかかるかどうか予測
- 多クラス分類
  - ある文書がどのカテゴリに属するか予測(y∈{政治、経済、スポーツ、芸能…})
  - 購買履歴から、ある人がどのジャンルの商品を買うか予測(y∈{食料品、本、雑貨、…})

用語

決定領域
(decision region)

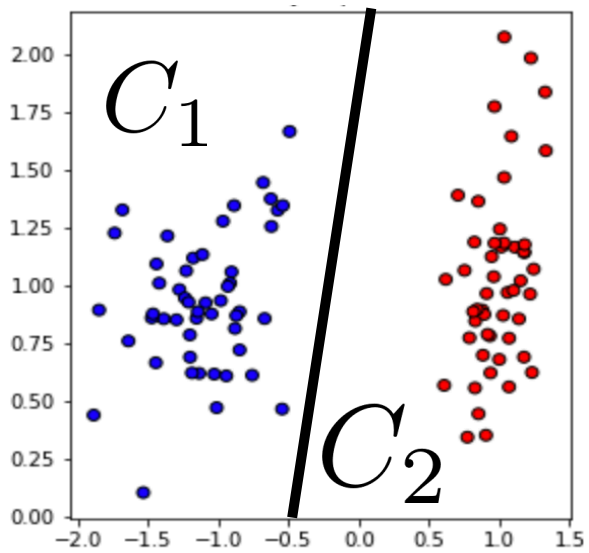
決定境界
(decision boundary)

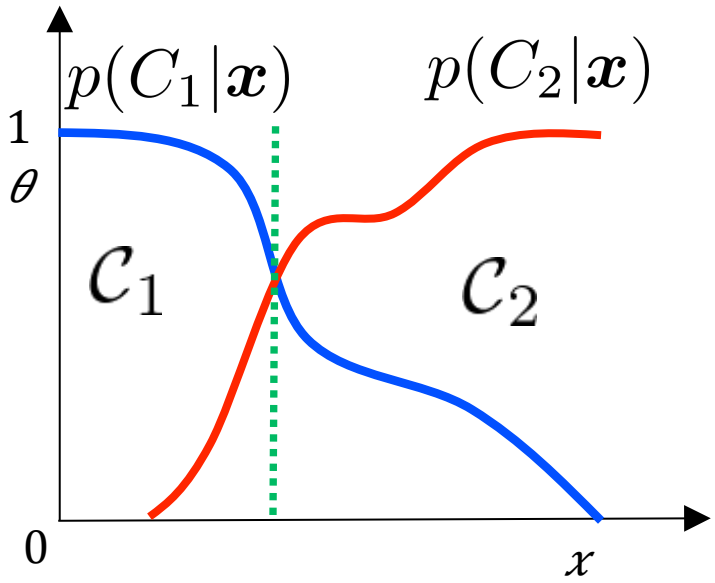


線形分類可能

線形分類不可能

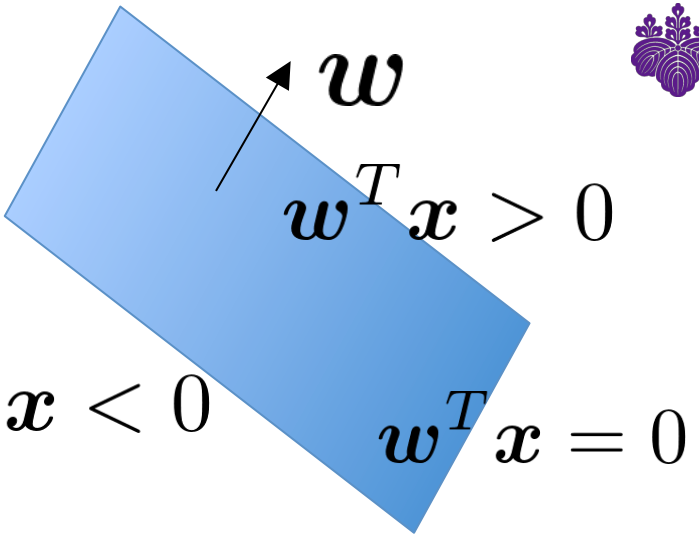
2つのアプローチ





- 決定的識別モデル
  - 決定境界 $f(\boldsymbol{x})$ を学習 $f(\boldsymbol{x}) > 0$ ならば $\boldsymbol{x}$ のラベルは C_1
- 確率的識別モデル
  - 条件付き確率 $p(C_1|\boldsymbol{x})$ を学習 $p(C_1|\boldsymbol{x}) > 0.5$ ならば $\boldsymbol{x}$ のラベルは C_1

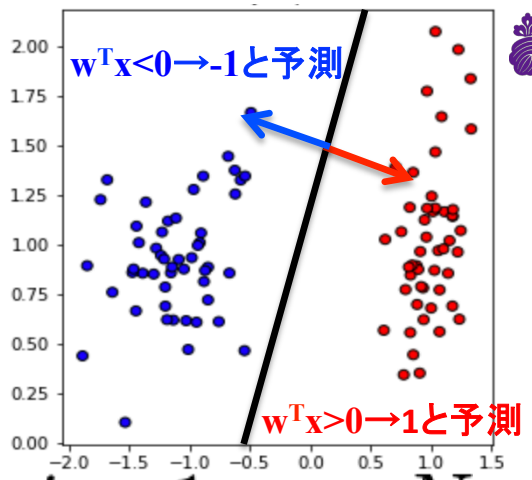
超平面



- D次元実数値空間 $\mathbb{R}^D$
- 超平面：$\mathbb{R}^D$ を二つの半空間に分割する平面
  - 超平面自体は $\mathbb{R}^{D-1}$ 次元の「平ら」な集合
- 例: 3次元空間の超平面は2次元平面
- $w_0 + w_1 x_1 + w_2 x_2 = 0$ パラメータ空間は $\mathbb{R}^3$
- 回帰と同様に,特徴量の1次元目を1とすれば

  $\boldsymbol{w}^T\boldsymbol{x} = 0$ D-1次元平面 (パラメータ空間は $\mathbb{R}^D$)
  - 半空間 $\boldsymbol{w}^T\boldsymbol{x} > 0$ $\boldsymbol{w}^T\boldsymbol{x} < 0$

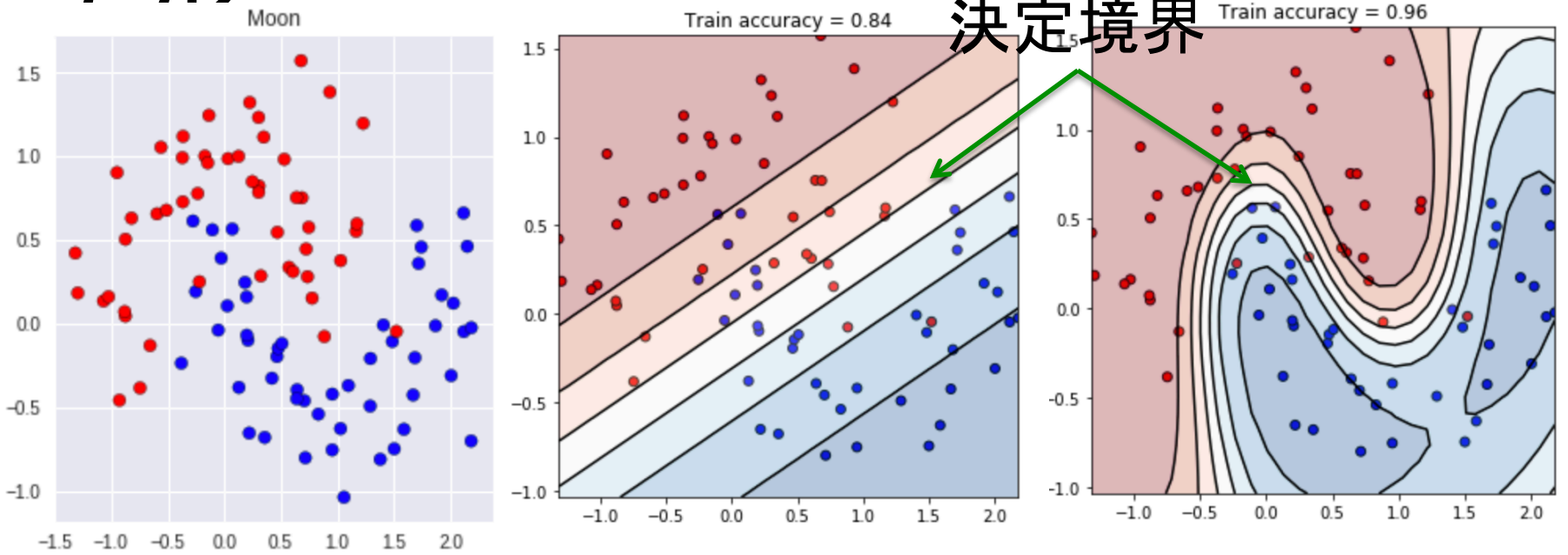
決定的識別モデル



- 教師あり学習
  - 事例　$x_i \in \mathbb{R}^D \quad t_i \in \{-1, 1\}, i = 1, \ldots, N$
  - 目標: 未知事例$\boldsymbol{x} \in \mathbb{R}^D$　のラベル　を当てる
  - 識別モデル　$f : \mathbb{R}^D \mapsto \{-1, 1\}$
- 最も単純な識別モデル
  - 超平面　$\boldsymbol{w}^T \boldsymbol{x} = 0$
    - 注意：回帰と異なり、w^Txそのものは分類の予測を与えない
  - 識別モデル

$$t_i = \mathrm{sgn}[\boldsymbol{w}^T \boldsymbol{x}_i] = \begin{cases} 1 \text{ if } \boldsymbol{w}^T \boldsymbol{x}_i > 0 \\ -1 \text{ o.w. } (\boldsymbol{w}^T \boldsymbol{x}_i \leq 0) \end{cases}$$

線形識別モデルと非線形識別モデル



線形モデルによる識別　　非線形モデルによる識別

- 決定的識別モデル：あるクラスと別のクラスを分ける境界を $f(\boldsymbol{x})=0$ で表す

1. まずは簡単な線形モデル $\boldsymbol{w}^T\boldsymbol{x}=0$ で
2. あとで非線形化(カーネル)します

最適な線形分類モデルとは？

- 回帰の場合：二乗誤差を最小化するような$\boldsymbol{w}$が最適なモデルであった

$$E(\boldsymbol{w}) = \sum_{i=1}^{N} (t_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2$$

正しい値 - 予測値

- 分類の場合：分類誤差を最小化するような$\boldsymbol{w}$が最適なモデルであるはず
  - 分類誤差＝「正しい分類」と「予測した分類」の不一致数

$$E(\boldsymbol{w}) = \sum_{i=1}^{N} I(t_i, \mathrm{sgn}(\boldsymbol{w}^T \boldsymbol{x}_i))$$

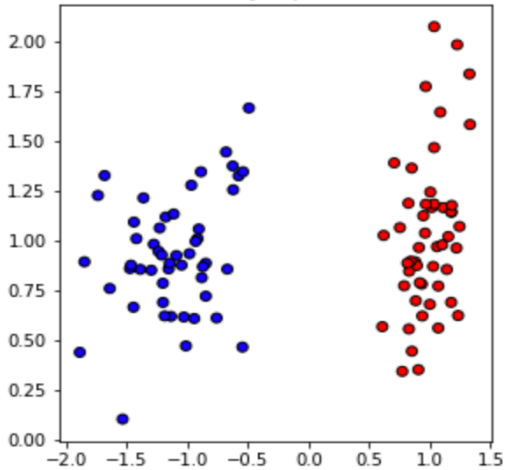
正しい分類　予測した分類

$$I(a,b) = \begin{cases} 0 \text{ if } a = b \\ 1 \text{ o.w. } (a \neq b) \end{cases}$$

一致するなら0

不一致なら1

「正しい分類」と「予測した分類」の不一致数をゼロにするには？

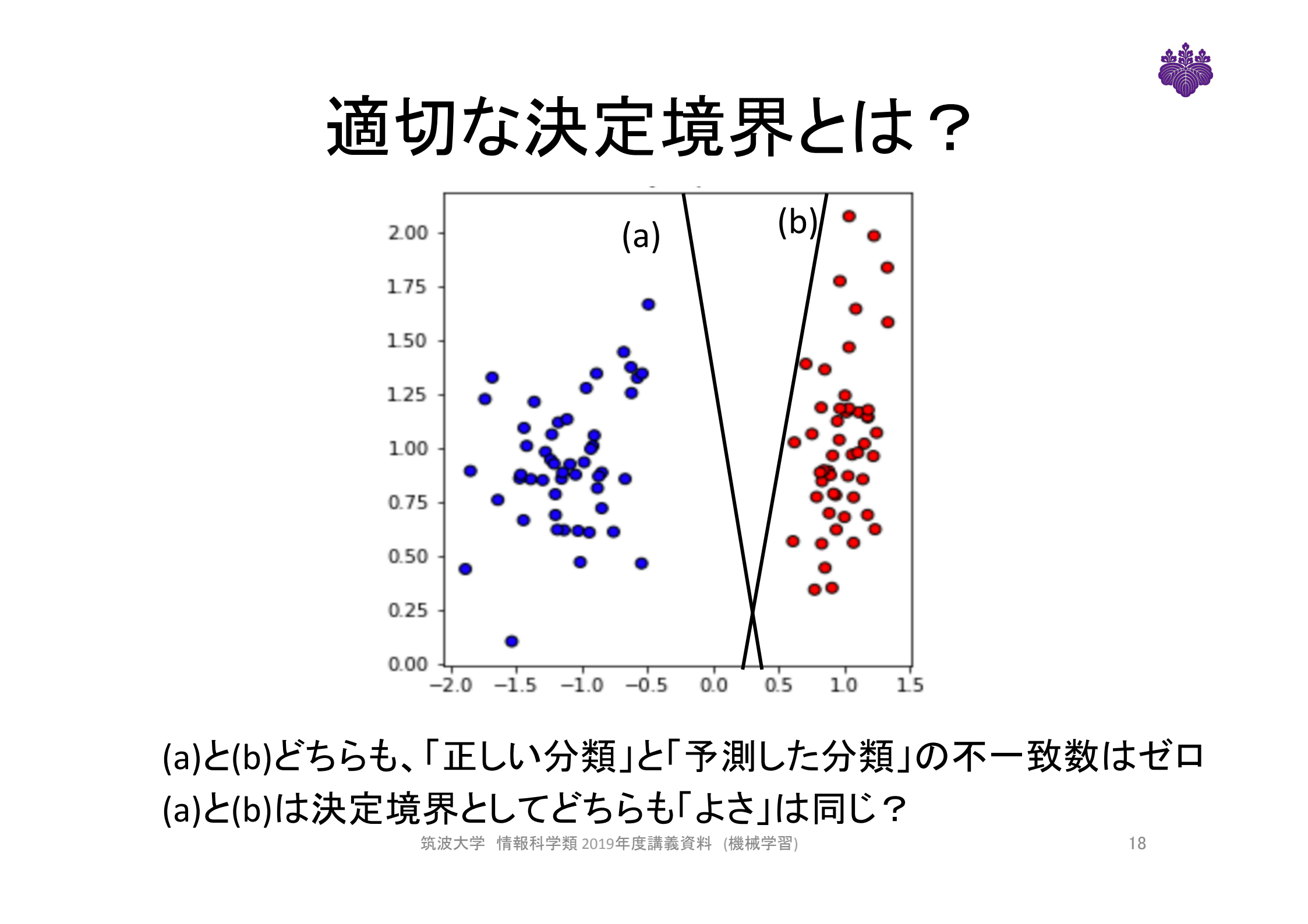


- 右図のようなデータについて、「正しい分類」と「予測した分類」の不一致数をゼロにするような超平面を考える
- $t_i(\boldsymbol{w}^T \boldsymbol{x}_i)$ の符号
  - 予測と正解が一致すれば正の値
  - 予測と正解が不一致ならば負の値

正解ラベル	線形モデルの予測	正解と予測の一致	$t_i(\boldsymbol{w}^T \boldsymbol{x}_i)$
$t_i > 0$	$\boldsymbol{w}^T \boldsymbol{x}_i > 0$	一致	$t_i(\boldsymbol{w}^T \boldsymbol{x}_i) > 0$
$t_i > 0$	$\boldsymbol{w}^T \boldsymbol{x}_i \leq 0$	不一致	$t_i(\boldsymbol{w}^T \boldsymbol{x}_i) \leq 0$
$t_i \leq 0$	$\boldsymbol{w}^T \boldsymbol{x}_i > 0$	不一致	$t_i(\boldsymbol{w}^T \boldsymbol{x}_i) \leq 0$
$t_i \leq 0$	$\boldsymbol{w}^T \boldsymbol{x}_i \leq 0$	一致	$t_i(\boldsymbol{w}^T \boldsymbol{x}_i) > 0$

- つまり、「正しい分類」を与える超平面を得るためには、以下の条件を満たすwを求めれば良い　(これを満たせば不一致数ゼロ)

$$t_i(\boldsymbol{w}^T \boldsymbol{x}_i) > 0, \forall i = 1, \ldots, N$$

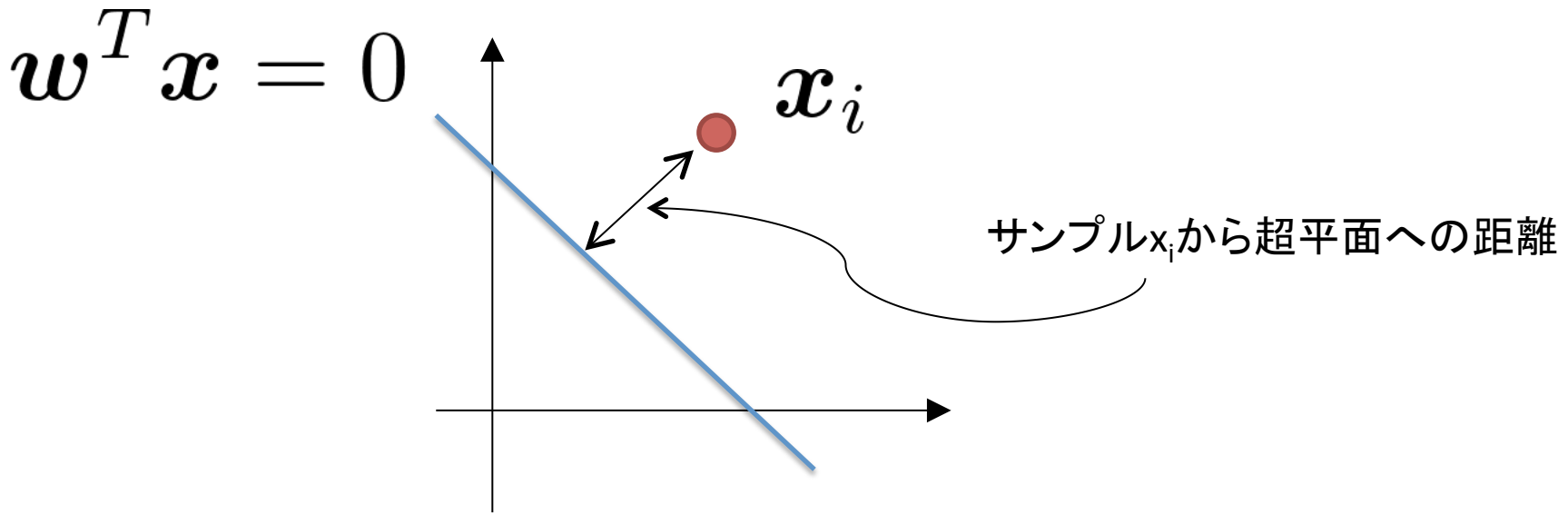


適切な決定境界とは？

(a)と(b)どちらも、「正しい分類」と「予測した分類」の不一致数はゼロ

(a)と(b)は決定境界としてどちらも「よさ」は同じ？

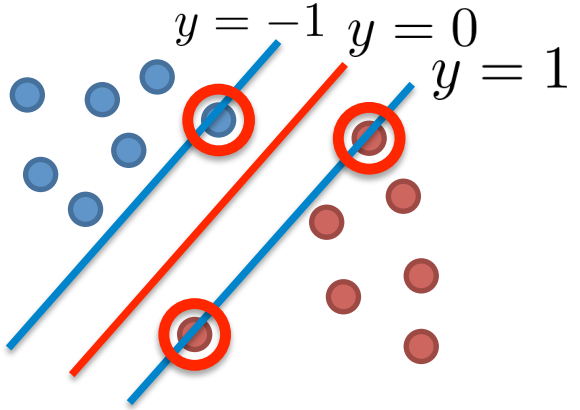
問題 マージン



5.1　マージン

超平面 $\boldsymbol{w}^T\boldsymbol{x} = 0$ と点 $\boldsymbol{x}_i$ の距離を求めなさい。(ヒント：点 $\boldsymbol{x}_i$ と超平面 $\boldsymbol{w}^T\boldsymbol{x} = 0$ の距離を d とおき, 点 $\boldsymbol{x}_i$ から超平面 $\boldsymbol{w}^T\boldsymbol{x} = 0$ への垂線を考える)

決定的識別の超平面が満たすべき条件



- 識別平面が満たすべき条件

1. 識別平面が異なるラベルのデータを分離
   - $\boldsymbol{w}^T\boldsymbol{x} > 0$ 側にラベル$t = +1$ のデータ
   - $\boldsymbol{w}^T\boldsymbol{x} \leq 0$ 側にラベル$t = -1$ のデータ
   - まとめると $i = 1, \ldots, N, (\boldsymbol{w}^T\boldsymbol{x}_i)t_i > 0$
2. 識別平面とデータの距離(マージン) $\dfrac{|\boldsymbol{w}^T\boldsymbol{x}|}{\|\boldsymbol{w}\|}$ の最大化

まとめると…

$$\max_{\boldsymbol{w}} \min_{i\in\{1,2,\ldots,N\}} \frac{|\boldsymbol{w}^T\boldsymbol{x}_i|}{\|\boldsymbol{w}\|_2} \text{ subject to } t_i(\boldsymbol{w}^T\boldsymbol{x}_i) > 0$$

最小のマージンを最大にしたい　どのデータも正しく分類して欲しい

サポートベクターマシン (ハードマージン)

- 前ページの最適化問題:

$$\max_{\boldsymbol{w}} \min_{i\in\{1,2,\ldots,N\}} \frac{|\boldsymbol{w}^T\boldsymbol{x}_i|}{\|\boldsymbol{w}\|_2} \text{ subject to } t_i(\boldsymbol{w}^T\boldsymbol{x}_i) > 0$$

- 超平面 $\boldsymbol{w}^T\boldsymbol{x} = 0$ は定数倍に対して不変($cw^Tx=0$)なので、$\min_i t_i(\boldsymbol{w}^T\boldsymbol{x}_i) = 1$ となるように、$c>0$を定めスケーリング
- このとき、目的関数は以下のように変換できる

$$\max_{\boldsymbol{w}} \min_{i\in\{1,2,\ldots,N\}} \frac{|\boldsymbol{w}^T\boldsymbol{x}_i|}{\|\boldsymbol{w}\|_2} \Rightarrow \min_{\boldsymbol{w}} \|\boldsymbol{w}\|_2^2$$

$t_i=+1/-1$ゆえ、$|w^Tx_i| = t_iw^Tx_i$

- 最終的には、以下を解けば良い

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_2^2 \text{ subject to } \min_i t_i(\boldsymbol{w}^T\boldsymbol{x}_i) \geq 1, i = 1, \ldots, N$$

マージンを最大にしたい　　どのデータも正しく分類して欲しい

凸二次計画問題

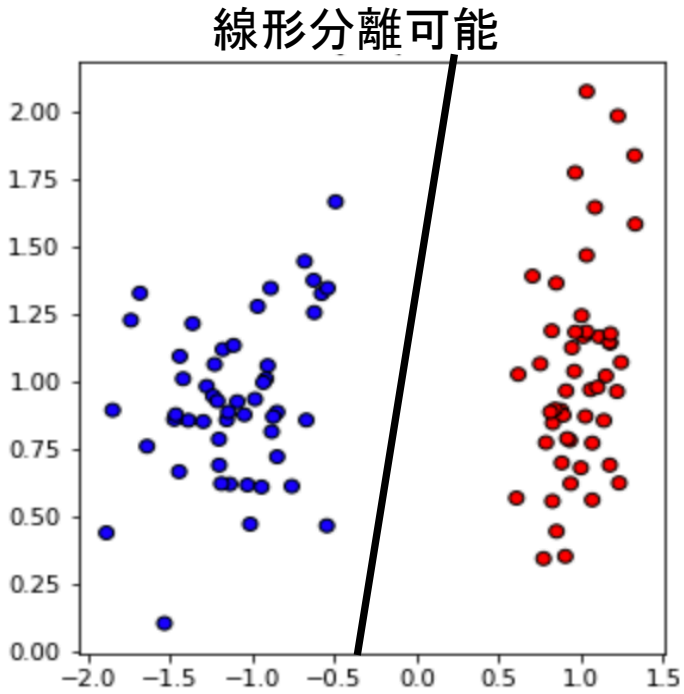
- 凸二次計画問題(QP, quadratic programming)
  - 線形等式・不等式制約のある二次形式の最適化

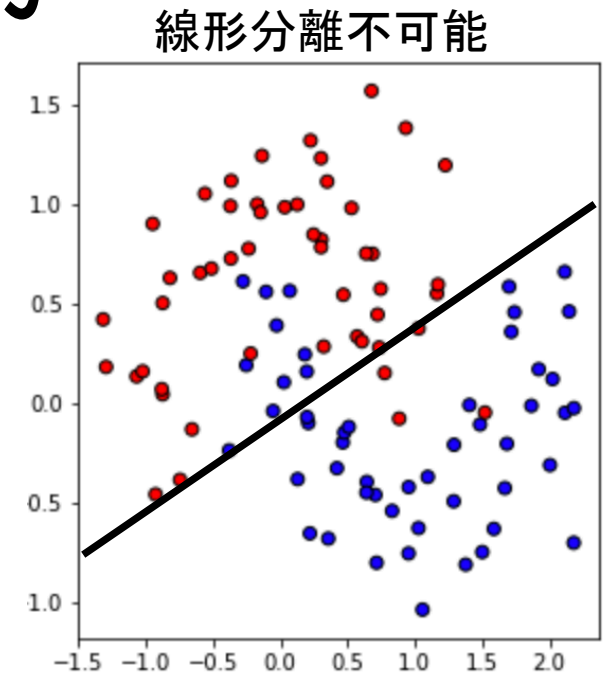
$$\text{minimize } \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{c}$$

$$\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}, \boldsymbol{C}\boldsymbol{x} \geq \boldsymbol{d}$$

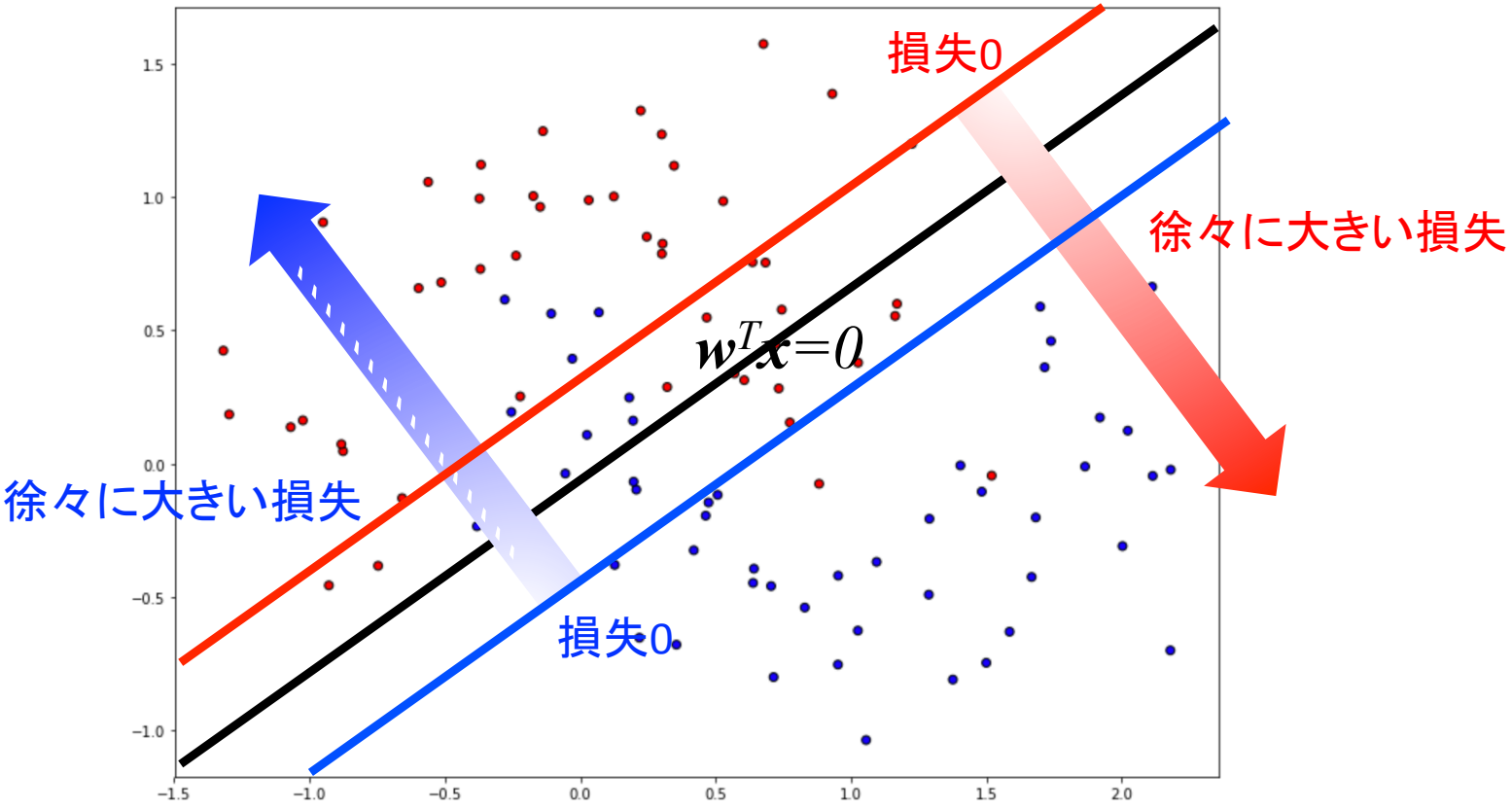
  - 先ほどの最適化はQPに持ち込める
  - QPになれば、QP最適化パッケージはいろいろある
- ハードマージンSVMはQPに帰着できる
- ただし、QPは基本的に重い計算

線形分離不可能な場合にどうするか





- どのような超平面でも「正しい分類」と「予測した分類」の不一致数をゼロにできない＝線形分離不可能
- 識別平面からの距離に応じたペナルティを与える


損失0
徐々に大きい損失
$w^T x = 0$
徐々に大きい損失
損失0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

ソフトマージン

サポートベクターマシン (ソフトマージン)

- ハードマージンSVMの最適化問題:

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_2^2 \text{ subject to } \min_i t_i(\boldsymbol{w}^T \boldsymbol{x}_i) \geq 1, i = 1, \ldots, N$$

- 制約の代わりに、正しく分類されない点にはペナルティを与える
  - $t_i(\boldsymbol{w}^T \boldsymbol{x}_i) > 1$ (マージン1以上) → ペナルティ0
  - $t_i(\boldsymbol{w}^T \boldsymbol{x}_i) \leq 1$ (マージンは1より小さい) →ペナルティ $1 - t_i(\boldsymbol{w}^T \boldsymbol{x}_i)$
- まとめると...
  - データ $\boldsymbol{x}_i$ に以下のペナルティを与える $\max\{0, 1 - t_i(\boldsymbol{w}^T \boldsymbol{x}_i)\}$
  - 全データのペナルティが最小になるような $\boldsymbol{w}$ を求める

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_2^2 + C \sum_{i=1}^{N} \max\{0, 1 - t(\boldsymbol{w}^T \boldsymbol{x}_i)\}$$

マージンを最大にしたい　データの分類についてのペナルティを小さくしたい

回帰とSVMを見比べる

- 回帰 $\min_{\boldsymbol{w}} \lambda \|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{N} (t_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2$

  誤差を小さく

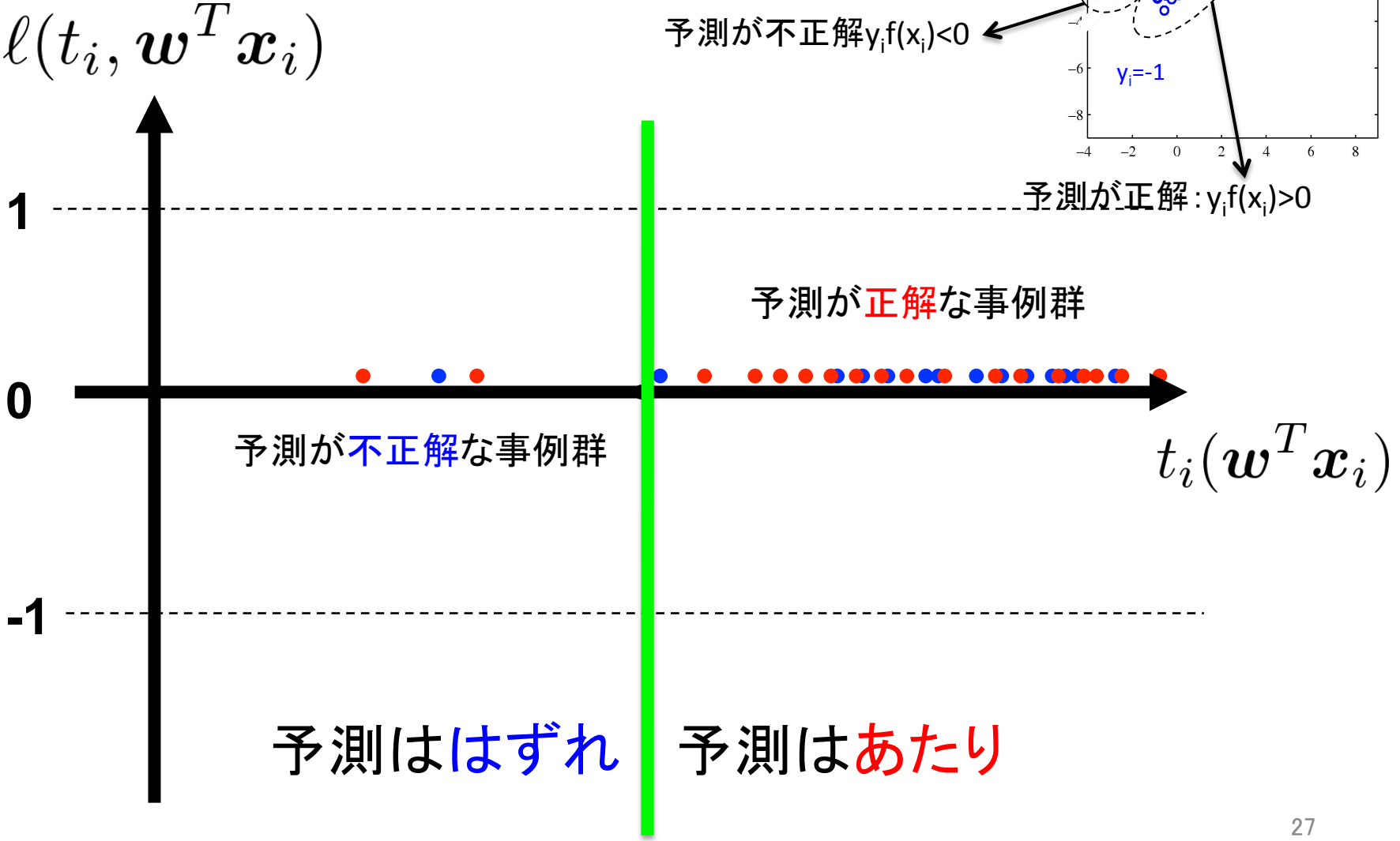
- SVM $\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_2^2 + C \sum_{i=1}^{N} \max\{0, 1 - t_i \boldsymbol{w}^T \boldsymbol{x}_i\}$

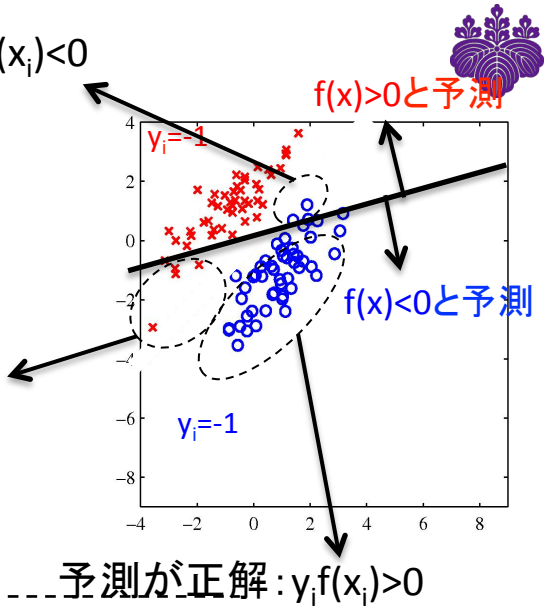
  ペナルティを小さく

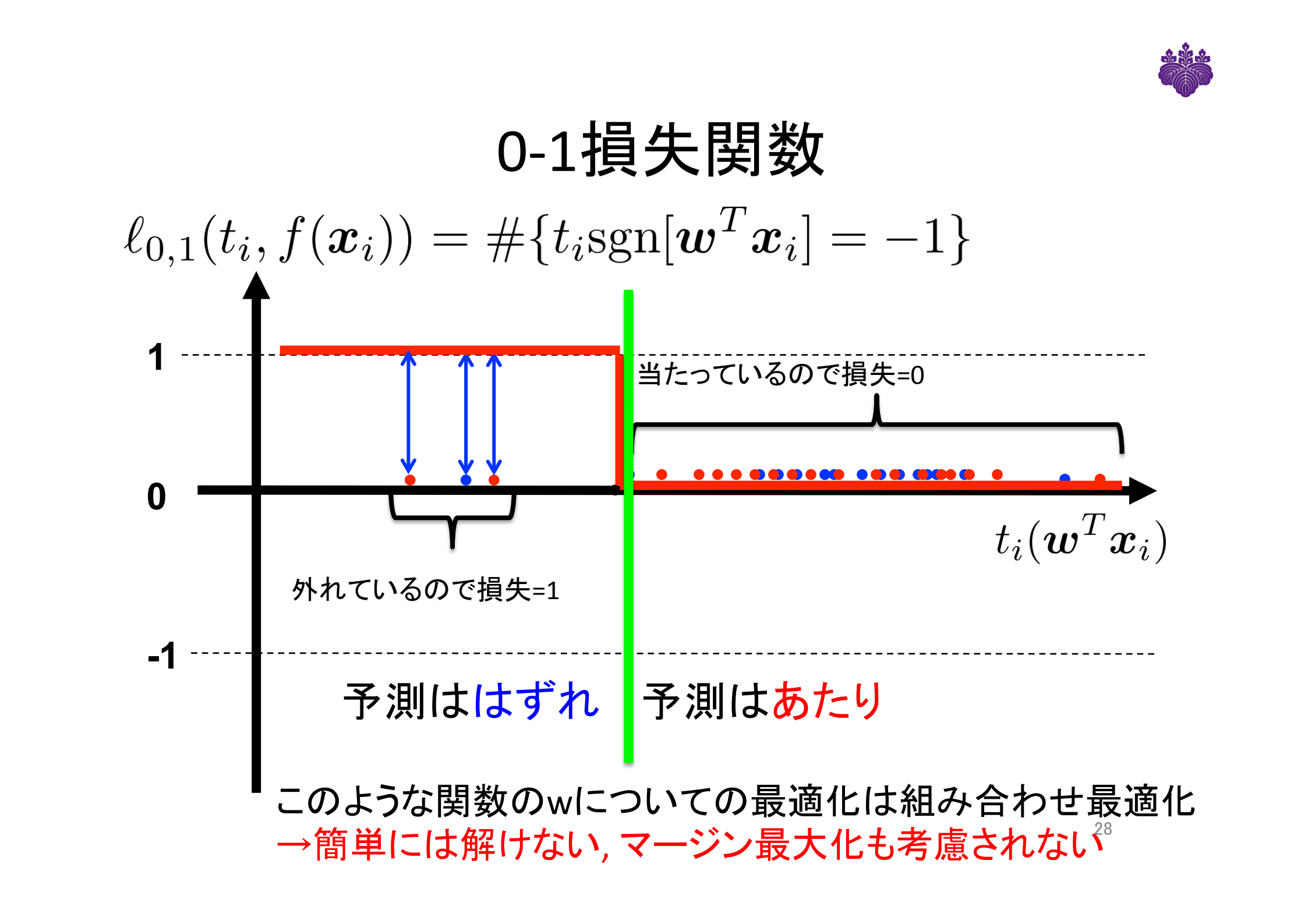
L2正則化項　各データにおける個別の損失の和

- 各データの個別のペナルティ=損失関数で一般化
  - 二乗損失(回帰) $\ell_{\mathrm{sq}}(t_i, \boldsymbol{w}^T \boldsymbol{x}_i) = (t_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2$
  - ヒンジ損失(SVM) $\ell_{\mathrm{hinge}}(t_i, \boldsymbol{w}^T \boldsymbol{x}_i) = \max\{0, 1 - t_i(\boldsymbol{w}^T \boldsymbol{x}_i)\}$

$t_i(\boldsymbol{w}^T\boldsymbol{x}_i)$で損失を測る







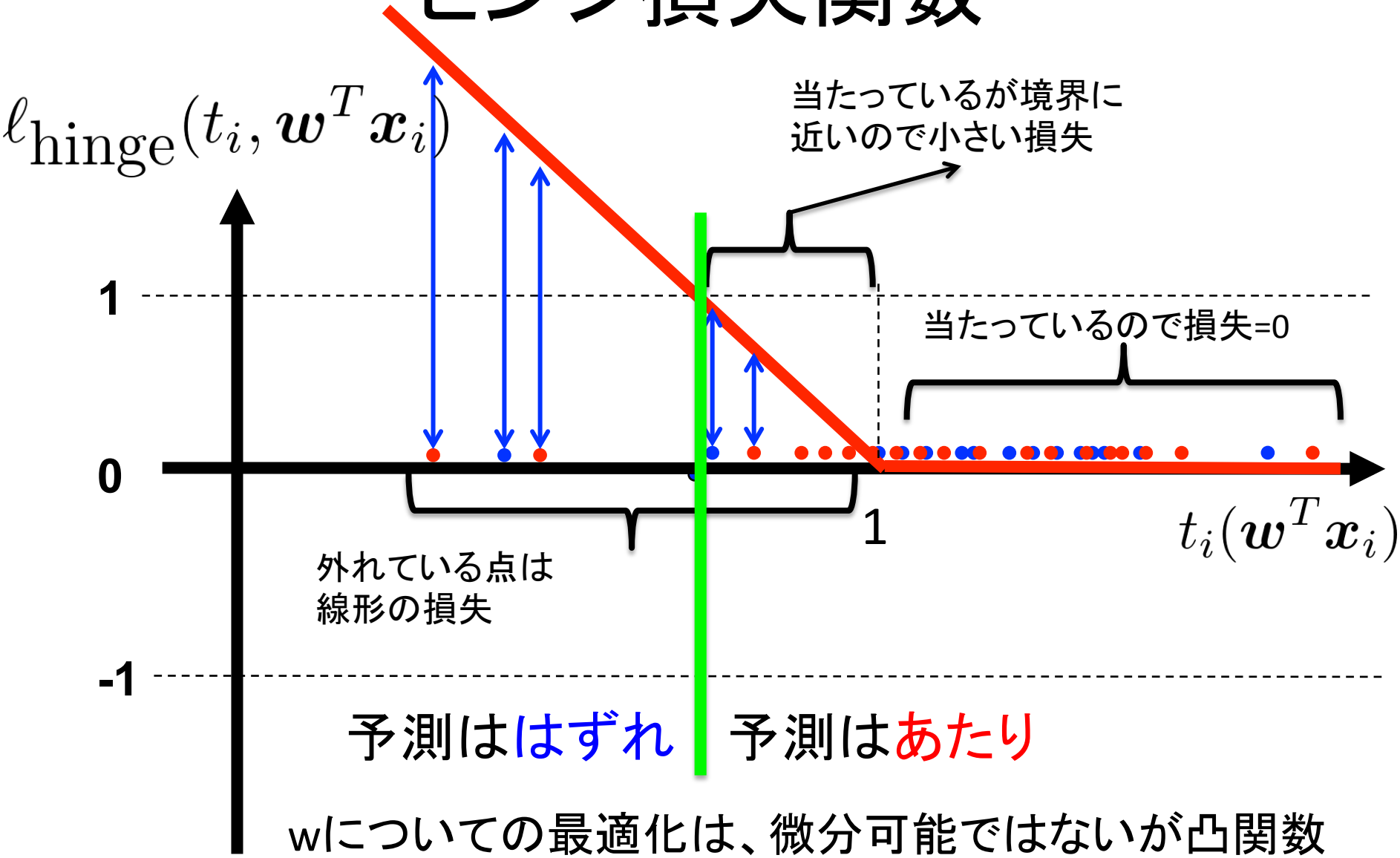
0-1損失関数

$$\ell_{0,1}(t_i, f(\boldsymbol{x}_i)) = \#\{t_i \mathrm{sgn}[\boldsymbol{w}^T \boldsymbol{x}_i] = -1\}$$

このような関数のwについての最適化は組み合わせ最適化

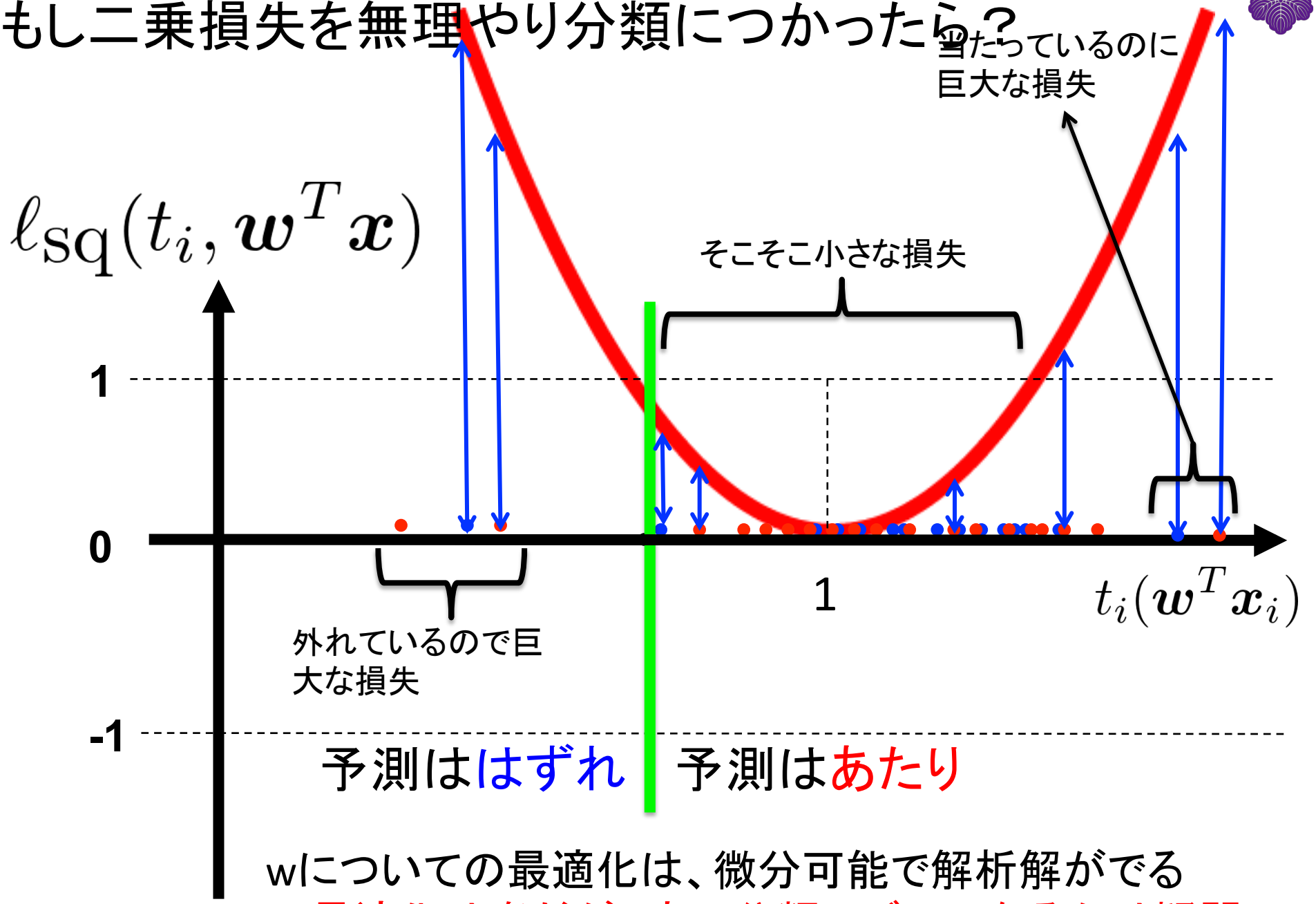
→簡単には解けない, マージン最大化も考慮されない

ヒンジ損失関数



wについての最適化は、微分可能ではないが凸関数
→それなりに効率的に解く方法が存在

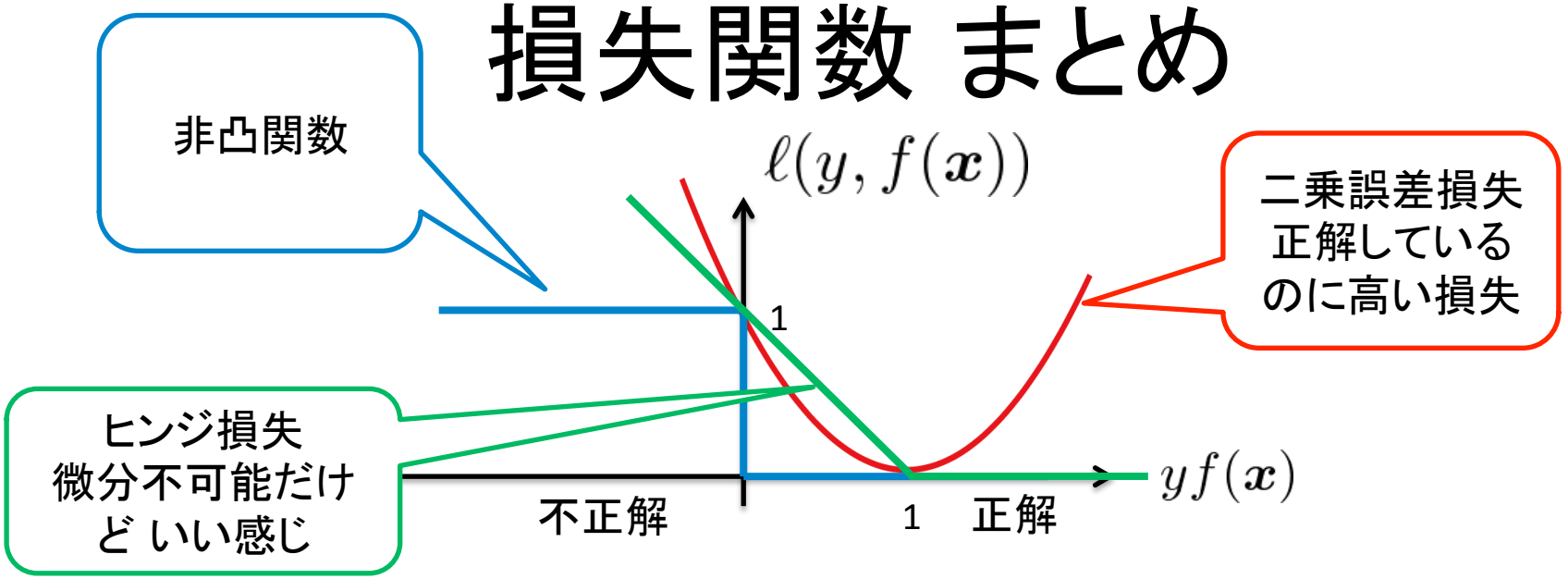
もし二乗損失を無理やり分類につかったら？



wについての最適化は、微分可能で解析解がでる

→最適化は楽だが、良い分類モデルになるかは疑問

損失関数 まとめ



- ヒンジ損失 $\ell_{\mathrm{hinge}}(y, f(\boldsymbol{x})) = \max\{0, 1 - yf(\boldsymbol{x})\}$
  - 凸関数: 誤差関数も凸関数→大域的最適化可能
  - 誤差の増え方が緩やか→外れ値に過剰適合しない
- ではどうやってヒンジ損失で定義された目的関数を最適化するのか？

損失関数の性質

- 二乗損失 $\ell_{\mathrm{sq}}(t_i, \boldsymbol{w}^T\boldsymbol{x}) = (t_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2$
  - 目標値との差の二乗、回帰に利用
  - 凸関数、微分可能
  - 総損失の最小化
    - 解析的に求まる, 厳密な最適解
    - 最急降下法でも解ける、近似的な解

- ヒンジ損失 $\ell_{\mathrm{hinge}}(t_i, \boldsymbol{w}^T\boldsymbol{x}_i) = \max\{0, 1 - t_i(\boldsymbol{w}^T\boldsymbol{x}_i)\}$
  - 超平面からの距離(正しく分類されない側にどれだけ離れているか)
  - 凸関数、ただし微分不可能
  - 総損失の最小化
    - QP（ソフトマージンSVMもQPで解ける）, 厳密な最適解
    - 劣勾配降下法, 近似的な解

再掲

劣勾配(sub-gradient))

- 微分不可能な項を含む目的関数をどうやって最適化する？

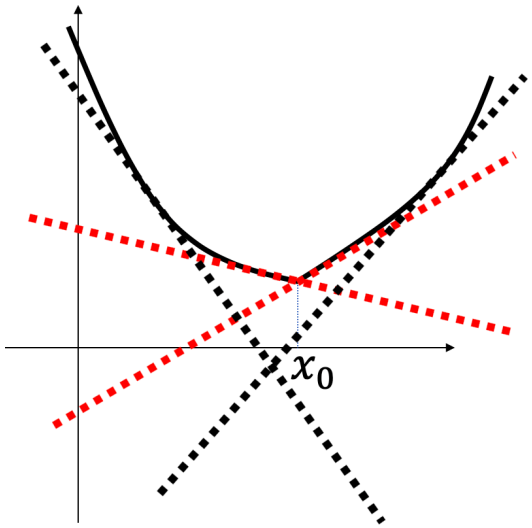
$$E(\boldsymbol{w}) = \sum_{i=1}^{N}(t_i - \boldsymbol{w}^T\boldsymbol{x})^2 + \lambda\sum_{i=1}^{D}|w_i|$$

L1正則化項は微分不可能

- 劣勾配 $\partial f(\mathbf{w}_0) = \{\mathbf{g} \in \mathbf{R}^D | \forall \mathbf{w}, f(\mathbf{w}) - f(\mathbf{w}_0) \geq (\mathbf{w} - \mathbf{w}_0)^T\mathbf{g}\}$
  - 例　$f(\mathbf{w}) = |w_1| + 2|w_2|$

$\mathbf{w}_0 = (1,0)^T$での劣微分

$$\partial f(\mathbf{w}_0) = \left\{ \begin{pmatrix} 1 \\ 2z \end{pmatrix} \middle| z \in [1,-1] \right\}$$



再掲

劣勾配降下法(sub-gradient descent)

- 勾配降下法における勾配を劣勾配に置き換え

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \partial E(\mathbf{w}^{(t)})$$

- 効率がよくない(ことがある)
- wがスパースにならない(ことがある)
  - 劣勾配は一意に定まらないことに関係

確率的劣勾配降下法によるSVMの解法

1. $\boldsymbol{w}^{(0)}$を適当に初期化
2. For i=1,…:

$$\boldsymbol{w}^{(t)} = \boldsymbol{w}^{(t-1)} - \begin{cases} \eta(-t_i \boldsymbol{x}_i + \frac{2}{C}\boldsymbol{w}) & \text{if } 1 - t_i \boldsymbol{w}^T \boldsymbol{x}_i \geq 0 \\ \eta(\frac{2}{C}\boldsymbol{w}) & \text{if } 1 - t_i \boldsymbol{w}^T \boldsymbol{x}_i < 0 \end{cases}$$

5.2　確率的劣勾配降下法によるSVMの最適化

演習

$\boldsymbol{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{iD} \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix}, \boldsymbol{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix}, \boldsymbol{w} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{pmatrix}$ とする。 $t_i \in \{-1, 1\}$ である。事例群 $\{(\boldsymbol{x}_i, t_i)\}_{i=1}^N$ を使ってSVMによる線形分類モデルを確率的列勾配降下法を使って求めることを考える。目的関数は以下の通り。

$$E(\boldsymbol{w}) = \frac{1}{C}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{N} \max\{0, 1 - t_i(\boldsymbol{w}^T \boldsymbol{x}_i)\} \tag{2}$$

目的関数は損失関数項 (hinge 損失) と正則化項 (L2 正則化) からなる。確率的劣勾配法では、パラメータを、一つの事例に対する損失項と正則化項の劣勾配を用いて更新する.

1. 一つの事例に対する損失 $\ell_i(\boldsymbol{w}) = \max\{0, 1 - t_i(\boldsymbol{w}^T \boldsymbol{x}_i)\}$ の劣勾配を求めなさい
2. 正則化項の勾配を求めなさい
3. 確率的劣勾配降下法によるSVMの最適化における更新式を導出しなさい。ただし、ステップサイズパラメータを η とする。

演習　ヒンジ損失と二乗損失

5.3　ヒンジ損失と二乗損失

hinge_and_squared_loss.ipynb と hinge_and_squared_loss_withOutlier.ipynb を実行して以下の問いに答えなさい。この二つのプログラムは、ランダムに生成された 40 点の 2 クラス分類問題のための訓練事例について、ヒンジ損失と二乗損失で線形分類モデルを学習し、その決定境界を表示している。前者は外れ値を含まない訓練事例、後者は外れ値を含む訓練事例で学習させた結果である。両プログラムとも、一つ目のセクションではヒンジ損失を用いて、二つ目のセクションでは二乗損失を用いて分類のための超平面を学習させている。以下の問いに答えよ。

1. 外れ値を含まない場合の、ヒンジ損失と二乗損失で学習させた場合の線形モデルの違いについて（違いがあれば）説明せよ

2. 外れ値を含む場合の、ヒンジ損失と二乗損失で学習させた場合の線形モデルの違いについて（違いがあれば）説明せよ

3. 問題 1 および問題 2 において、ヒンジ損失と二乗損失で違いが生じた場合、なぜその違いが生じたのか説明せよ

二値分類問題の評価

- 事例　　$(\boldsymbol{x}_i, y_i)$ where $\boldsymbol{x}_i \in \mathbb{R}^D, y_i \in \{\mathcal{C}_1, \mathcal{C}_2\}$
- 予測　　$\hat{y}_i = f(\boldsymbol{x}_i)$
- 二値分類問題
  - 単にAかBか予測するケース(写真の動物はネコか犬か？)
  - Aならばアラートを出したいケース(この患者はガンか?)
  - 後者のケースでは、より多様な観点での精度の評価が必要

混同行列

		正解	
		0	1
予測	0真(P)	TP	FP
	1偽(N)	FN	TN

アラートを出した

アラートを出さなかった

アラートを出すべき　アラートを出してはならない

- 混同行列
  - 予測と真のラベルが一致=True(T)、不一致=False(F)
  - 真と予測すること=Positive(P)、偽と予測すること=Negative(N)
  - True positive(TP), false positive(FP), false negative(FN), true negative (TN)

分類器の性能評価

		正解	
		0	1
予測	0真(P)	TP	FP
	1偽(N)	FN	TN

- 正解率 (accuracy)=(TP+TN)/(TP+FP+FN+TN)
  - 予測がどの程度当たったか？
- 精度(precision) =TP/(TP+FP)
  - 0と予測したデータのうち，実際に0である割合
  - 取りこぼしてもいいが、間違いが許されない場合に使う
  - Aさんは犯人かどうか？
- 偽陽性率(false positive rate)
  - FP/(TN+FP)
  - アラートを出してはいけない事例のうち、アラートを出してしまった事例の割合
- 再現率 (recall, true positive rate) =TP/(TP+FN)
  - 実際に0であるもののうち，0と予測した割合
  - 間違ってもいいが、取りこぼしがゆるされない場合に使う
  - Aさんは癌かどうか？
- F値 (F-measure)=2 Recall*Precision/(Recall + Precision)
  - 精度と再現率の調和平均
  - 総合的な評価につかう
- 精度と再現率を両方高くするのは難しい

分類器の性能評価例

- 正解率が大事な例
  - この写真の動物は犬か猫か?
  - 正解率=(48+40)/100=0.88

		正解	
		犬	猫
予測	犬	48	2
	猫	10	40

- 精度が大事な例
  - 特徴量x=1の人は、犯罪を犯す?
  - 精度=5/(5+0)=1.0

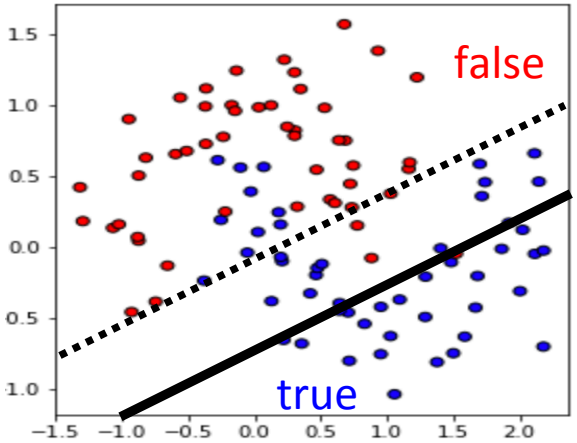
		正解	
		犯罪を犯す	犯さない
予測	犯罪を犯す	5	0
	犯さない	20	75

- 再現率が大事な例
  - 検査結果x=1の人は、癌であろ

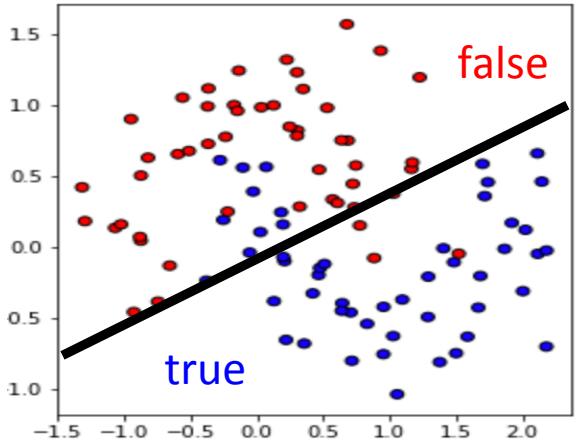
		正解	
		癌である	癌でない
予測	癌である	9	23
	癌でない	1	67

精度と再現率のコントロール

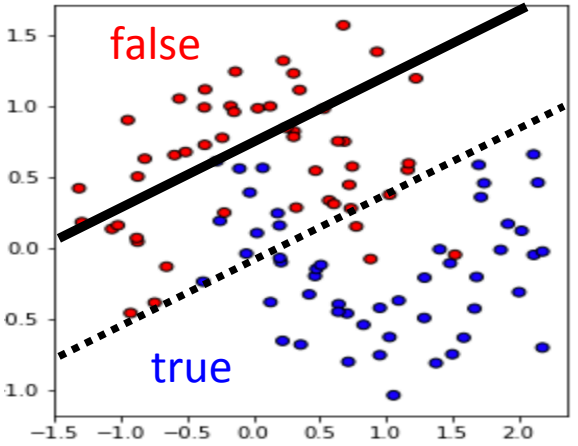
青をtrue, 赤をfalseとする



精度は高い
再現率は低い



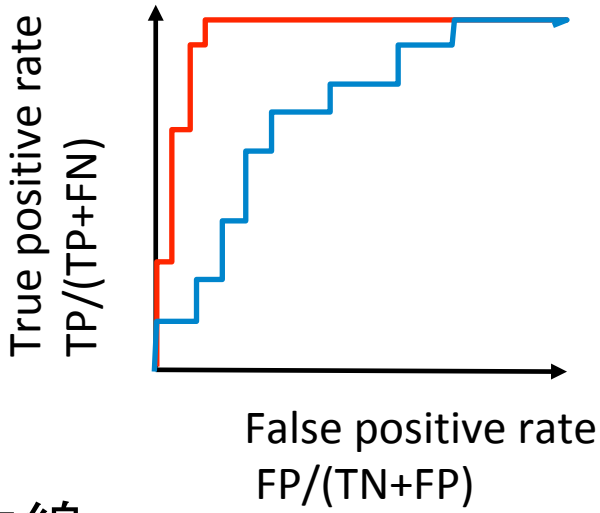
精度も、再現率もほどほど
正解率が高い



精度は低い
再現率は高い

線形分類器の場合、正解率を最大にする分類器を表す超平面と同じ法線ベクトルを持つ平面について、正解率を最大にする分類器からの距離をコントロールすることで、精度と再現率のバランスをコントロールできる
(2値分類のロジスティック回帰の場合、確率の閾値を0.5より低く(or 高く)する)

ROC曲線とAUC



		正解	
		0	1
予測	0真(P)	TP	FP
	1偽(N)	FN	TN

- ROC曲線
  - 閾値を変えながらFP vs TPをプロット
  - その下の面積が広いほど、「良い」分類器 (赤の分類器が良い分類器)
  - ROC曲線の下の面積= AUC (area under the curve)を性能評価として用いる
  - 正解率の評価と同様、訓練事例で学習し、テスト事例でAUCを評価する
    - K-fold CVを使っても良い

→演習 ROC曲線
→演習 SVMによる分類

性能評価指標の使い分け

- 分類器の性能を閾値に依存せずに評価したい
  - AUC
- 分類器の性能を特定の閾値について評価したい
  - t=1/-1の重要度が同じである場合　e.g. 写真に写っているのは犬か猫か
    - 精度
  - t=1がより重要である場合
    - 予測の質が重要: precision e.g. この人は犯罪者か
    - 取りこぼさないことが重要: recall　e.g. この人は肺がんか
  - どちらも着目したい場合
    - F-measure

5.4　ROC 曲線の作成と正則化パラメータのチューニング

IRIS データを SVM で分類するプログラム svm_iris_roc_visualization_auc_tuning.ipynb を実行して以下の問いに答えなさい。ここでは, 与えられたあやめの特徴から、それが virginica か (true)、そうでないか (false) を予測するモデルを作成している。損失関数にはヒンジ損失を用いて、C=0.01 と固定する (C はここでは $L2$ 正則化パラメータの逆数)。切片を変化させながら、ROC 曲線を作成し、AUC を評価する。以下の問いに答えよ。

1. この分類モデルにおける false positive とはどのような状況か

2. virginica のデータを分類モデルで予測した時に、実際に virginica と予測する確率をなんと呼ぶか

3. セクション ”Next, we will learn a SVM classifier, and visualize the ROC curve ” のコードを実行して以下の問いに答えよ。このプログラムで学習した分類器において, あやめデータを予測した時に、virginica でないにも関わらず、virginica と予測してしまう確率を 0.2 以下に抑えた時に、virginica のデータを virginica と正しく予測する確率は最大でどの程度か

4. セクション ”Now, we learn how to tune the parameter using the AUC scores” のコードを実行して以下の問いに答えよ。$C = 10^{-1}$ におけるモデルの AUC は 0.945 より大きく、それ以外の C におけるモデルの AUC は 0.945 より小さい。このとき、$C = 10^{-1}$ と設定した場合、任意の false positive rate において, そのモデルの true positive rate を他のモデルの true positive rate 以上にすることが可能か？理由とともに答えなさい。

まとめ

- 最小マージンを最大化する識別超平面による分類を、サポートベクターマシンと呼ぶ
- サポートベクターマシンの目的関数は、各サンプルにおける予測結果に対するヒンジ損失の総和＋正則化で表現される
- サポートベクターマシンの目的関数は微分可能ではないため、列勾配に基づく最適化が用いられる
- 二値分類のための分類器の性能評価は、目的に応じて様々な指標が用いられる