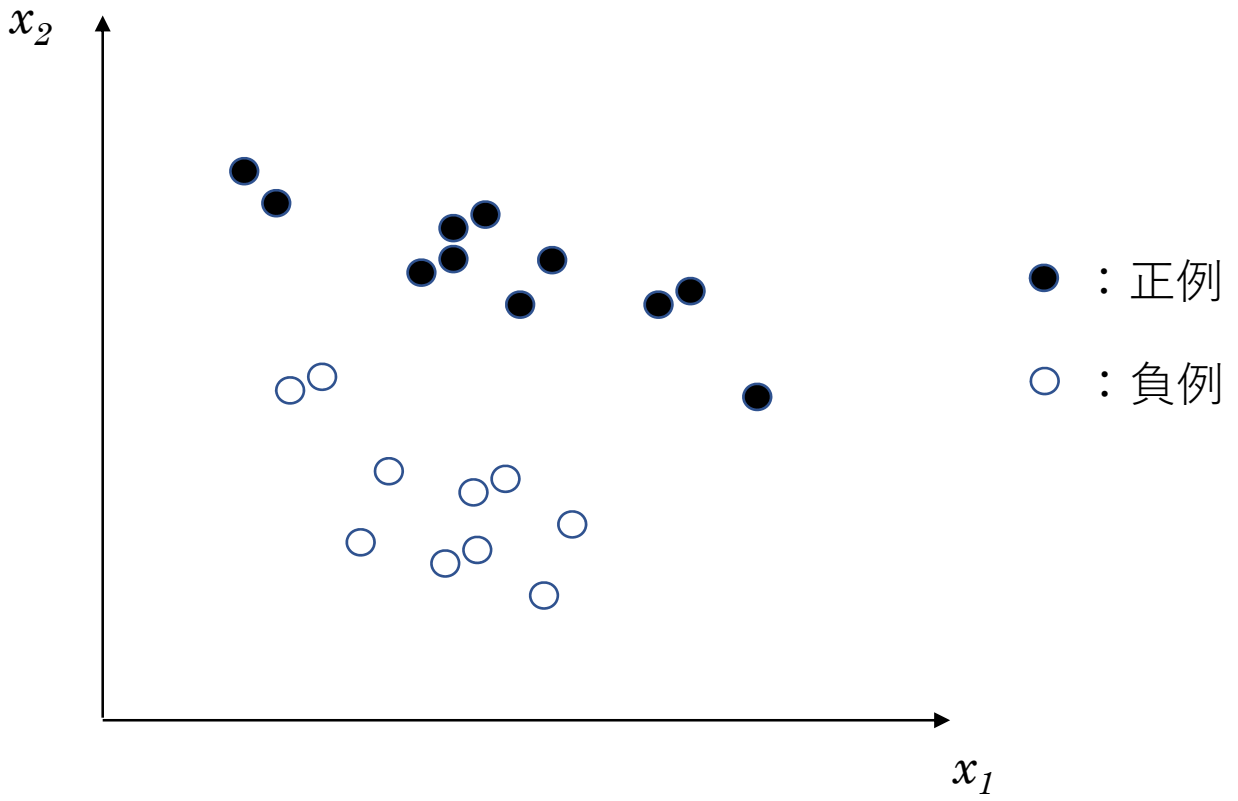
基礎的な識別（2章）

p.47　5コマ目

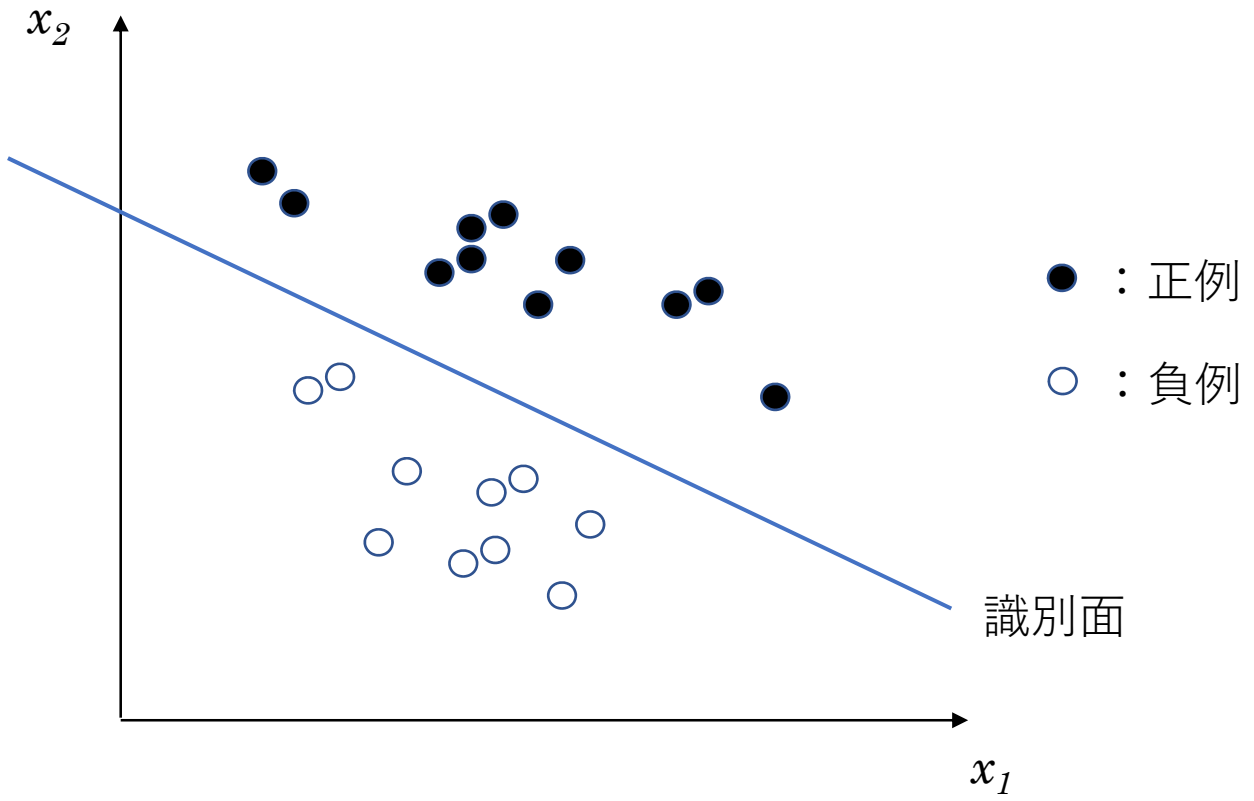
識別

- 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



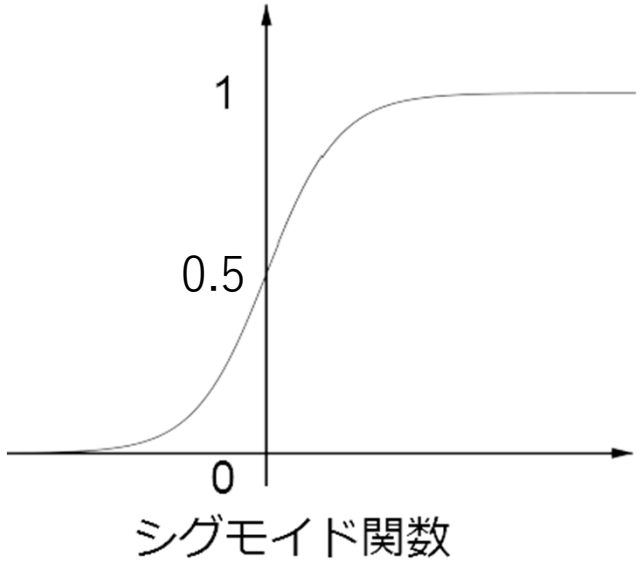
ロジスティック識別

- 識別面の式

$$\hat{g}(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots + w_d x_d + w_0 = \boldsymbol{w}^T \boldsymbol{x} = 0$$

  - 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 ⇒ シグモイド関数

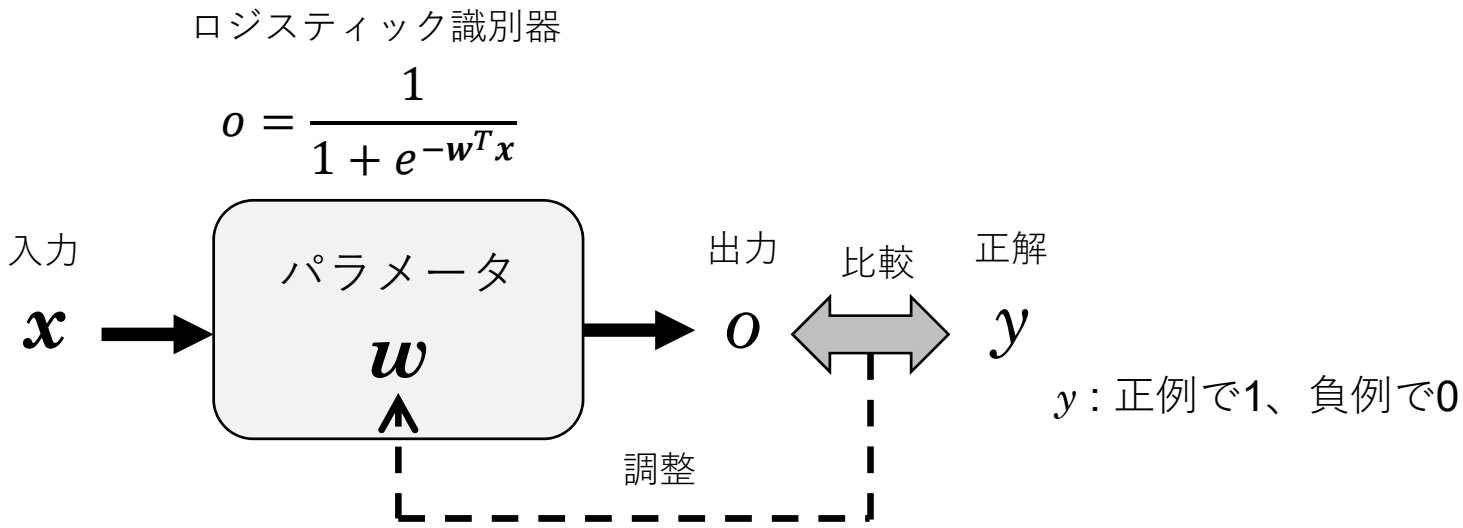
$$P(\text{正}|\boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{w}^T \boldsymbol{x})}$$



シグモイド関数

ロジスティック識別

- 係数 $\boldsymbol{w}$ の求め方



y : 正例で1、負例で0

- 尤度（モデルのもっともらしさ）の定義

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

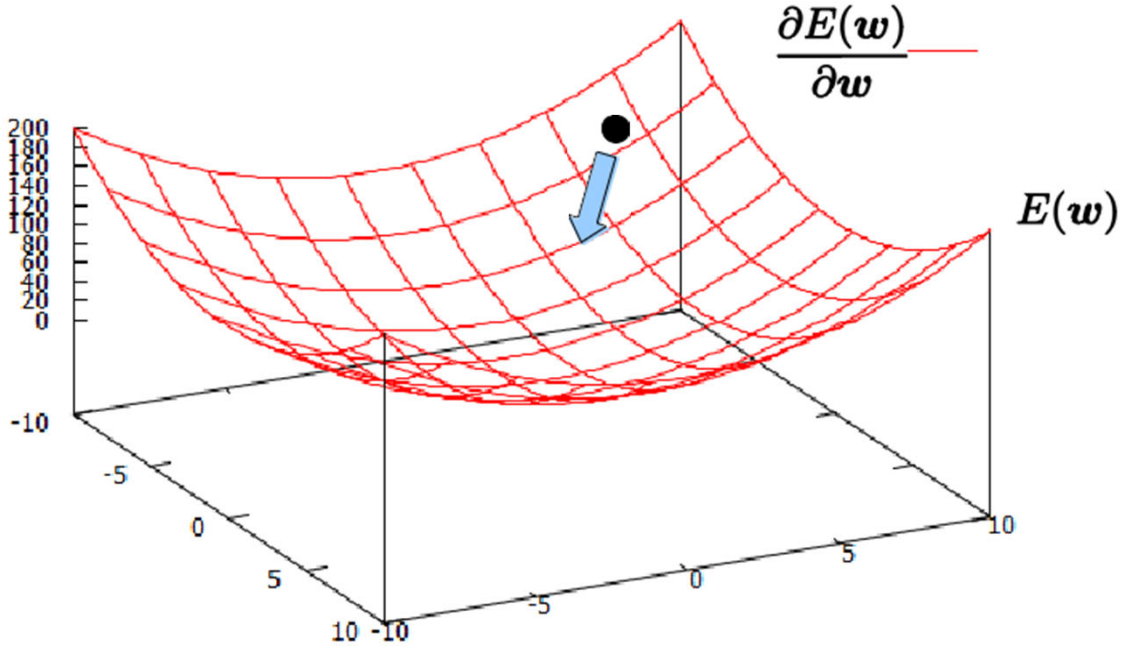
D : 全データ

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 $E(\boldsymbol{w}) = -\log P(D|\boldsymbol{w})$

⇒ 最急勾配法による最適化 $\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \dfrac{\partial E(\boldsymbol{w})}{\partial \boldsymbol{w}}$



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

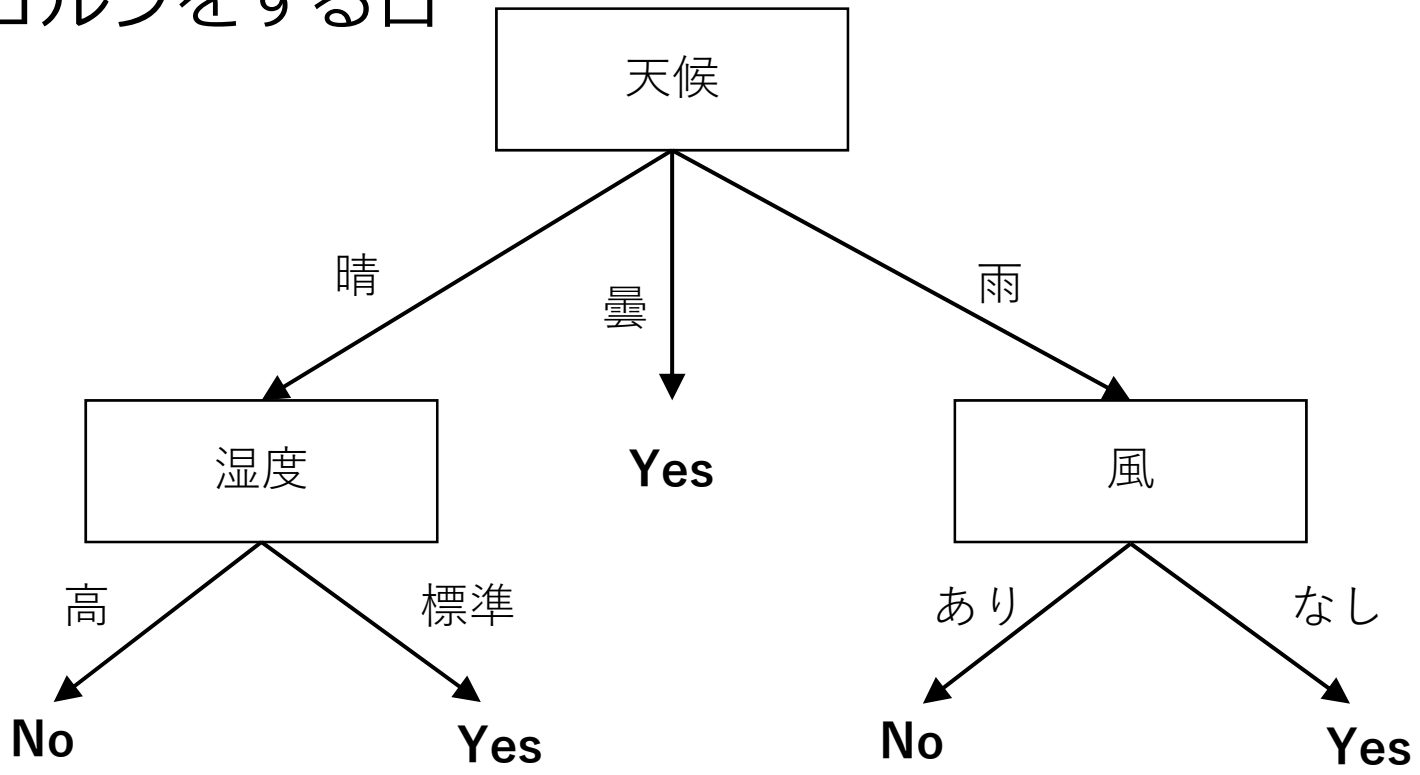
No.	1: preg Numeric	2: plas Numeric	3: pres Numeric	4: skin Numeric	5: insu Numeric	6: mass Numeric	7: pedi Numeric	8: age Numeric	9: **class** Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	tested_positive
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	tested_negative
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	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

予測式

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

決定木

- 決定木とは
  - 事例を分類する質問を繰り返す
  - 例）ゴルフをする日



決定木

学習データの例（カテゴリ特徴）

	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

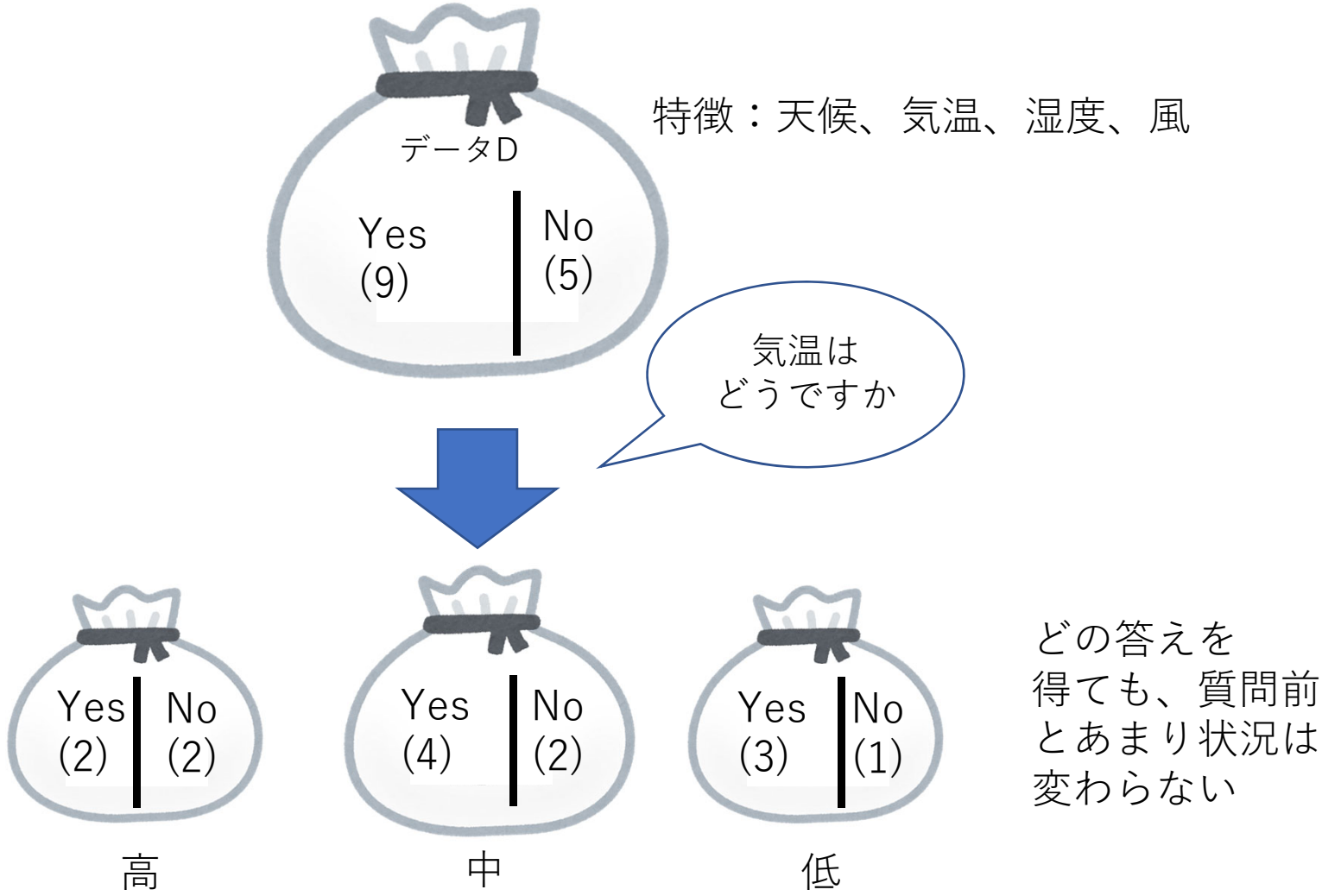
- 決定木の作り方
  - 大きな木を作れば（原理的には）データを100%正しく識別できる
  - 小さな木で多くのデータが正しく識別できれば、その木は未知のデータに対しても正しい識別を行う可能性が高い

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

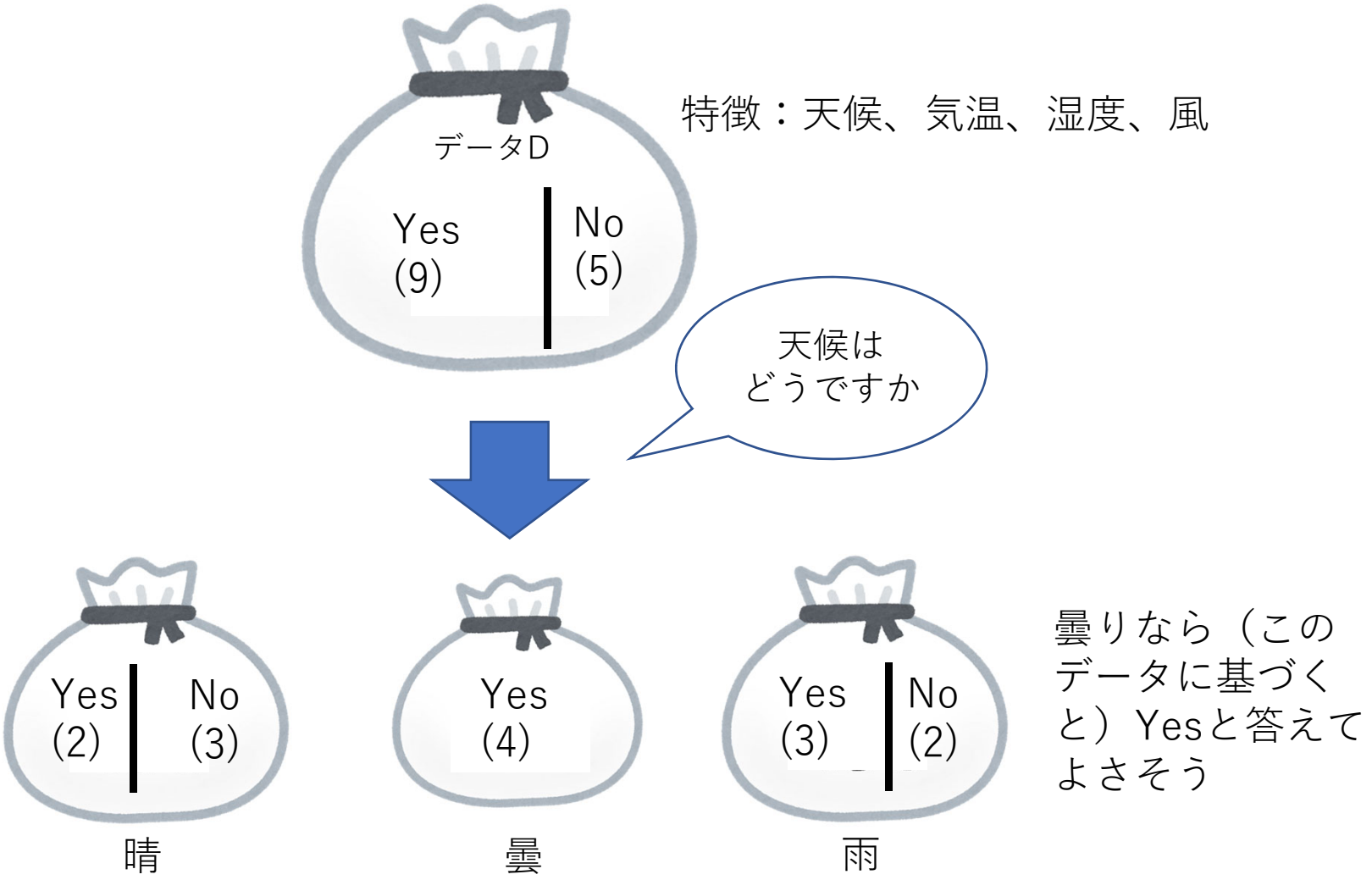
決定木

- 分類能力の低い質問



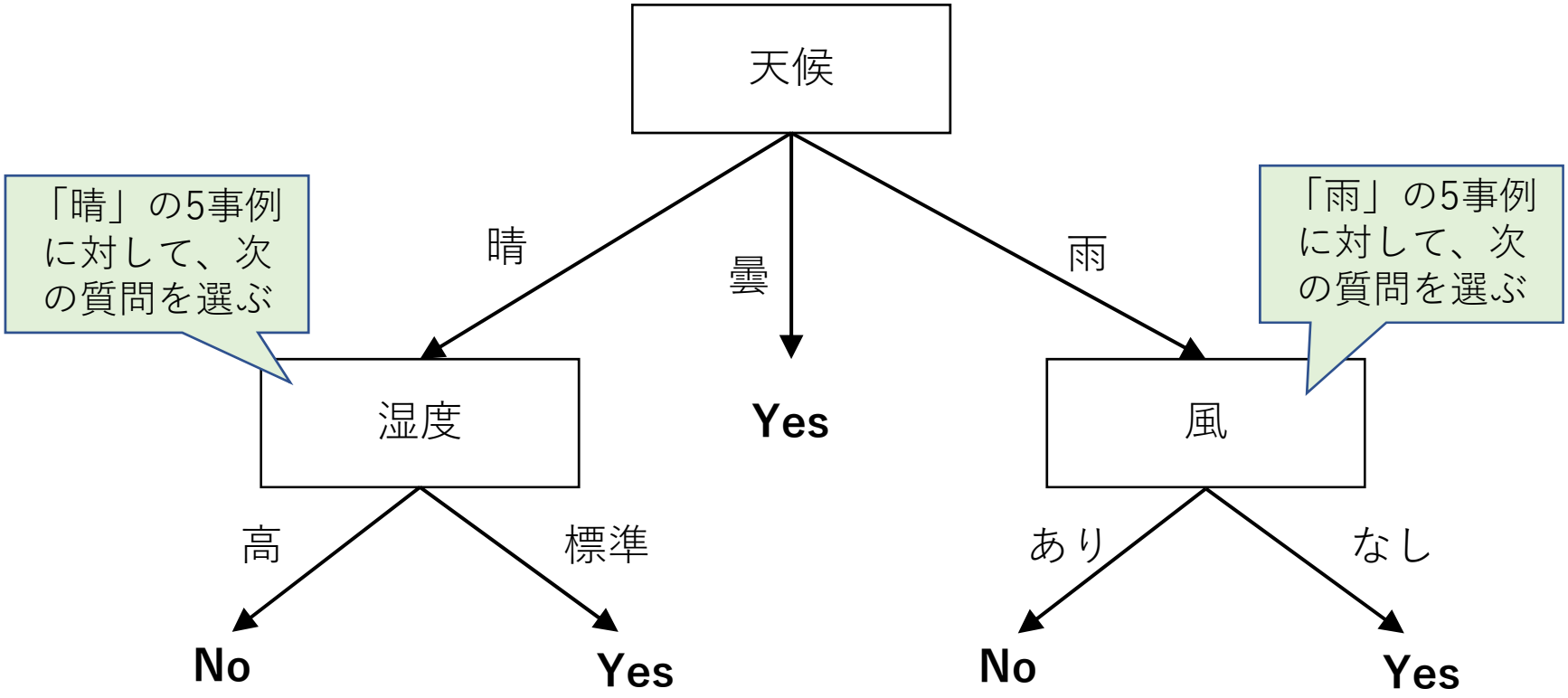
決定木

- 分類能力の高い質問



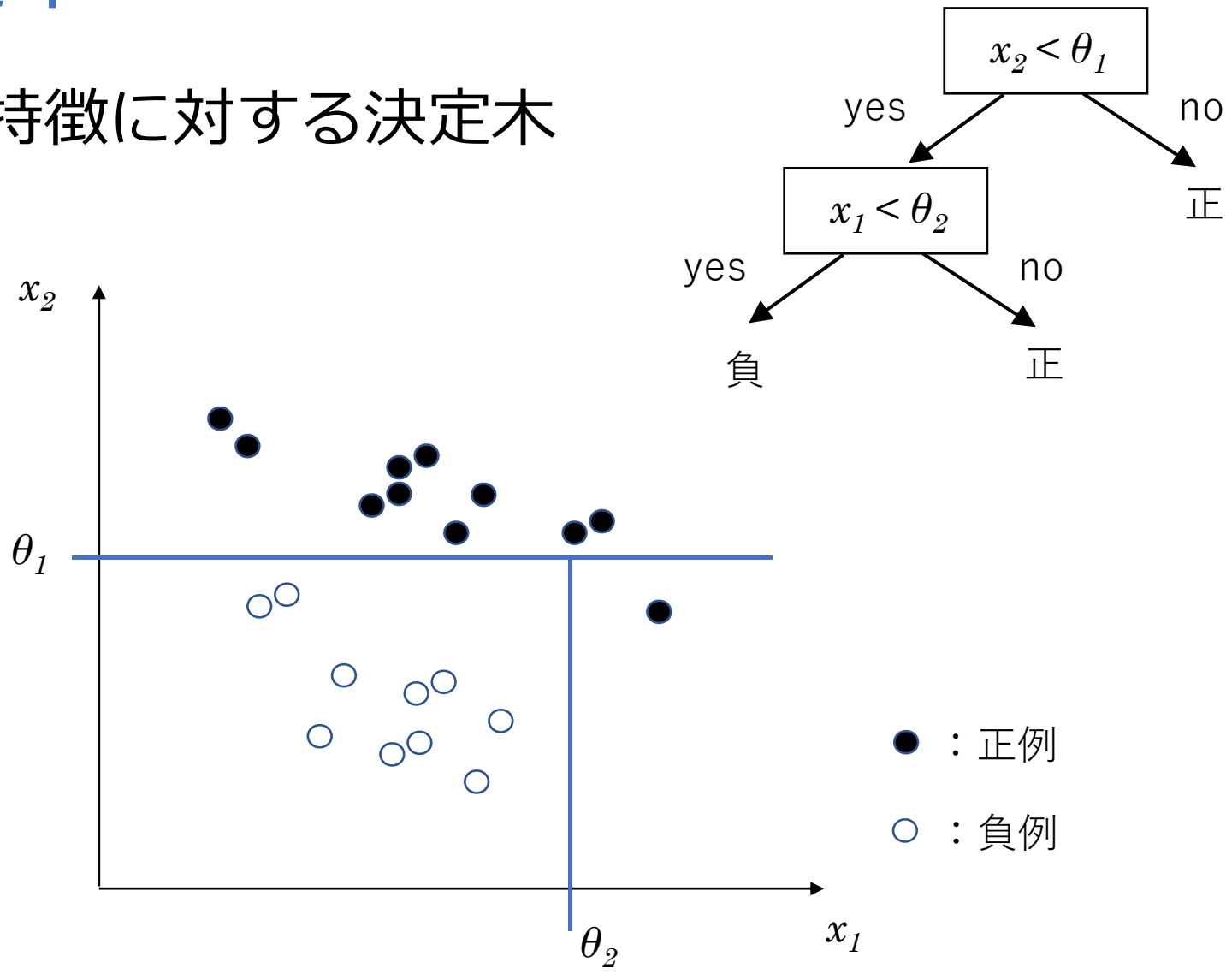
決定木

- 得られた決定木


天候
晴
曇
雨
「晴」の5事例に対して、次の質問を選ぶ
「雨」の5事例に対して、次の質問を選ぶ
湿度
Yes
風
高
標準
あり
なし
No
Yes
No
Yes

決定木

- 数値特徴に対する決定木

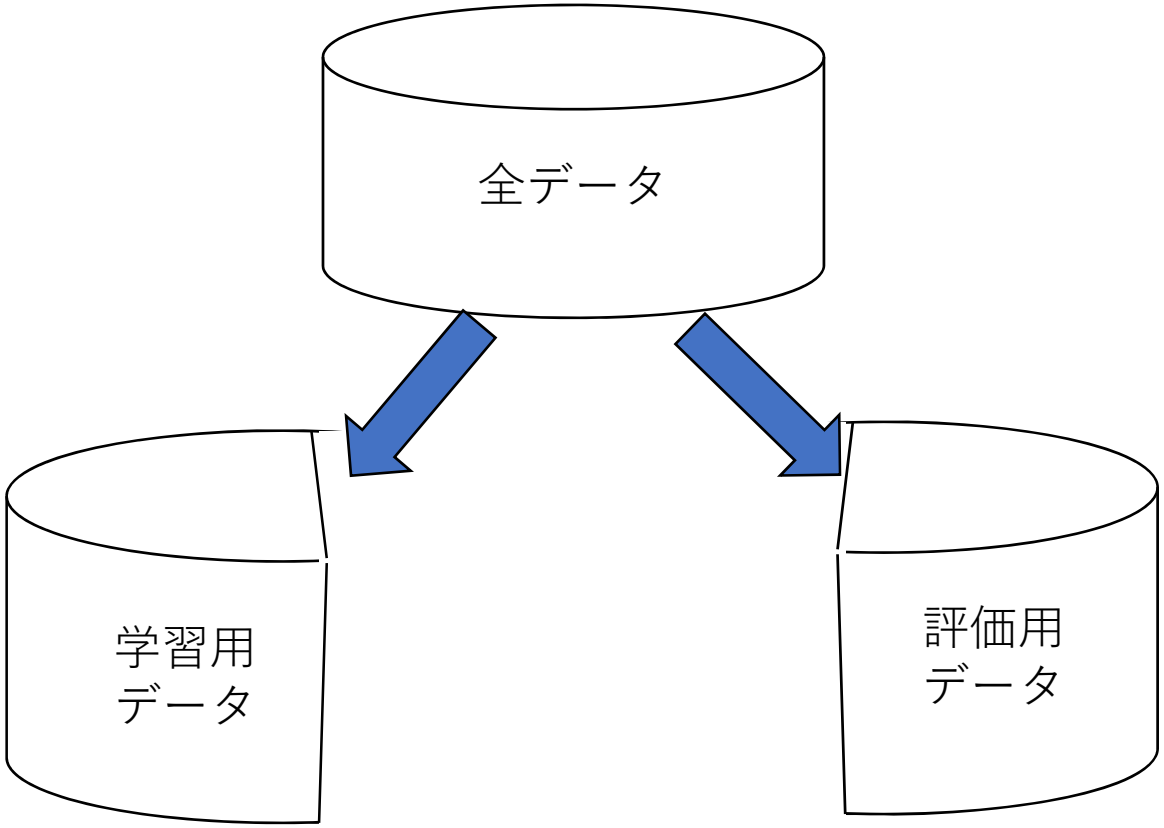


学習結果の評価（3章）

p.80　7コマ目

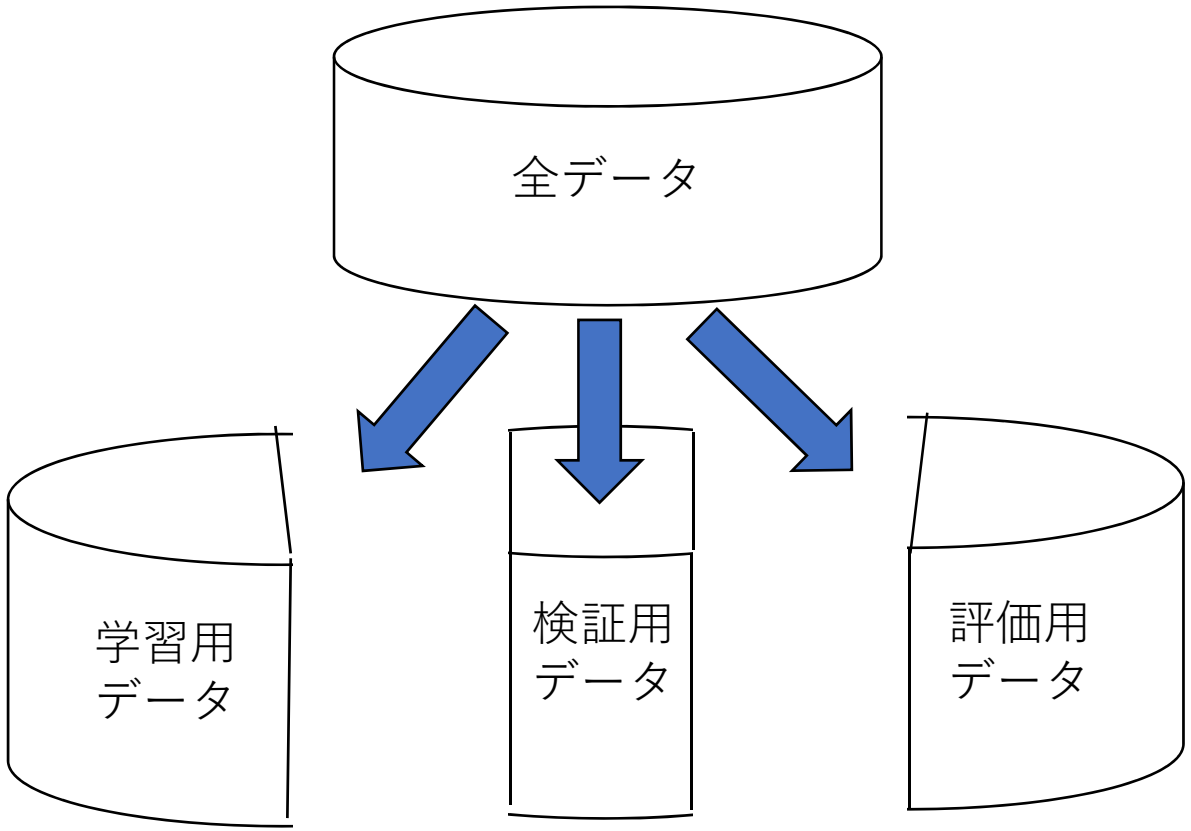
分割学習法

- 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



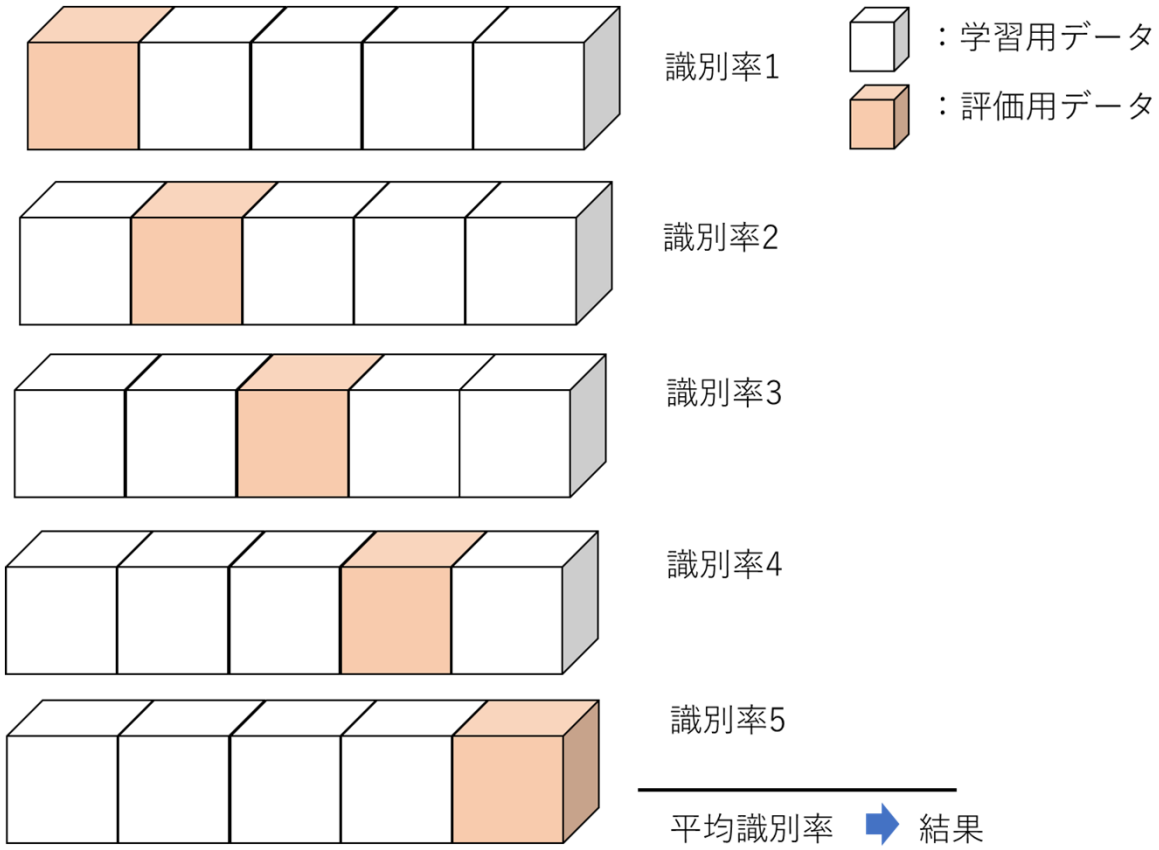
分割学習法

- 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

識別器の出力

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

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

正解の割合
クラスの出現率に
偏りがある場合は不適

- 精度 $Precision = \frac{TP}{TP + FP}$

正例の判定が
正しい割合

- 再現率 $Recall = \frac{TP}{TP + FN}$

正しく判定された
正例の割合

トレードオフ

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

精度と再現率の
調和平均

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/