2章のストーリー

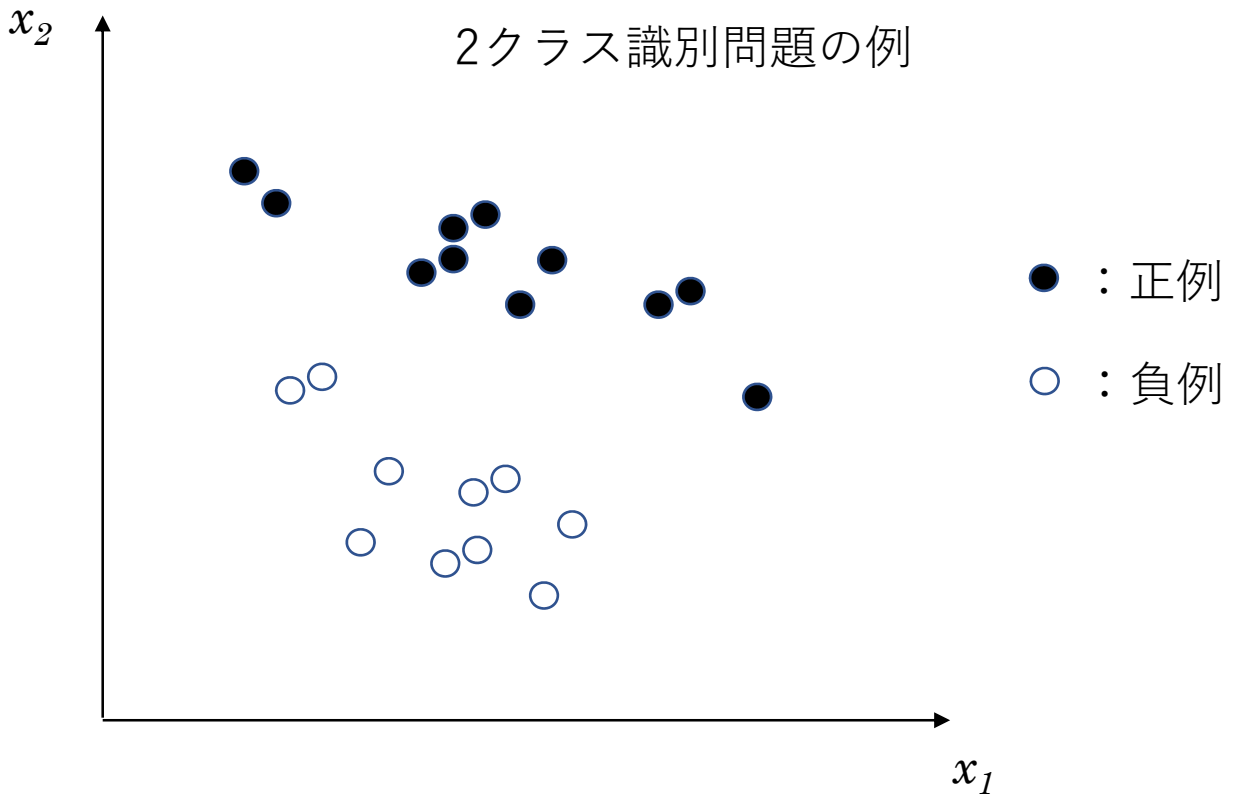
- 清原は、市の医療費削減のために健診結果から糖尿病の発病を予測するサービスを立ち上げたいと考える
- さやかは識別問題の解法として、ロジスティック識別と決定木について教える

基礎的な識別（2章）

p.47　5コマ目

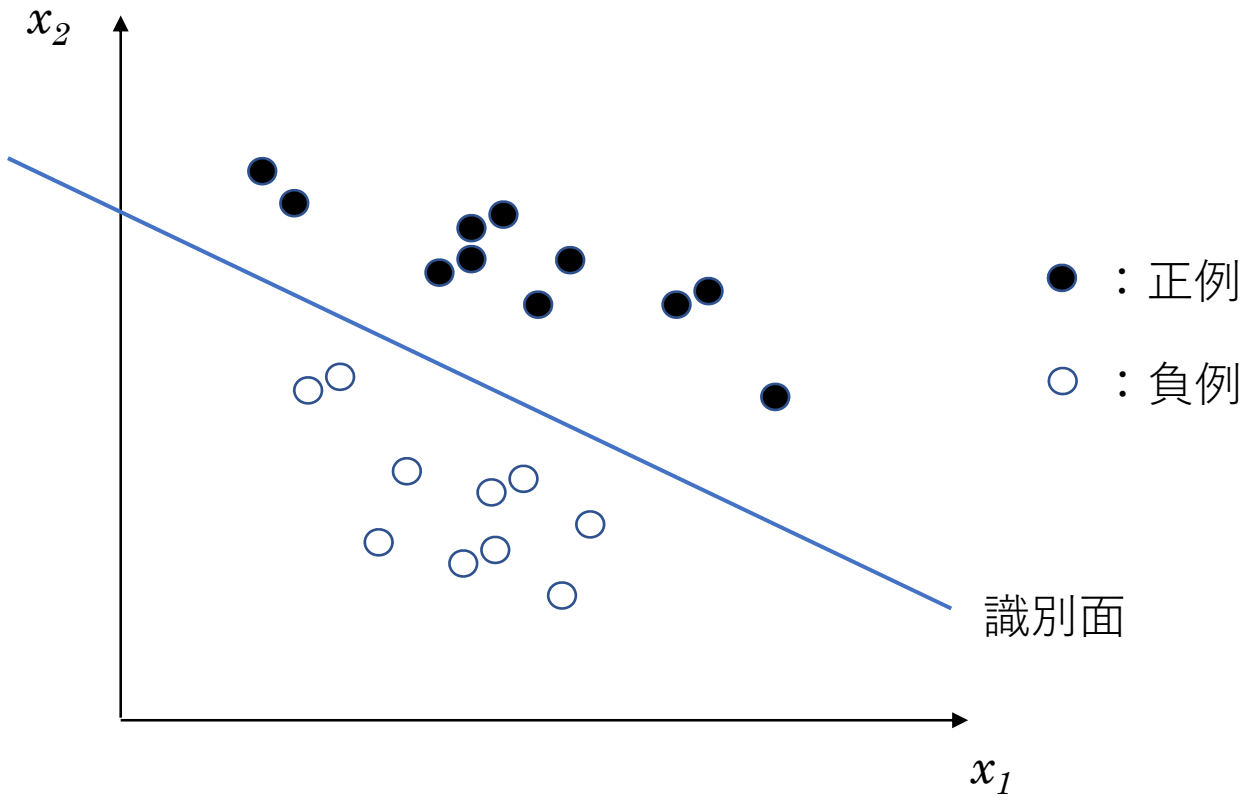
識別

- 識別とは
  - 教師あり学習のひとつ
  - 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クラス識別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

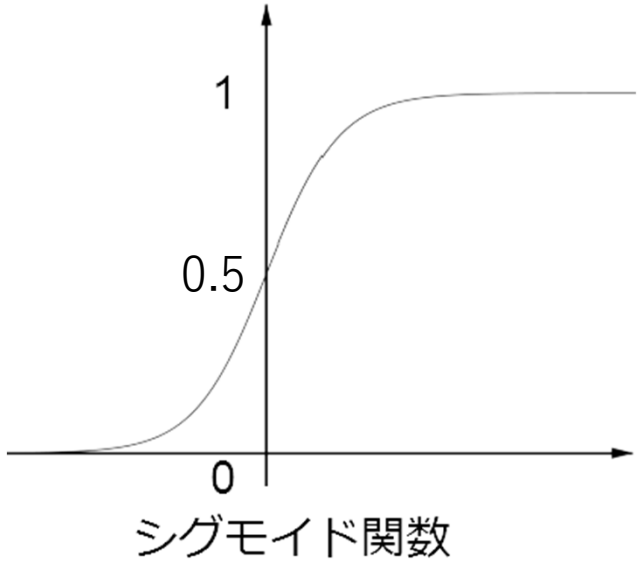
- 識別面の式

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

  - 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

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

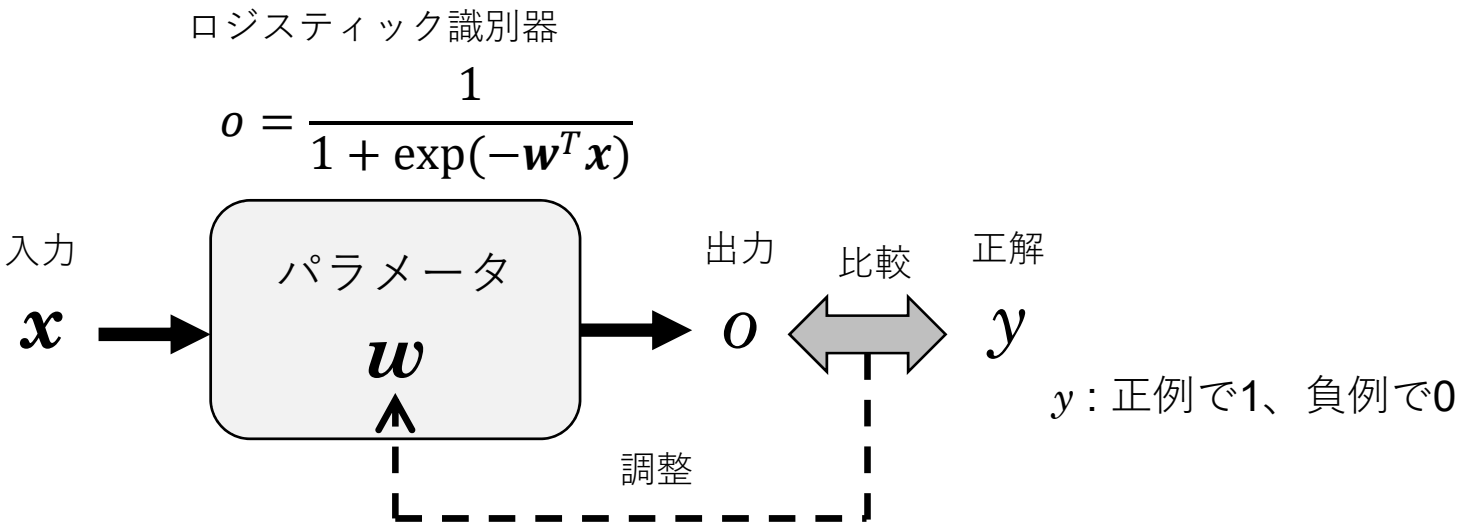
$\exp(x)$: e^x
e=2.71828...



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）が最大となるよう調整

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

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

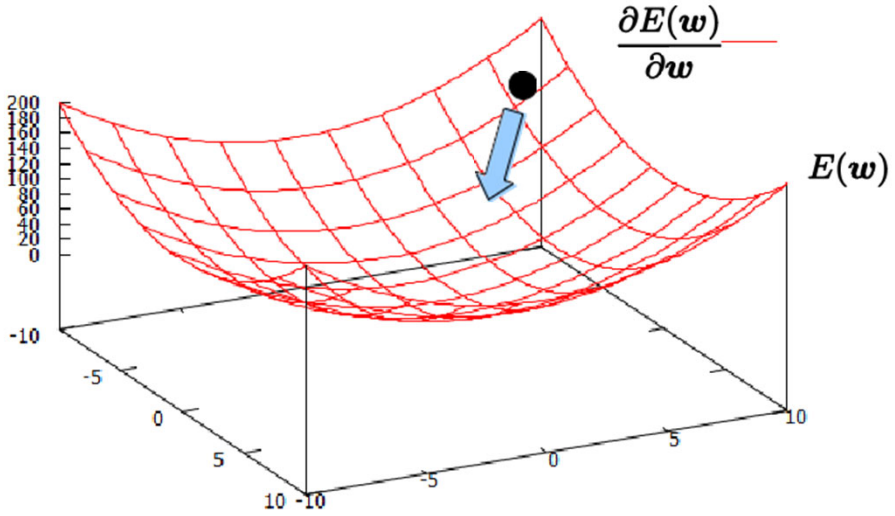
⇒ 最急勾配法による最適化

1. $\boldsymbol{w}$の初期値を適当に設定
2. 以下の式で$\boldsymbol{w}$の更新を繰り返す

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

η：学習係数

3. $\boldsymbol{w}$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 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

妊娠回数　血糖値　血圧　皮下脂肪　インスリン　BMI 家系 年齢　検査結果

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
```

係数$\boldsymbol{w}$

ロジスティック識別の具体例

- データの標準化
  - 各特徴のスケールを平均0、分散1に揃える
  - 特徴が結果に寄与する度合いが係数の大きさでわかる

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	0.639...	0.847...	0.149...	0.906...	-0.692...	0.203...	0.468...	1.42...	tested_positive
2	-0.84...	-1.12...	-0.16...	0.530...	-0.692...	-0.683...	-0.36...	-0.19...	tested_negative
3	1.233...	1.942...	-0.26...	-1.28...	-0.692...	-1.102...	0.604...	-0.10...	tested_positive
4	-0.84...	-0.99...	-0.16...	0.154...	0.123...	-0.493...	-0.92...	-1.04...	tested_negative
5	-1.14...	0.503...	-1.50...	0.906...	0.765...	1.408...	5.481...	-0.02...	tested_positive
6	0.342...	-0.15...	0.252...	-1.28...	-0.692...	-0.810...	-0.81...	-0.27...	tested_negative

妊娠回数　血糖値　血圧　皮下脂肪　インスリン　BMI　家系　年齢　検査結果

```
-0.43 +
[preg] *  0.2  +
[plas] *  0.56 +
[pres] * -0.13 +
[insu] * -0.07 +
[mass] *  0.35 +
[pedi] *  0.16 +
[age] *   0.09
```

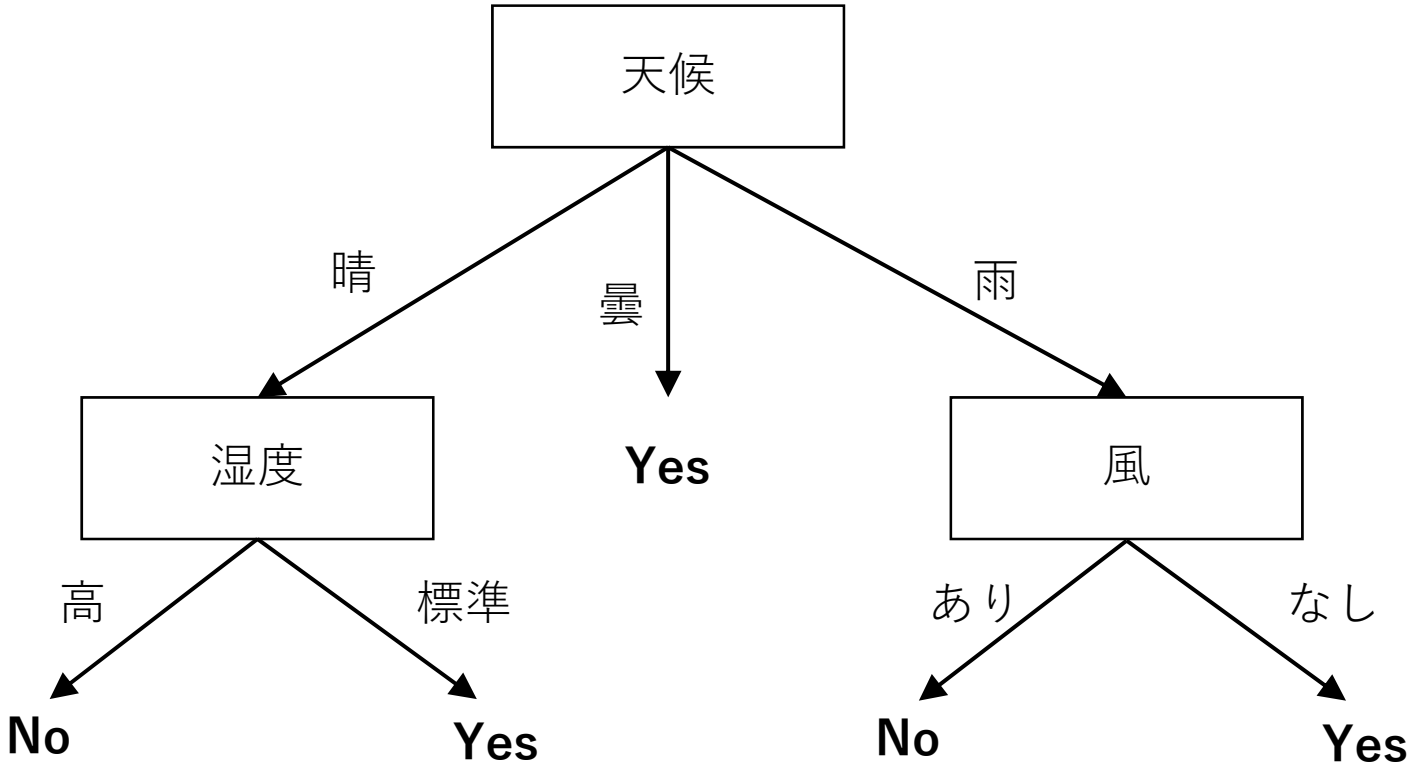
カテゴリ特徴に対する識別

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
  - 事例を分類する質問を繰り返す



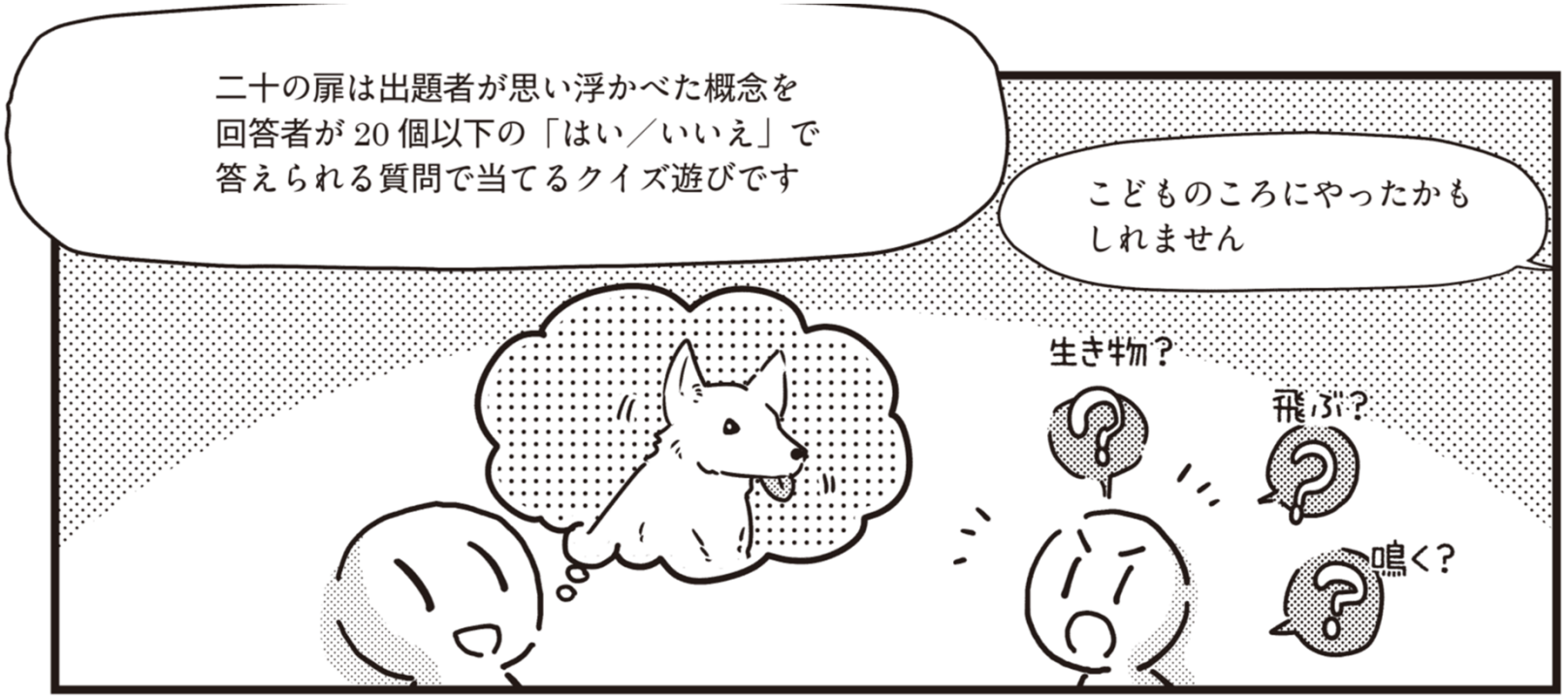
決定木

- 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

p.65　2コマ目

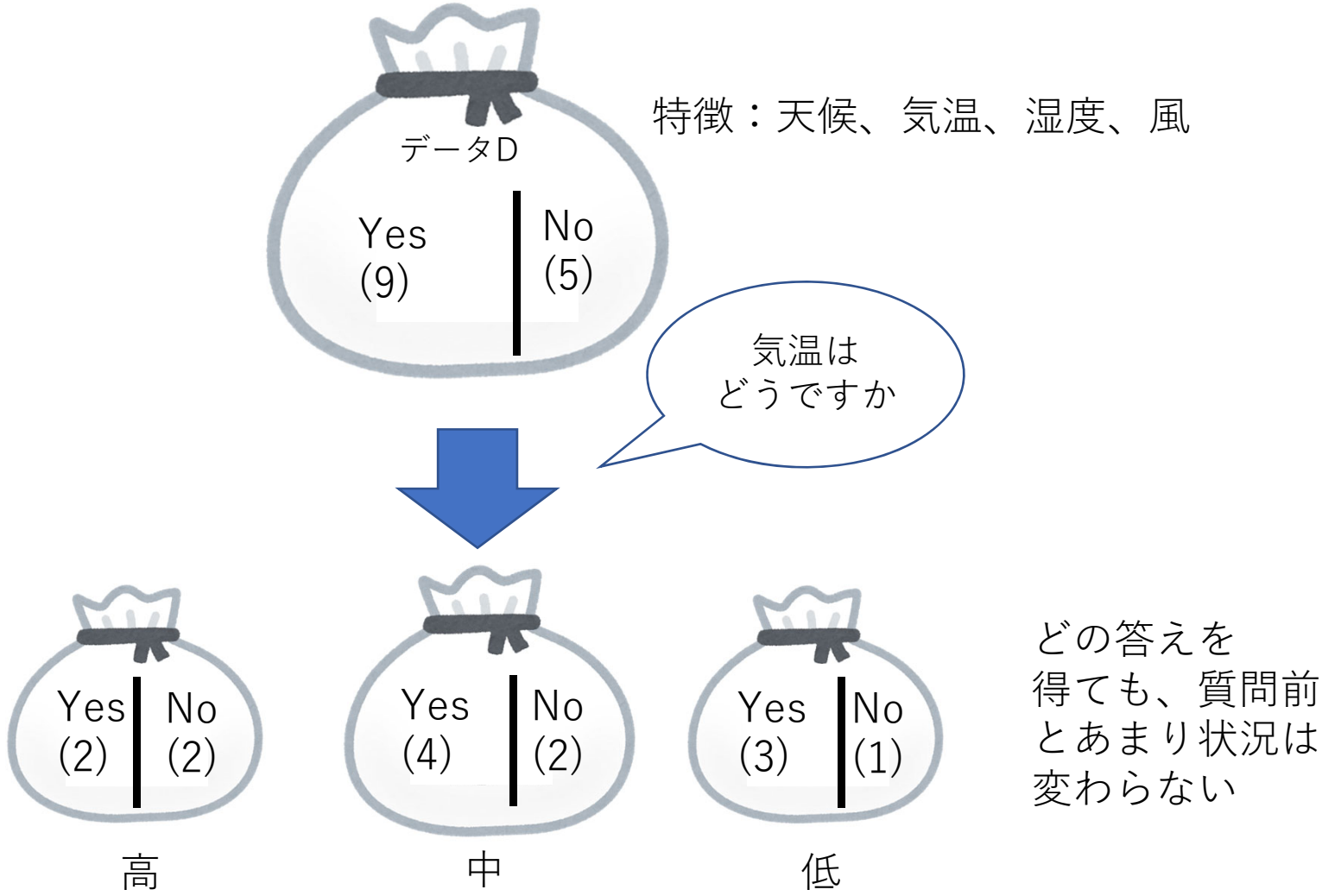
決定木

- 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



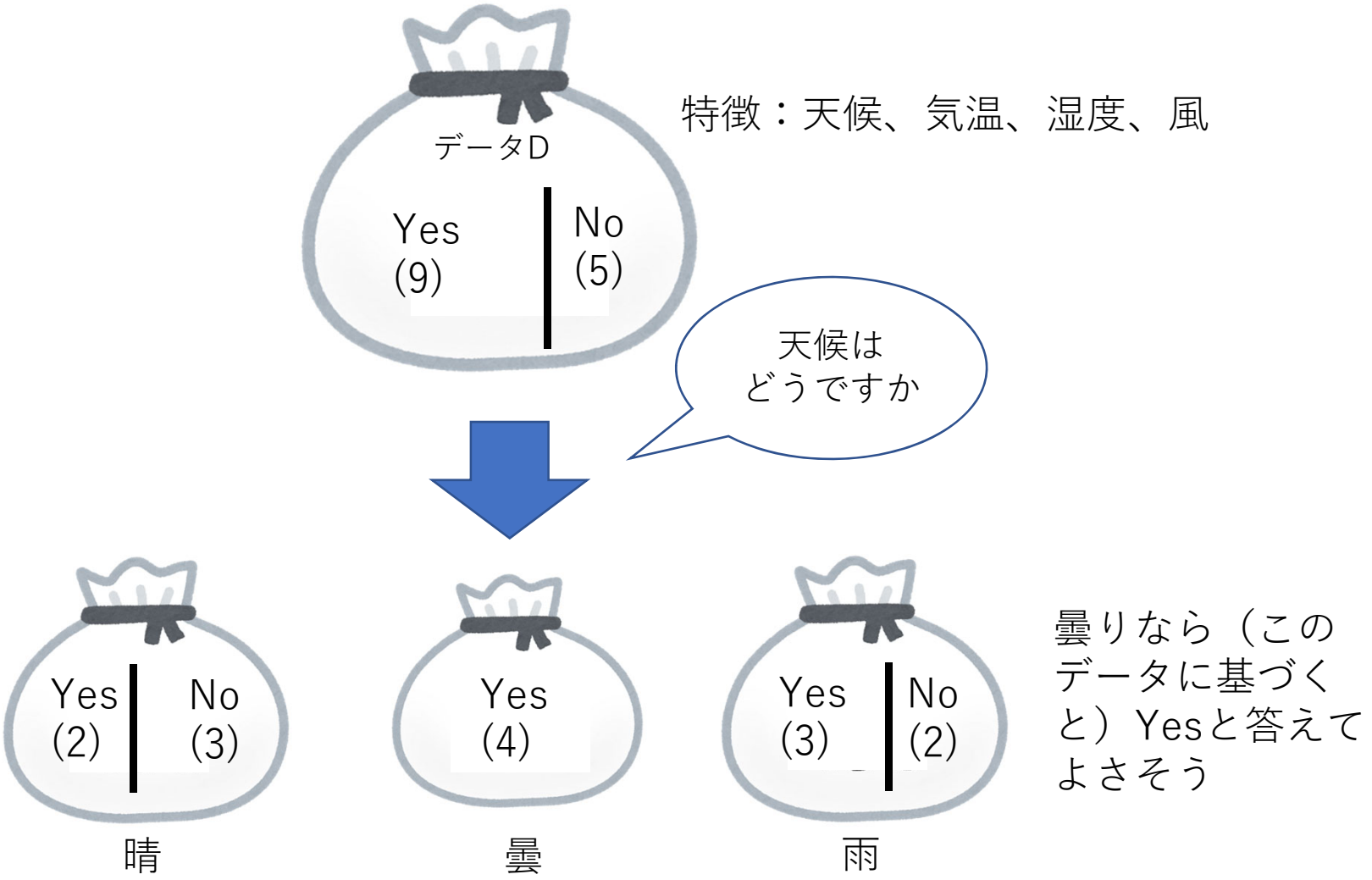
決定木

- 分類能力の低い質問



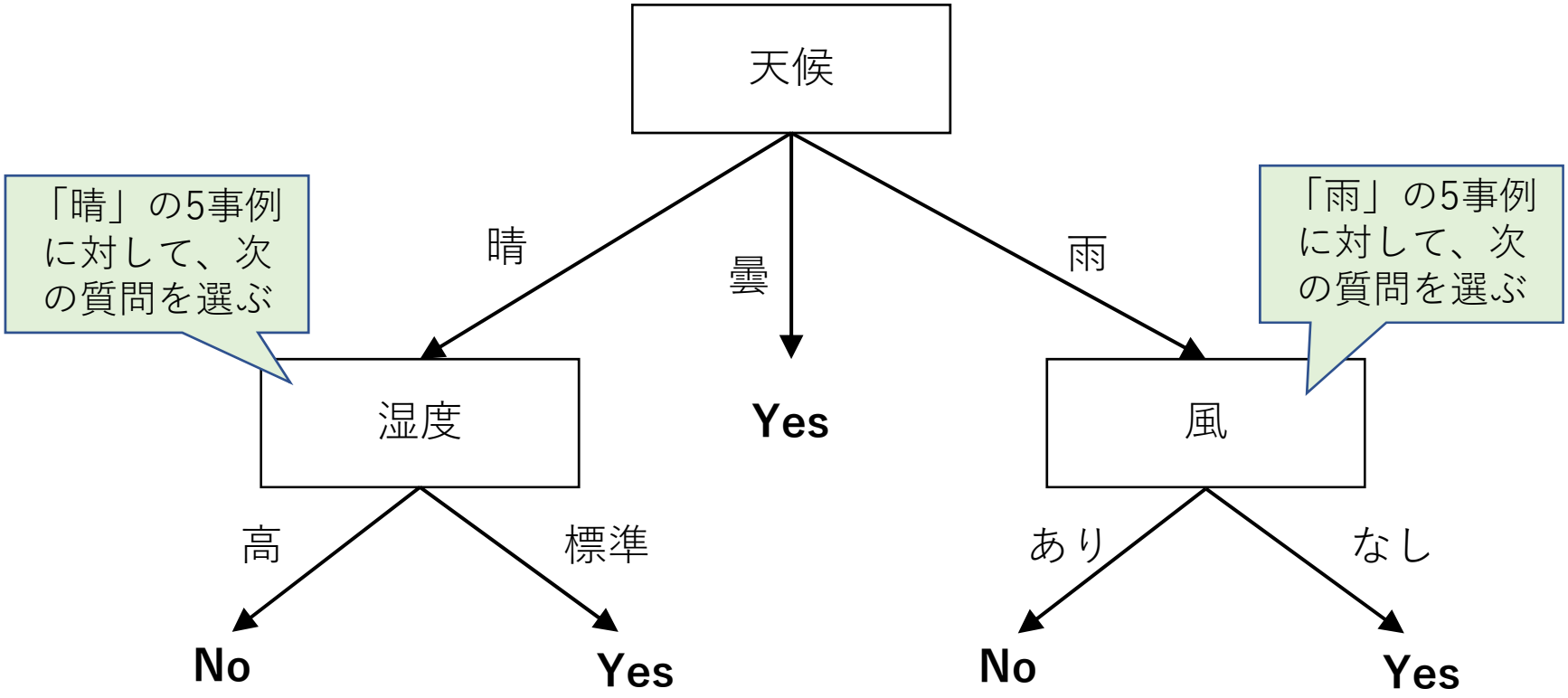
決定木

- 分類能力の高い質問



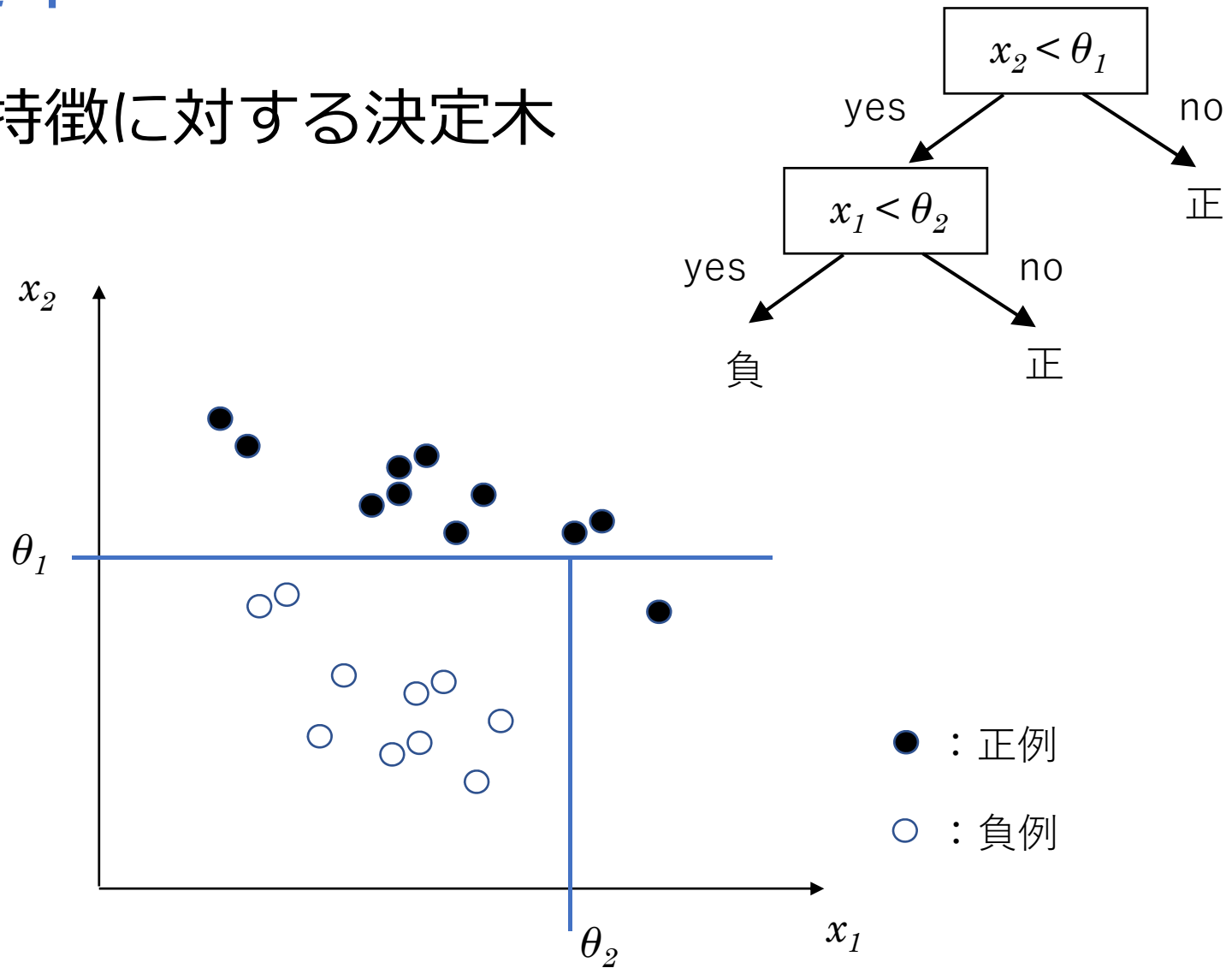
決定木

- 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



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/