

2章のストーリー

- 清原は、市の医療費削減のために健診結果から糖尿病の発病を予測するサービスを立ち上げたいと考える
- さやかは識別問題の解法として、ロジスティック識別と決定木について教える

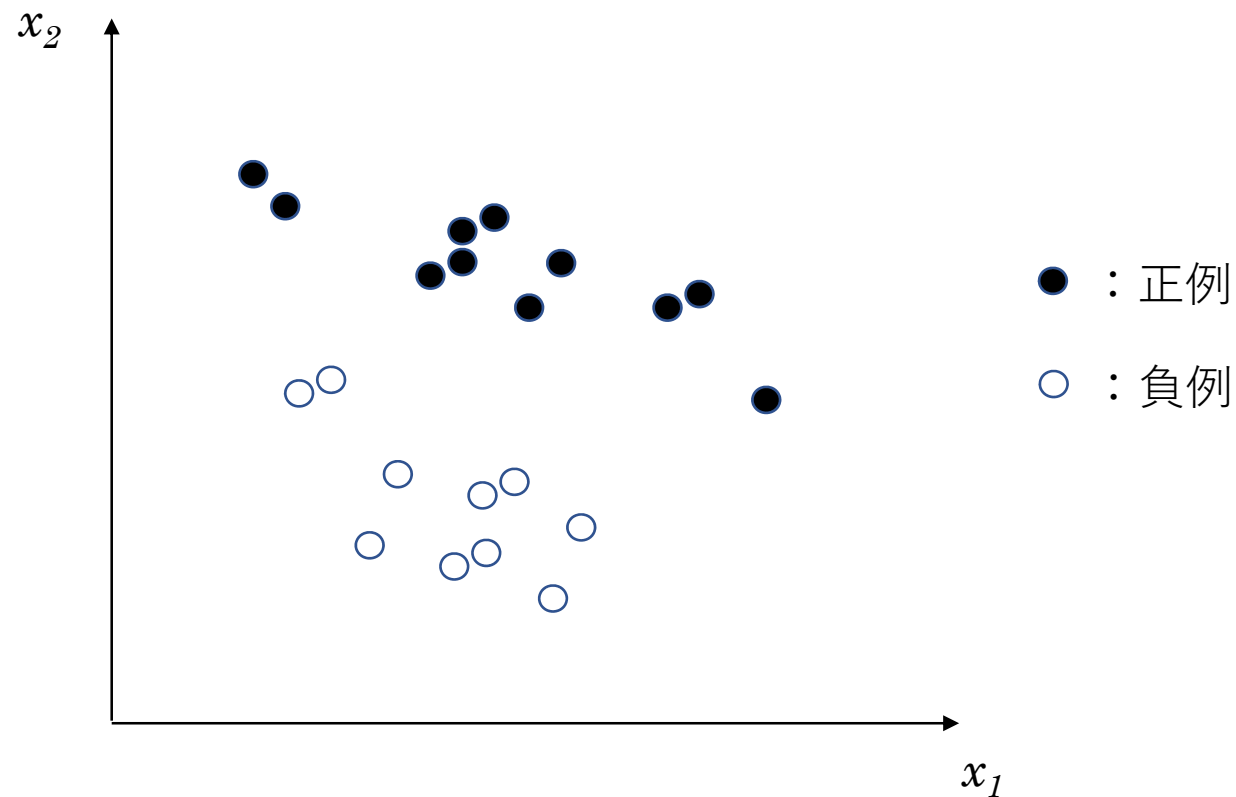
3章のストーリー

- 清原は、教わった方法で100%の正解率を実現し、同僚の九条の助けを借りて糖尿病診断のwebサイトを立ち上げる
- しかし判定精度が悪く、多くの苦情を受ける
- さやかは機械学習を使ったシステムの正しい性能予測法を教える

基礎的な識別（2章）

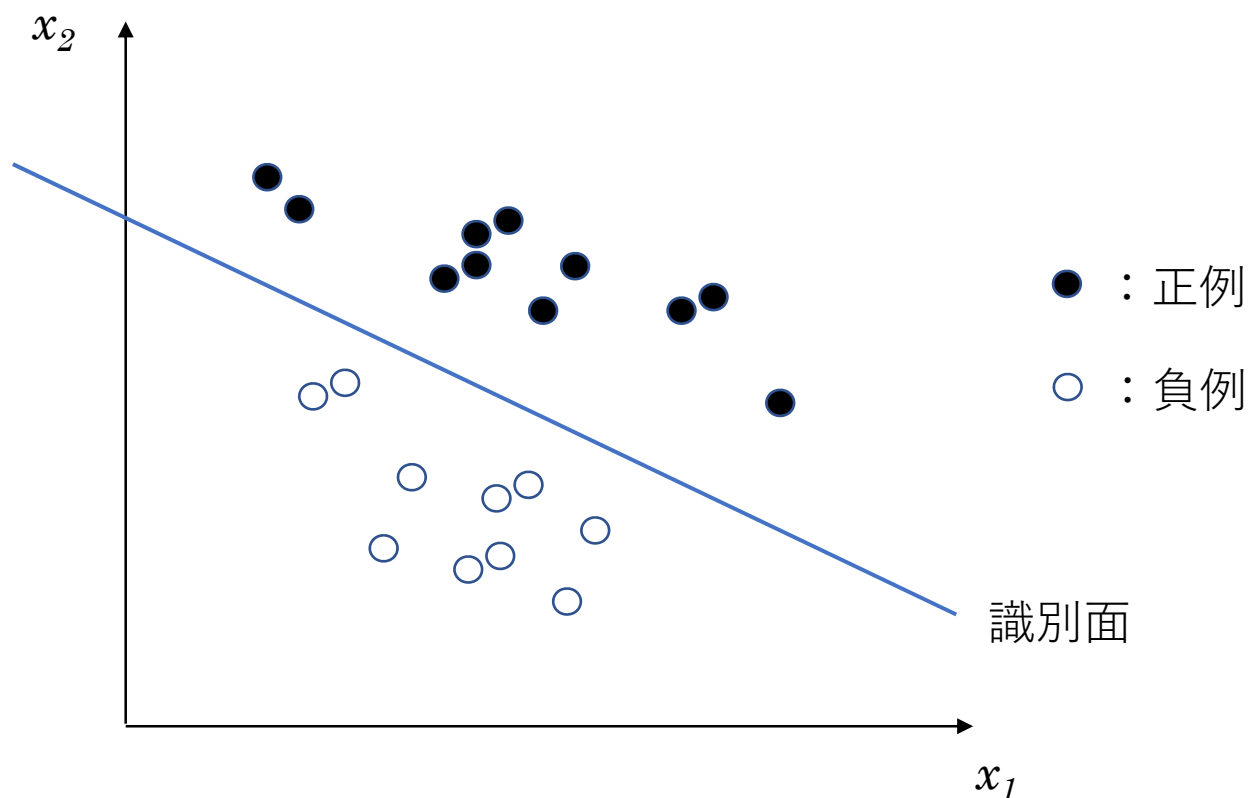
識別

- 識別とは
 - 教師あり学習問題
 - 特徴からクラスを予測する（できれば確率も得たい）



ロジスティック識別

- 2クラス分類でのロジスティック識別の考え方
 - 入力された特徴が正例である確率を得たい
 - 確率=0.5の点の集合を識別面と考える



ロジスティック識別

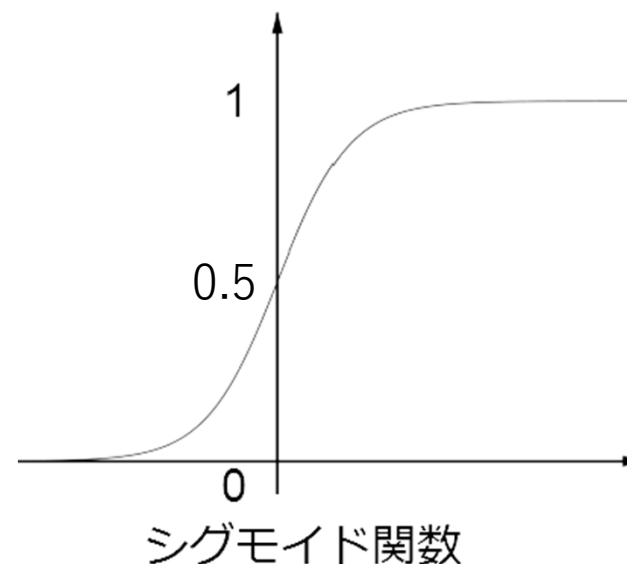
- 識別面の式

$$\hat{g}(\boldsymbol{x}) = w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_dx_d + w_0 = \boldsymbol{w}^T \boldsymbol{x} = 0$$

- 正例の \boldsymbol{x} に対しては $\hat{g}(\boldsymbol{x}) > 0$
- 負例の \boldsymbol{x} に対しては $\hat{g}(\boldsymbol{x}) < 0$
- これを確率と対応付けたい \Rightarrow シグモイド関数

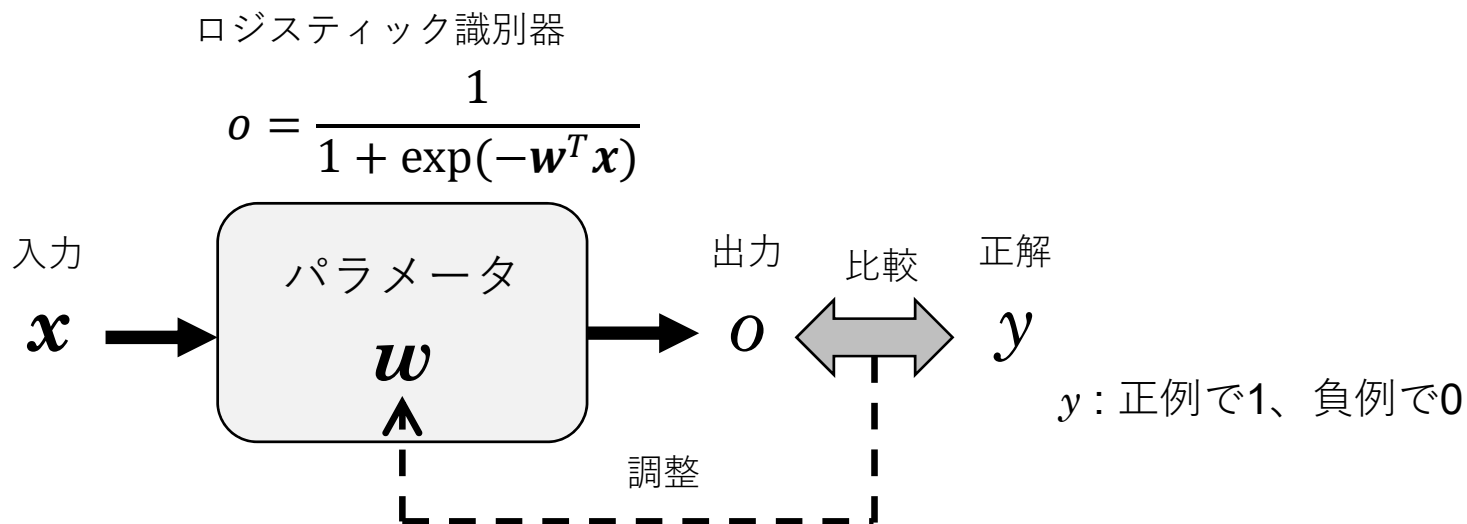
$$P(\text{正}|\boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{w}^T \boldsymbol{x})}$$

$\exp(x) : e^x$
 $e = 2.71828\dots$



ロジスティック識別

- 係数 w の求め方



- 尤度（モデルのもっともらしさ）の定義

$$P(D|w) = \prod_{x_i \in D} o_i^{y_i} (1 - o_i)^{(1-y_i)}$$

D : 全データ

ロジスティック識別

- 尤度の最大化

⇒ 対数尤度の最小化に読み替え $E(w) = -\log P(D|w)$

⇒ 最急勾配法による最適化

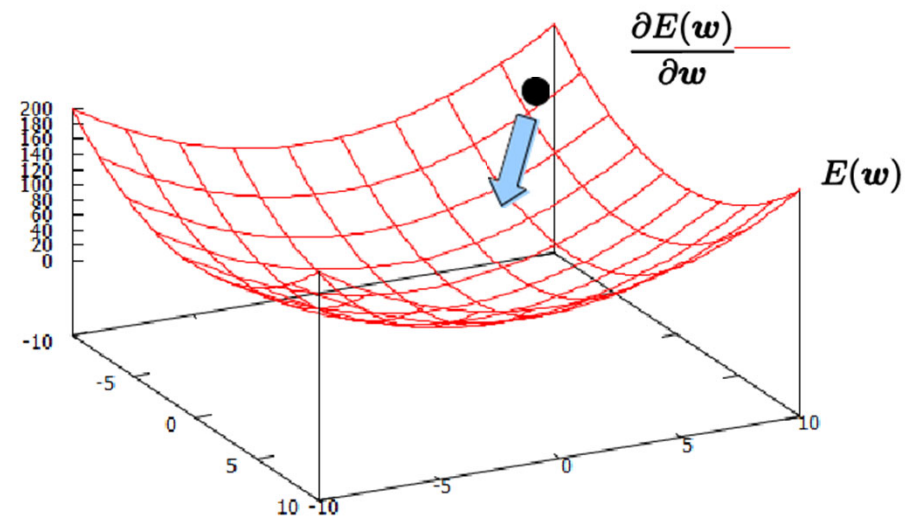
1. w の初期値を適当に設定

2. 以下の式で w の更新を
繰り返す

$$w \leftarrow w - \eta \frac{\partial E(w)}{\partial w}$$

η : 学習係数

3. w の変化量が一定以下になれば終了



ロジスティック識別の具体例

- Diabetesデータ

- 年齢・血圧・BMIなどから糖尿病検査結果を予測

No.	1: preg	2: plas	3: pres	4: skin	5: insu	6: mass	7: pedi	8: age	9: class
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	tested_positive
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	tested_negative
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	tested_positive
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	tested_negative
5	0.0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	tested_positive
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	tested_negative
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	tested_positive
...

パラメータ w

4.18 +
[preg] * -0.06 +
[plas] * -0.02 +
[pres] * 0.01 +
[insu] * 0 +
[mass] * -0.04 +
[pedi] * -0.47 +
[age] * -0.01

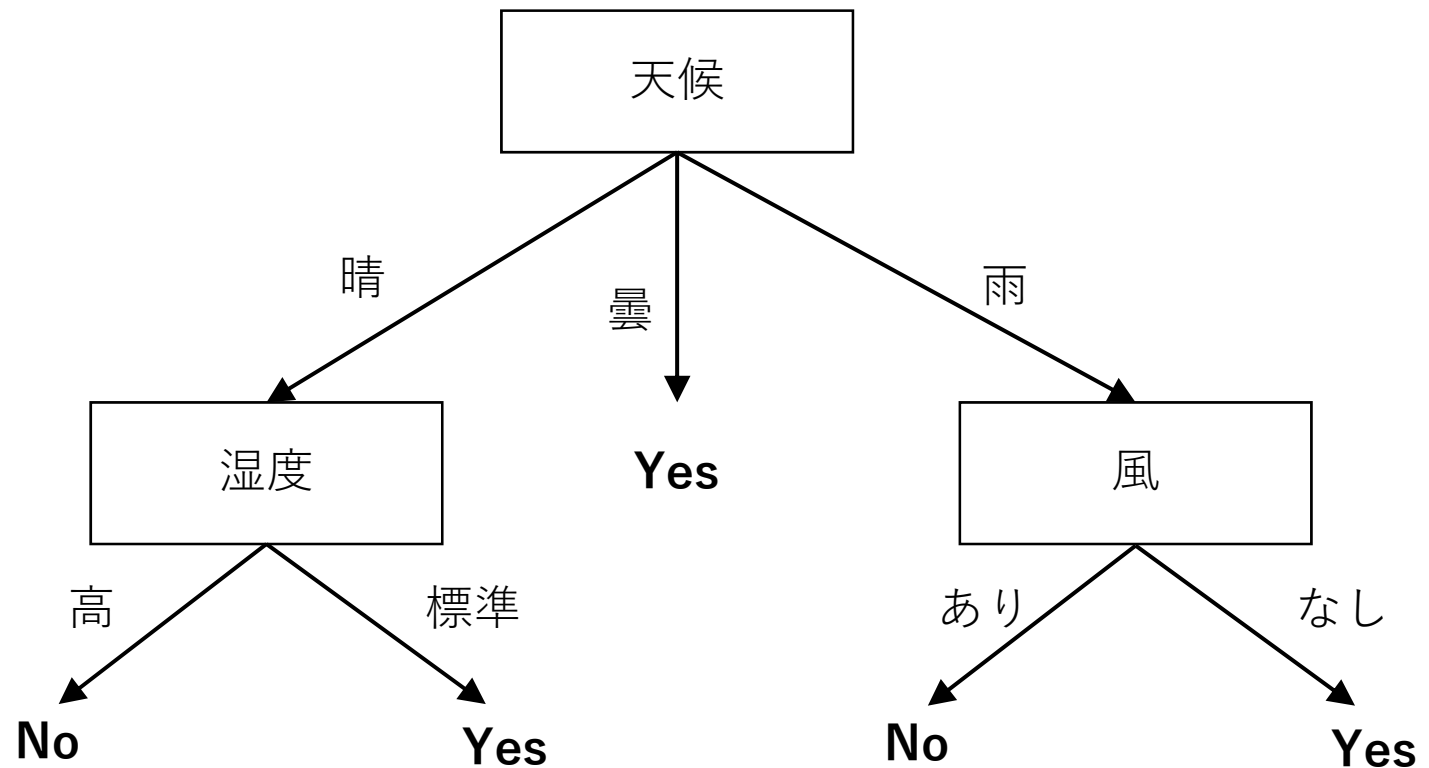
カテゴリ特徴に対する識別

ゴルフをする日のデータ

	天候	気温	湿度	風	play
1	晴	高	高	なし	no
2	晴	高	高	あり	no
3	曇	高	高	なし	yes
4	雨	中	高	なし	yes
5	雨	低	標準	なし	yes
6	雨	低	標準	あり	no
7	曇	低	標準	あり	yes
8	晴	中	高	なし	no
9	晴	低	標準	なし	yes
10	雨	中	標準	なし	yes
11	晴	中	標準	あり	yes
12	曇	中	高	あり	yes
13	曇	高	標準	なし	yes
14	雨	中	高	あり	no

決定木

- 決定木とは
 - 事例进行分类する質問を繰り返す



決定木

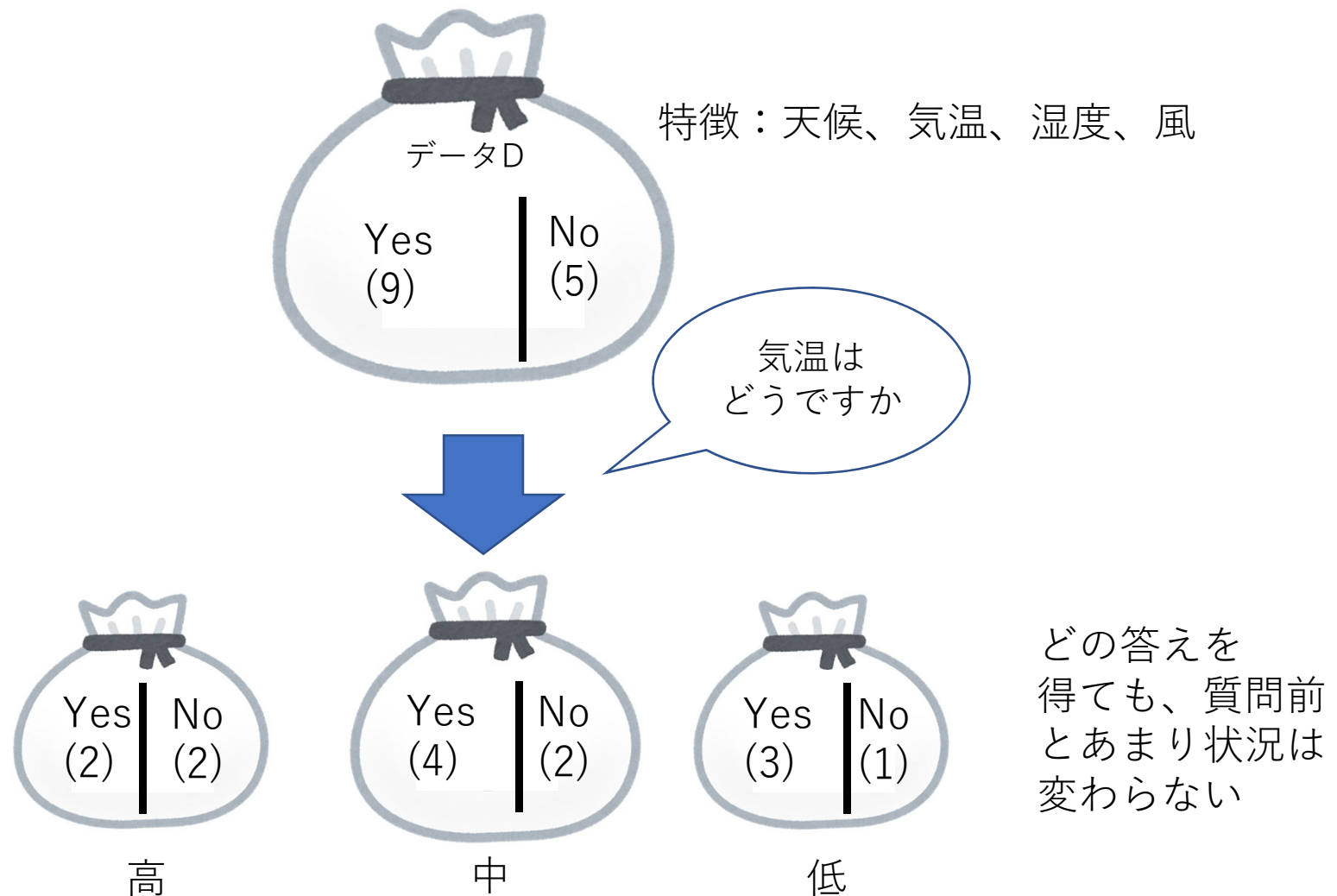
- 決定木の作り方
 - 大きな木を作れば（原理的には）データを100%正しく識別できる
 - 小さな木で多くのデータが正しく識別できれば、その木は未知のデータに対しても正しい識別を行う可能性が高い

決定木

- 小さな木の作り方
 - 分類能力の高い質問を、木の根に近いところに配置する

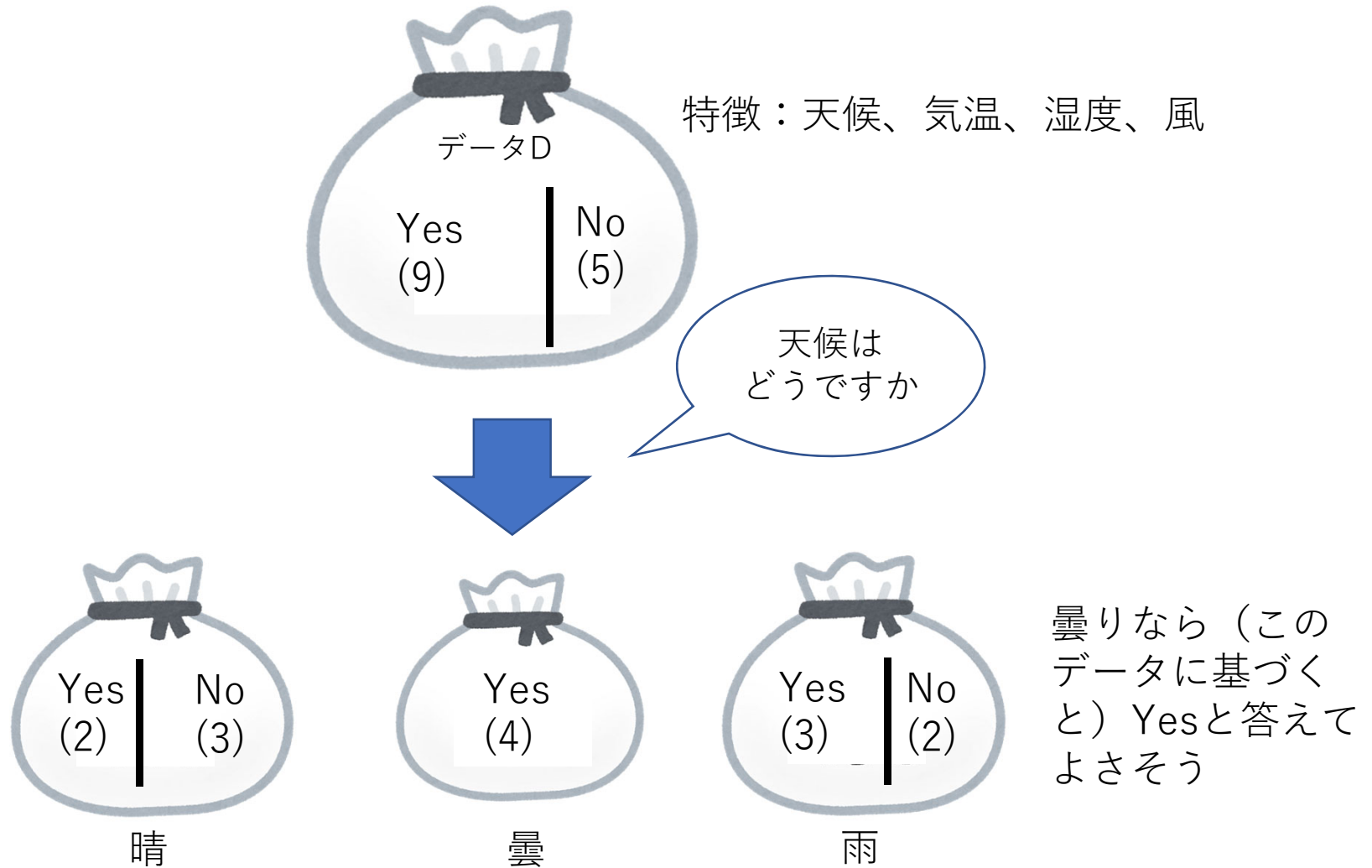
決定木

- 分類能力の低い質問



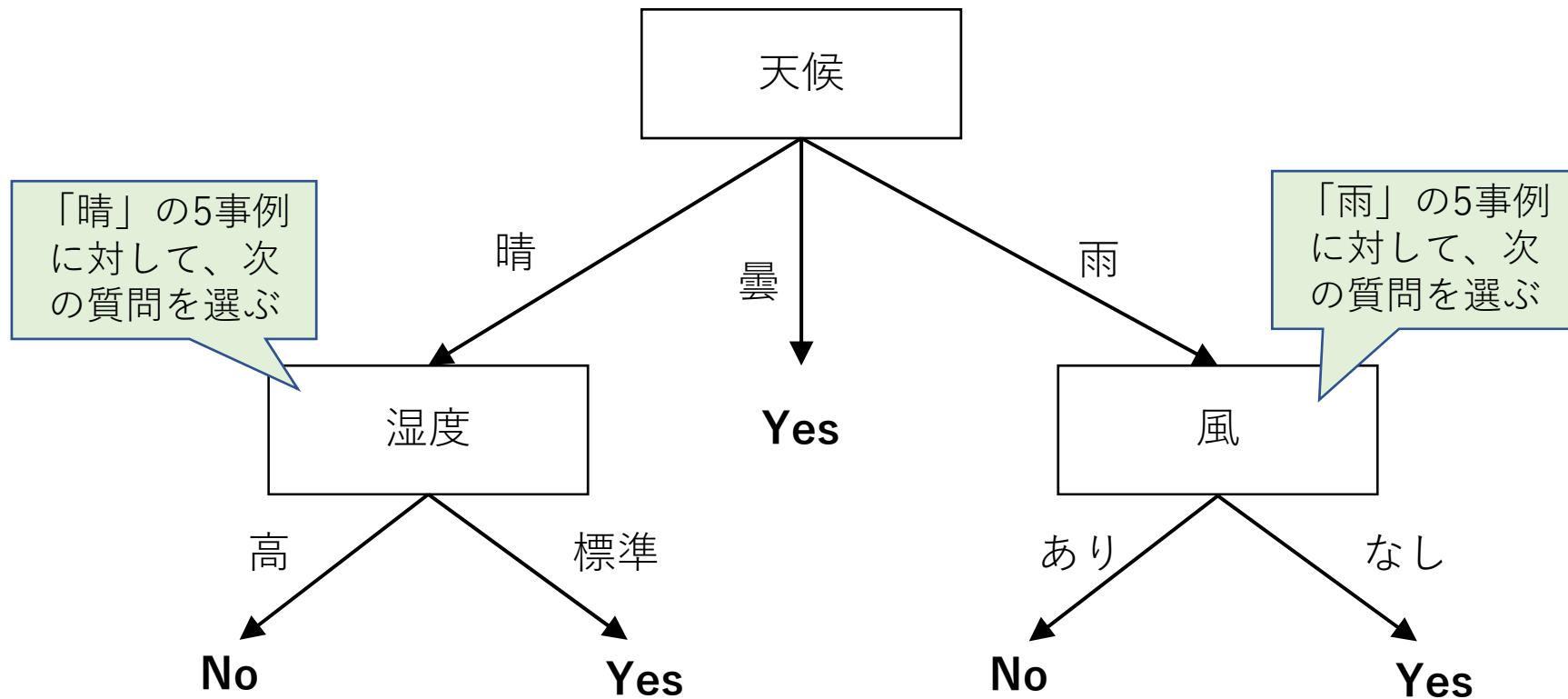
決定木

- 分類能力の高い質問



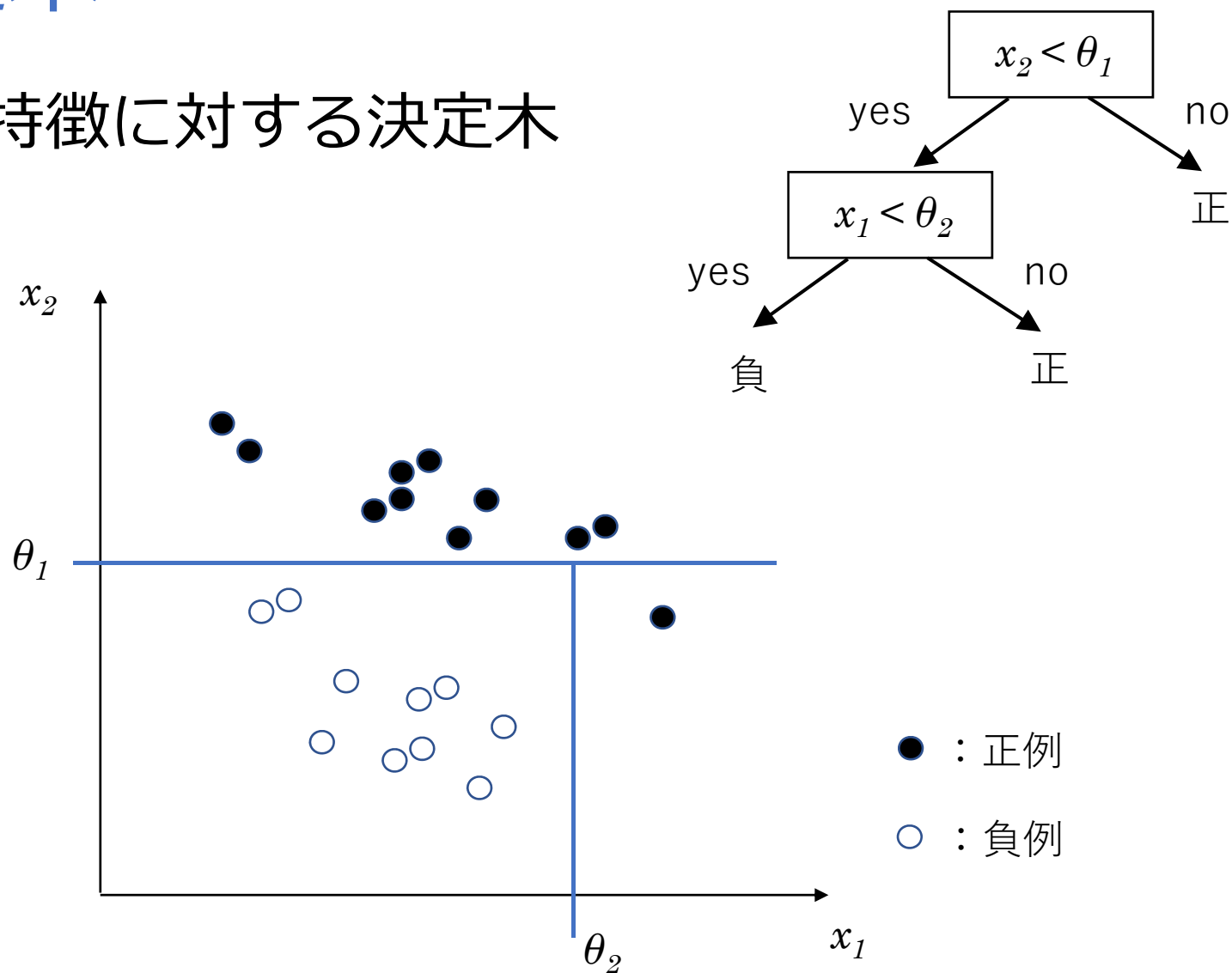
決定木

- 得られた決定木



決定木

- 数値特徴に対する決定木

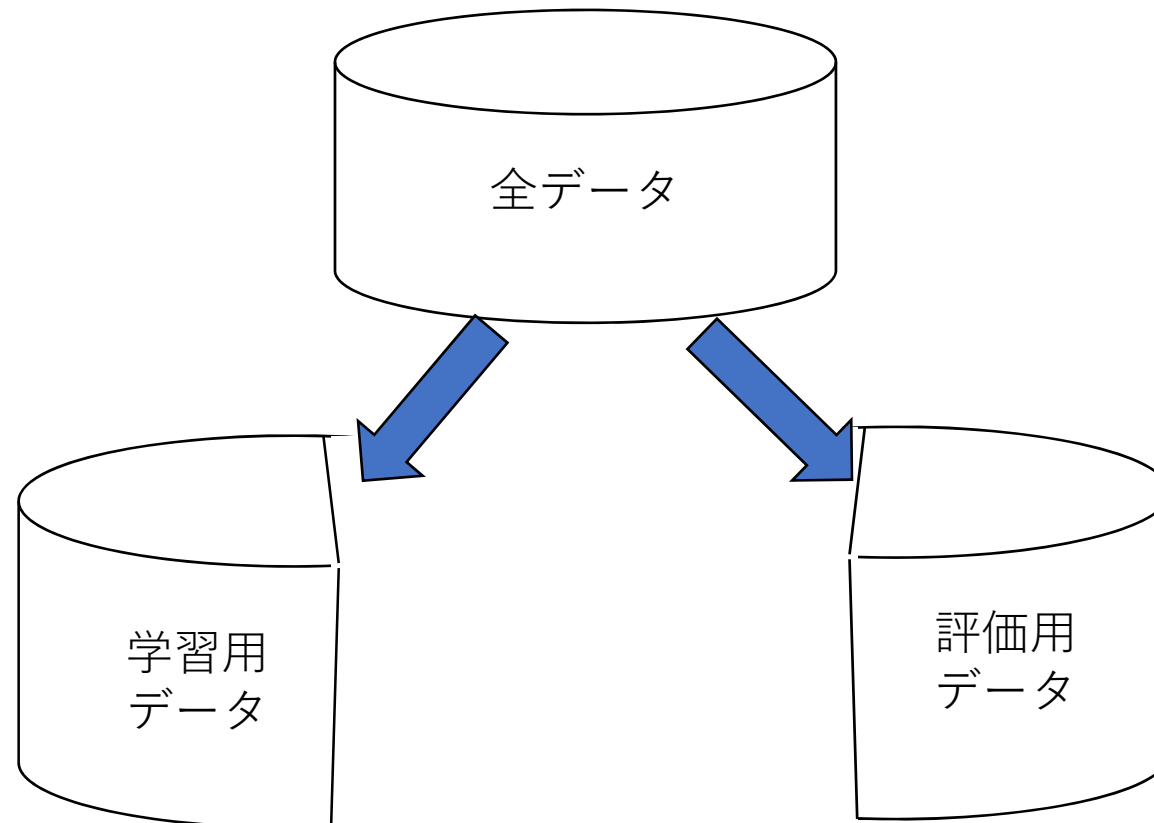


学習結果の評価（3章）

p.80 7コマ目

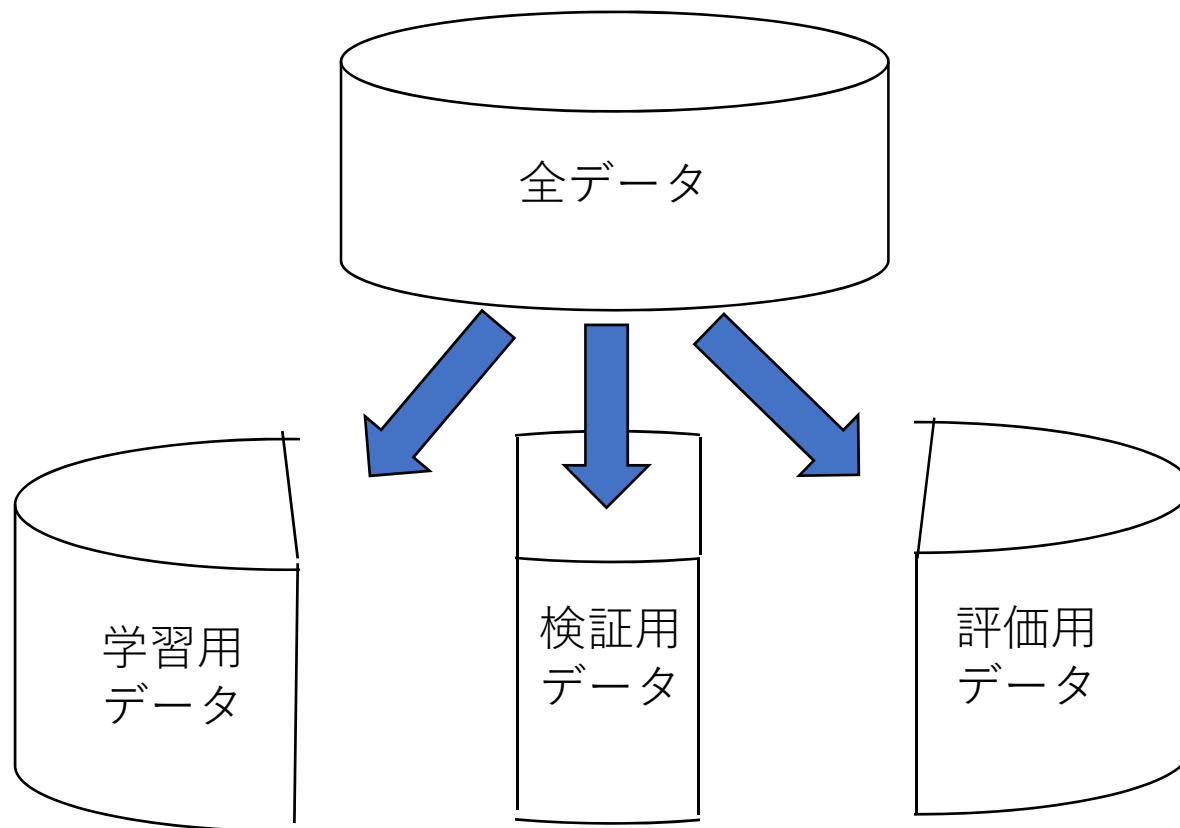
分割学習法

- 全データを学習用と評価用に分ける
 - データが多くあるときに有効



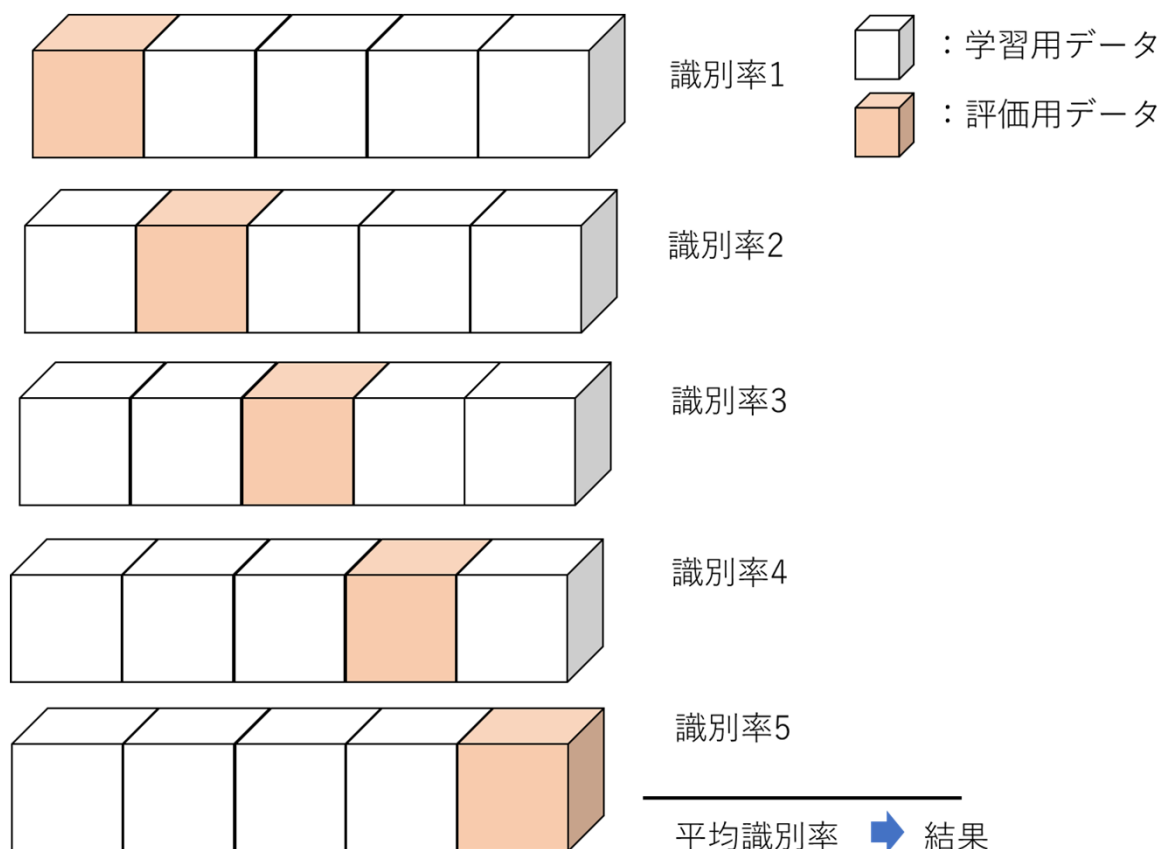
分割学習法

- パラメータチューニングを行うときは3分割
 - 検証用データでパラメータの良さを評価
 - 最終的な性能は評価用データで推測



交差確認法

- データをm分割して、m回の評価の平均をとる
 - 学習データが少ない場合に有効



評価指標

- 混同行列から算出

	予測 +	予測 -
正解 +	true positive (TP)	false negative (FN)
正解 -	false positive (FP)	true negative (TN)

識別器の出力

データに付いた正解

- 正解率

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

評価指標

- 目的に応じて適切な評価指標を選ぶ

	予測 +	予測 -
正解 +	TP	FN
正解 -	FP	TN

- 正解率 $Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$

- 精度 $Precision = \frac{TP}{TP + FP}$

- 再現率 $Recall = \frac{TP}{TP + FN}$

- F値 $F-measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

正解の割合
クラスの出現率に
偏りがある場合は不適

正例の判定が
正しい割合

正しく判定された
正例の割合

精度と再現率の
調和平均

↑
トレード
オフ
↓

識別の実用化事例

- オートマギ、NTTドコモ
 - 居眠り運転検知

<https://www.nikkei.com/article/DGXMZO38577940V01C18A2XY0000/>

- 国立国際医療研究センター
 - 糖尿病の発症リスク予測

<http://www.ncgm.go.jp/riskscore/>