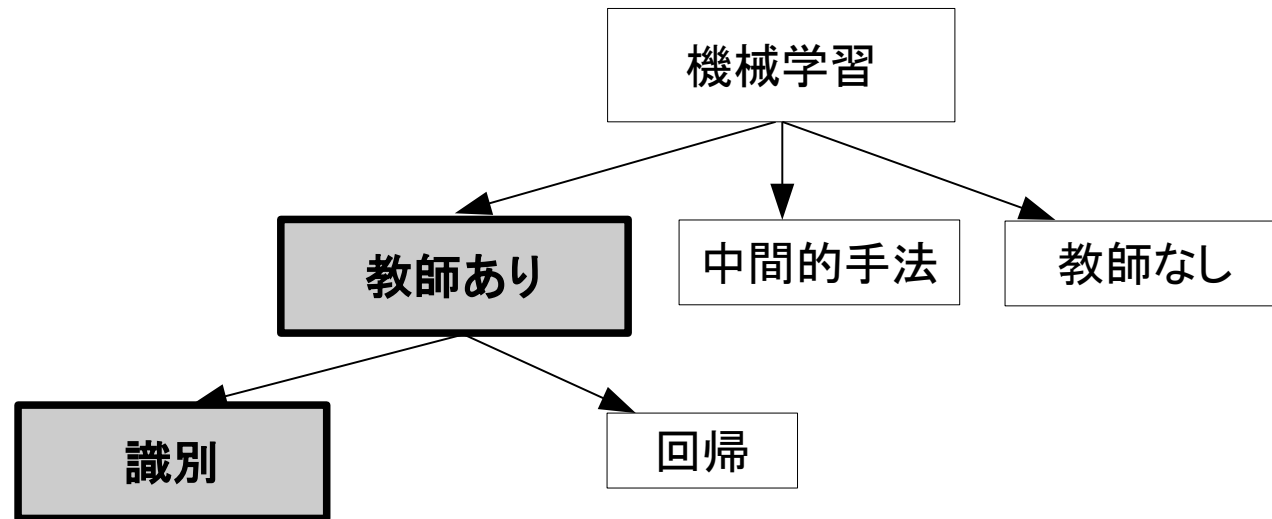
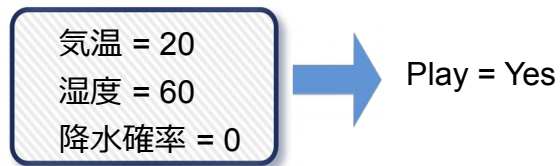


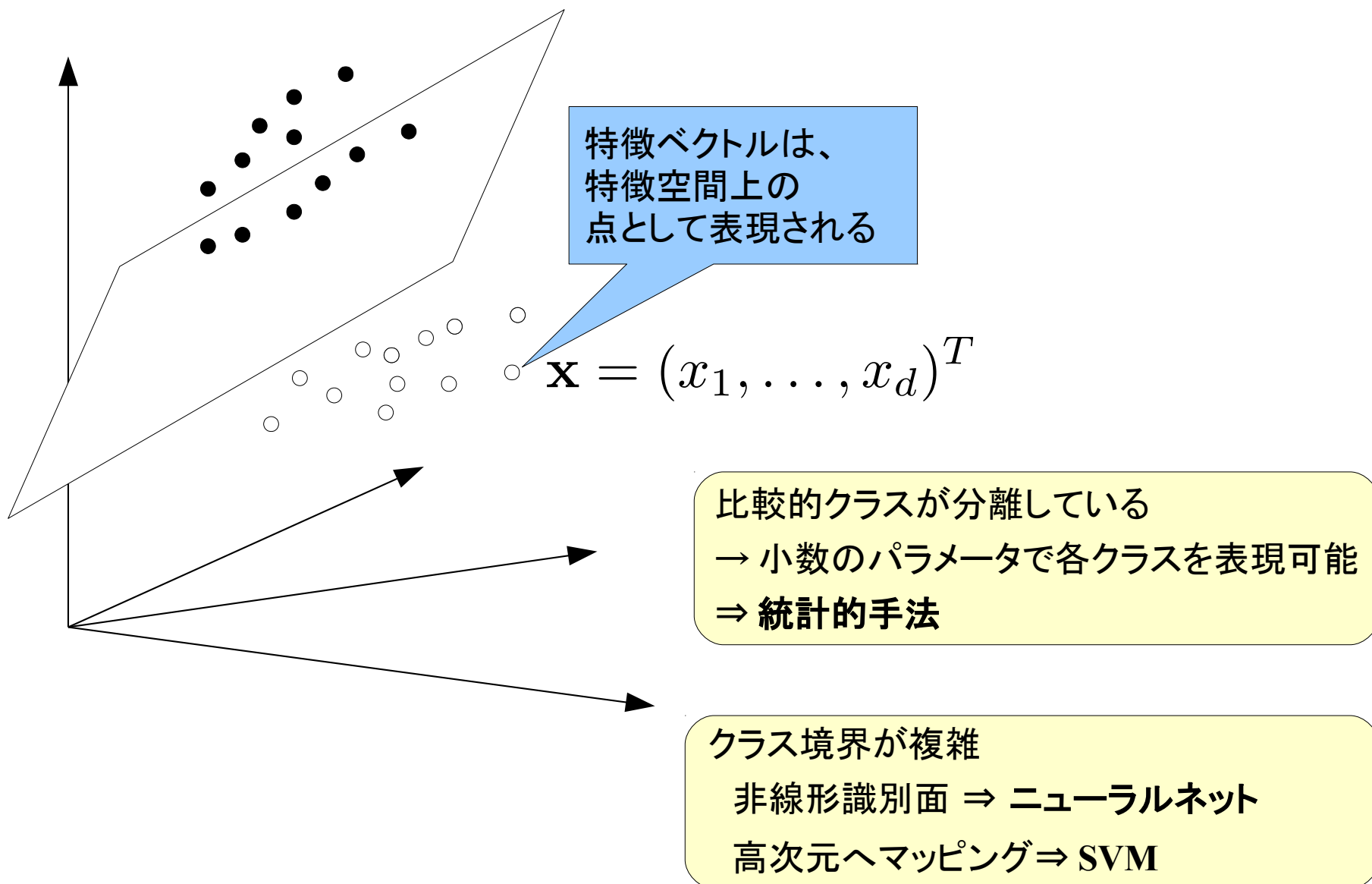
5. 識別 — 生成モデルと識別モデル —



- ラベル特徴
- 数値特徴



5.1 数値特徴に対する「教師あり・識別」問題の定義



5.2 数値特徴に対するベイズ識別

5.2.1 数値特徴に対するナイーブベイズ識別

$$C_{NB} = \arg \max_i P(\omega_i) \prod_{j=1}^d p(x_j | \omega_i)$$

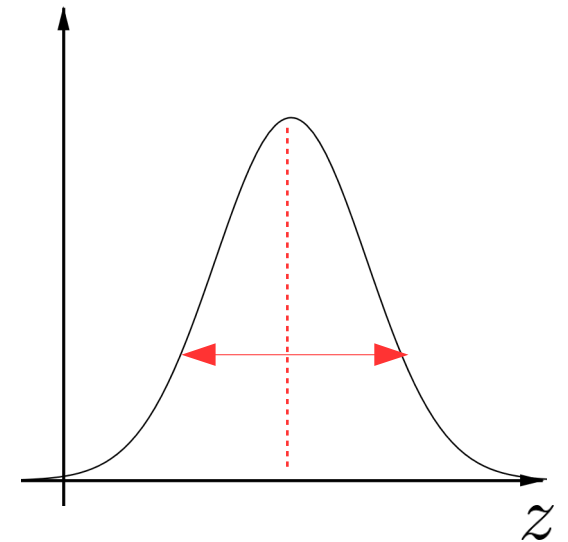
- 確率密度関数 $p(x_j | \omega_i)$ の推定

- 正規分布を仮定

$$p(z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z - \mu)^2}{2\sigma^2}\right)$$

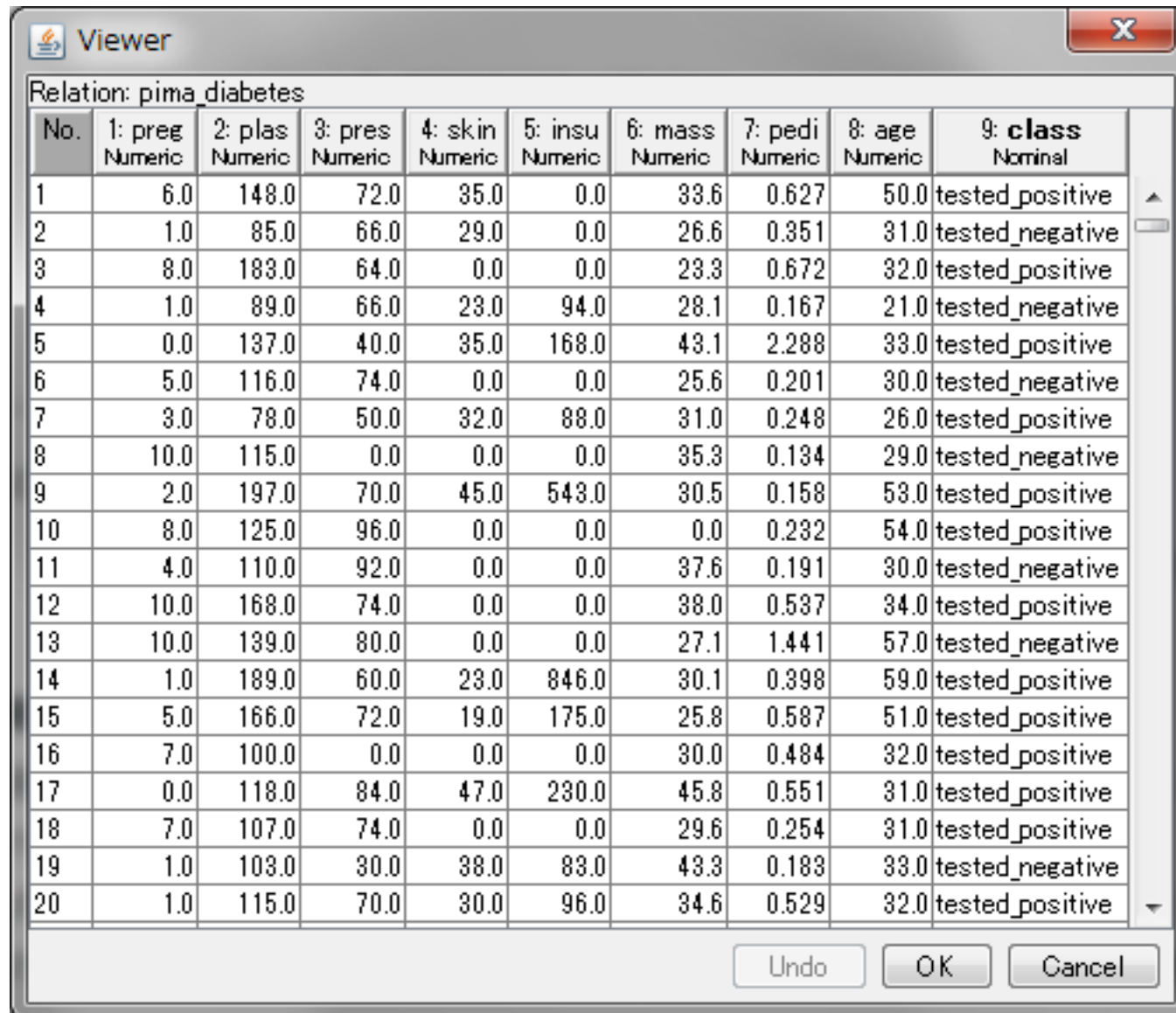
- 平均 μ と分散 σ を最尤推定

- それぞれ、学習データの平均と分散になる



5.2.1 数値特徴に対するナイーブベイズ識別

diabetes データ



Relation: pima diabetes

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
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	tested_negative
9	2.0	197.0	70.0	45.0	543.0	30.5	0.158	53.0	tested_positive
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	tested_positive
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	tested_negative
12	10.0	168.0	74.0	0.0	0.0	38.0	0.537	34.0	tested_positive
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	tested_negative
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	tested_positive
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	tested_positive
16	7.0	100.0	0.0	0.0	0.0	30.0	0.484	32.0	tested_positive
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	tested_positive
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	tested_positive
19	1.0	103.0	30.0	38.0	83.0	43.3	0.183	33.0	tested_negative
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	tested_positive

Undo OK Cancel

5.2.1 数値特徴に対するナイーブベイズ識別

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier: Choose NaiveBayes

Test options:
☒ Use training set
☐ Supplied test set (Set...)
☐ Cross-validation (Folds: 10)
☐ Percentage split (%: 66)
More options...

(Nom) class: [v]
Start Stop

Result list (right-click for options):
12:44:23 - trees.J48
13:10:03 - bayes.NaiveBayes
13:17:28 - bayes.NaiveBayes

Status: OK

Classifier output:
=== Classifier model (full training set) ===
Naive Bayes Classifier

Attribute	Class	tested negative (0.65)	tested positive (0.35)
preg			
mean		3.4234	4.9795
std. dev.		3.0166	3.6827
weight sum		500	268
precision		1.0625	1.0625

=====
Attribute negative positive
=====
preg
mean 3.4234 4.9795
std. dev. 3.0166 3.6827
weight sum 500 268
precision 1.0625 1.0625

5.2.2 生成モデルの考え方

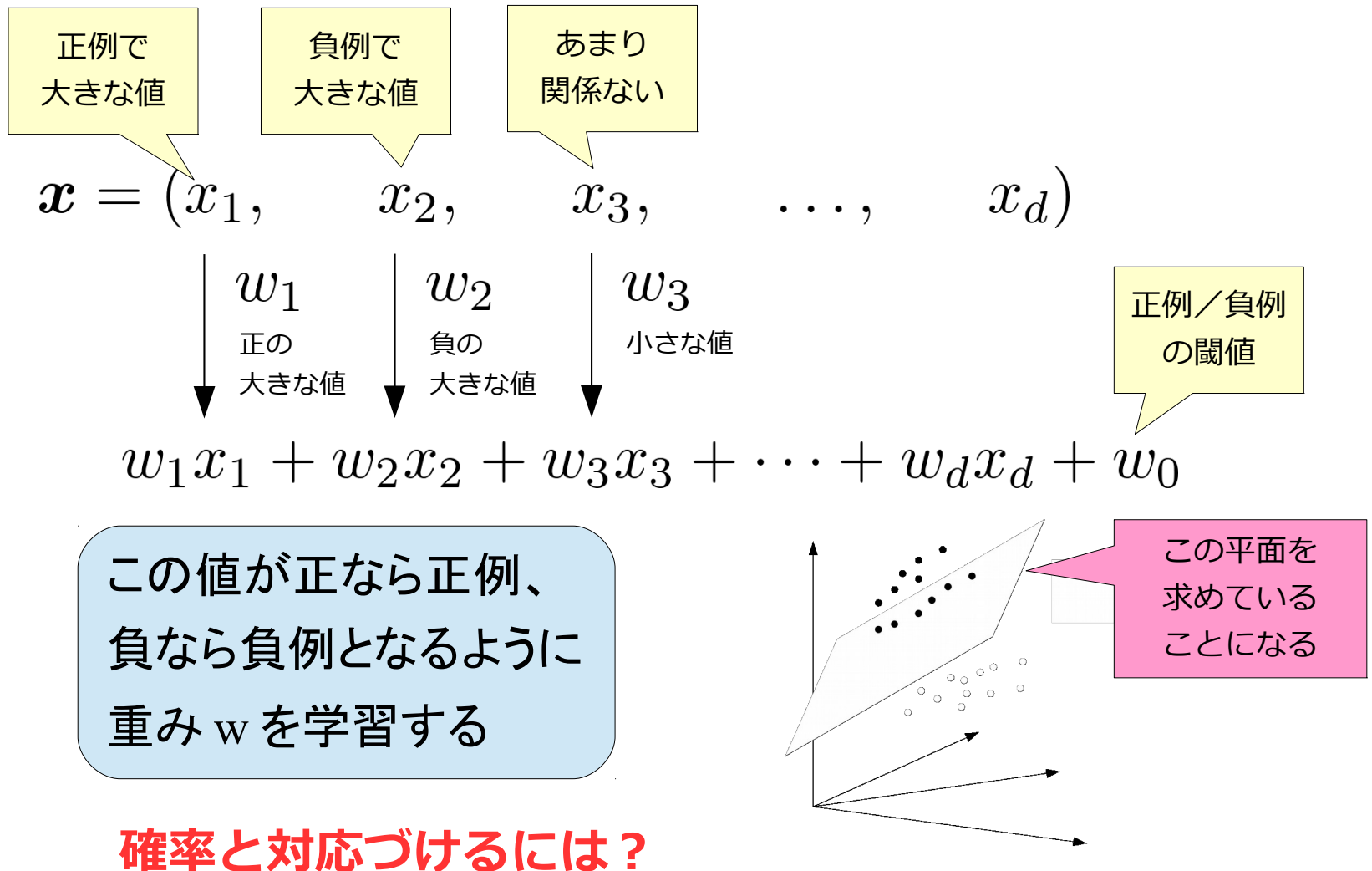
- 事後確率を求めるにあたって、同時確率を求めている
 - データが生成される様子をモデル化しているとも見ることが出来る
 - 事前確率に基づいてクラスを選ぶ
 - そのもとで、特徴ベクトルを出力する

$$\begin{aligned} P(\omega_i | \mathbf{x}) &= \frac{p(\mathbf{x} | \omega_i) P(\omega_i)}{p(\mathbf{x})} \\ &= \frac{p(\omega_i, \mathbf{x})}{p(\mathbf{x})} \end{aligned}$$

事後確率を求めるより、
難しい問題を解いている
のではないかな？

5.3.1 識別モデルの考え方

- 事後確率を直接求める

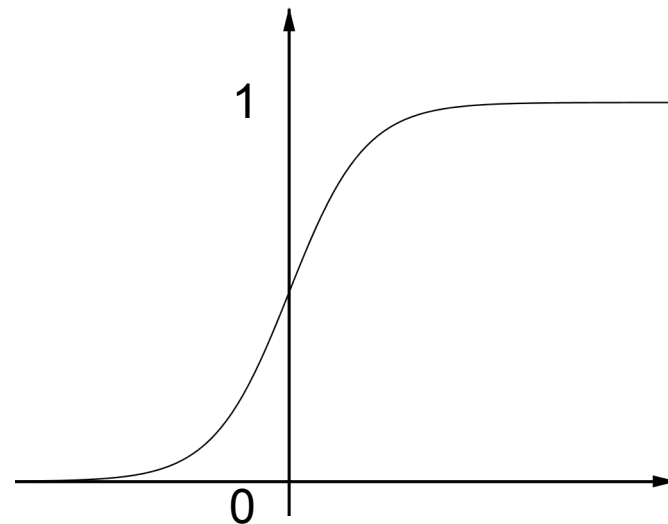


5.3.1 識別モデルの考え方

- ロジスティック識別
 - 入力が正例である確率

$$P(\oplus|\boldsymbol{x}) = \frac{1}{1 + \exp(-(\boldsymbol{w} \cdot \boldsymbol{x} + w_0))}$$

$-\infty \sim +\infty$ の値域を持つ
ものを、順序を変えずに
 $0 \sim 1$ にマッピング



シグモイド関数

5.3.2 ロジスティック識別器の学習

- 最適化対象 = モデルが学習データを生成する確率

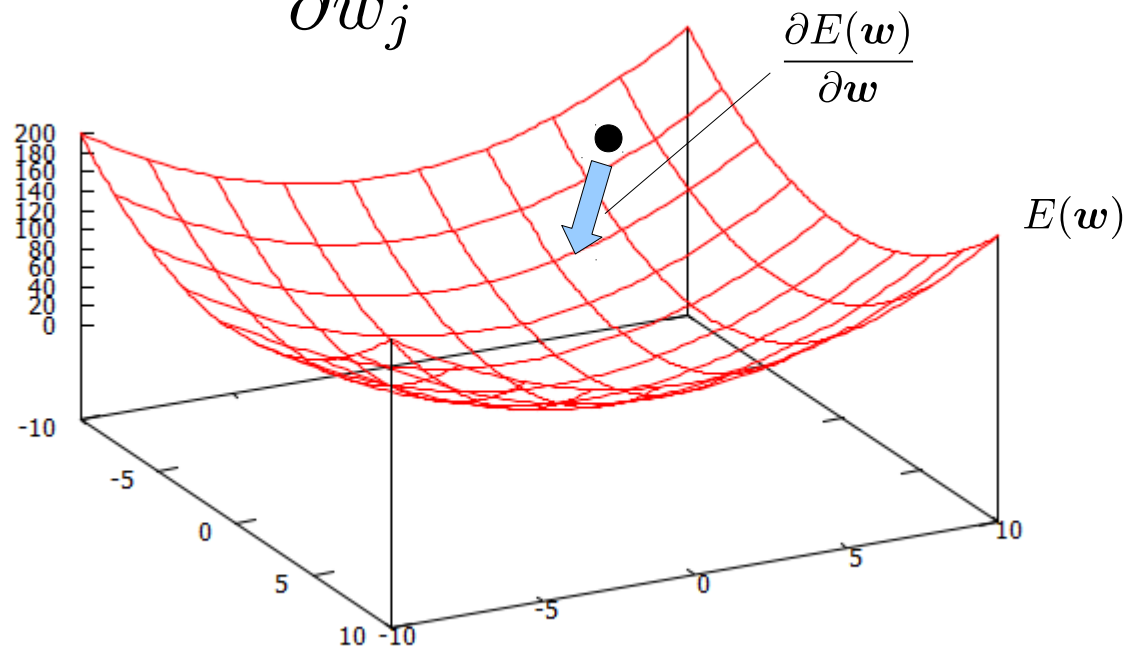
$$E(\mathbf{w}) = -\log P(D|\mathbf{w}) = -\log \prod_{\mathbf{x}_i \in D} o_i^{y_i} (1 - o_i)^{(1-y_i)}$$

- $E(\mathbf{w})$ を最急勾配法で最小化

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

$$o = P(\oplus | \mathbf{x})$$
$$y = o \text{ or } 1$$

正解ラベル



5.3.2 ロジスティック識別器の学習

- 重み更新量の計算

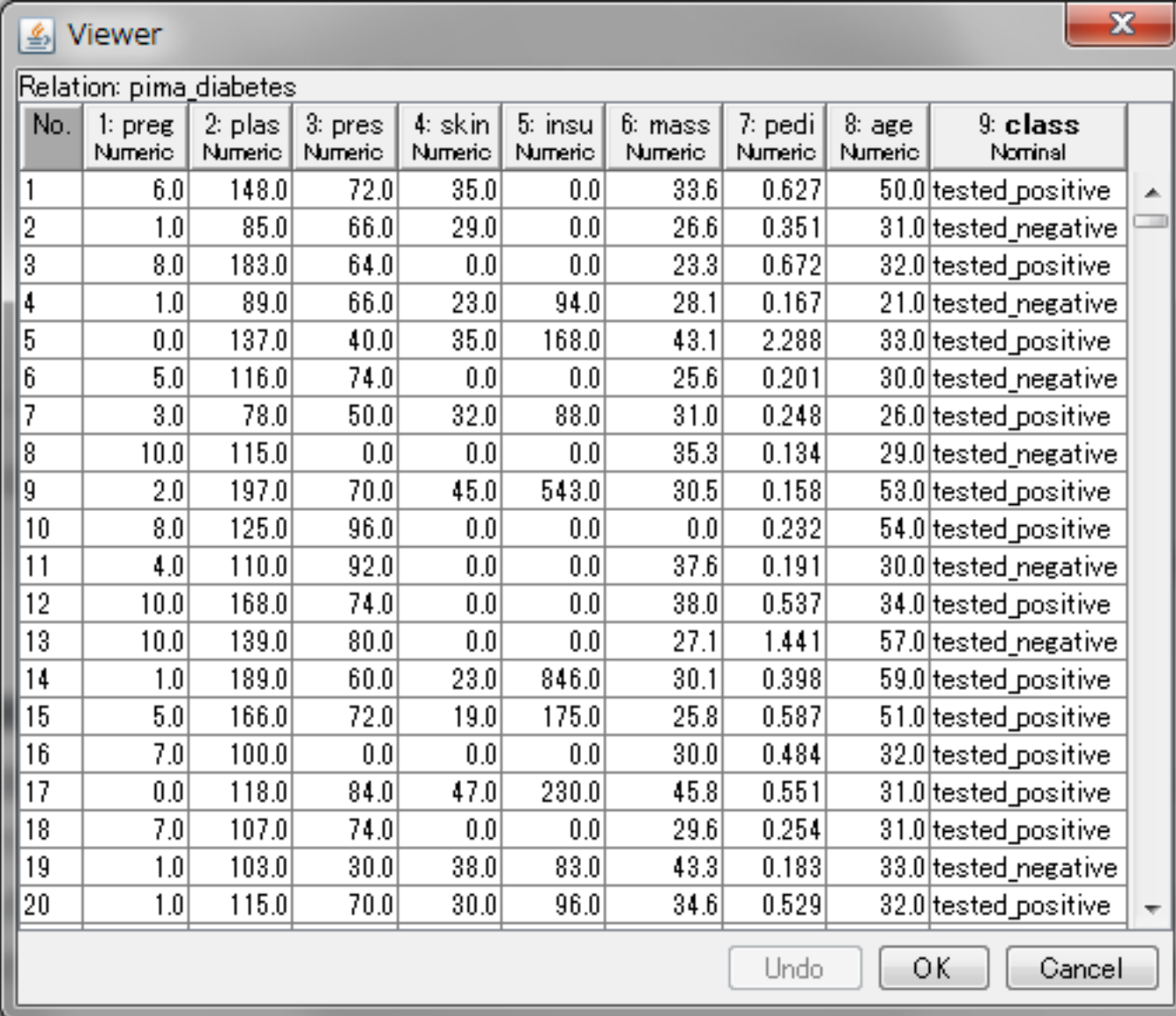
$$\begin{aligned}\frac{\partial E(\boldsymbol{w})}{\partial w_j} &= \sum_{\boldsymbol{x}_i \in D} \left(\frac{y_i}{o_i} - \frac{1 - y_i}{1 - o_i} \right) o_i (1 - o_i) x_{ij} \\ &= \sum_{\boldsymbol{x}_i \in D} (y_i - o_i) x_{ij}\end{aligned}$$

- 重みの更新式

$$w_j \leftarrow w_j - \eta \sum_{\boldsymbol{x}_i \in D} (y_i - o_i) x_{ij}$$

5.3.2 ロジスティック識別器の学習

diabetes データ



Relation: pima_diabetes

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
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	tested_negative
9	2.0	197.0	70.0	45.0	543.0	30.5	0.158	53.0	tested_positive
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	tested_positive
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	tested_negative
12	10.0	168.0	74.0	0.0	0.0	38.0	0.537	34.0	tested_positive
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	tested_negative
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	tested_positive
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	tested_positive
16	7.0	100.0	0.0	0.0	0.0	30.0	0.484	32.0	tested_positive
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	tested_positive
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	tested_positive
19	1.0	103.0	30.0	38.0	83.0	43.3	0.183	33.0	tested_negative
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	tested_positive

Undo OK Cancel

5.3.2 ロジスティック識別器の学習

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'SimpleLogistic -I 0 -M 500 -H 50 -W 0.0'. The 'Test options' section has 'Cross-validation' selected with 'Folds' set to 10. The 'Classifier output' pane shows the results of a 10-fold cross-validation. A red box highlights the coefficients for Class 0, which are then shown in a larger blue box on the right.

Classifier output

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

SimpleLogistic:

Class 0 :

4.18 +

[preg] * -0.06 +

[plas] * -0.02 +

[pres] * 0.01 +

[insu] * 0 +

[mass] * -0.04 +

[pedi] * -0.47 +

[age] * -0.01

Class 1 :

-4.18 +

[preg] * 0.06 +

[plas] * 0.02 +

[pres] * -0.01 +

[insu] * 0 +

[mass] * 0.04 +

[pedi] * 0.47 +

[age] * 0.01

Time taken to build model: 0.11 seconds

=== Stratified cross-validation ===

=== Summary ===

Result list (right-click for options)

13:22:41 - functions.SimpleLogistic

Status: OK

Log x 0

実行例

- 入力

妊娠 血糖値 血圧 ...
 $\mathbf{x}=(6, 148, 72, 35, 0, 33.6, 0.627, 50)$

$$\begin{aligned} g(\mathbf{x}) &= 6 \times 0.06 + 148 \times 0.02 - 72 \times 0.01 \\ &\quad + 33.6 \times 0.04 + 0.627 \times 0.47 + 50 \times 0.01 \\ &\quad - 4.18 \\ &= 0.559 \end{aligned}$$

$$P(\text{tested_positive}) = 1/(1 + \exp(-g(\mathbf{x}))) = 0.636$$

- 出力

tested_positive