機械学習
アンサンブル学習

管理工学科
篠沢佳久

資料の内容

- アンサンブル学習
  - ブースティング
  - アダブースト（AdaBoost）
- 実習
  - Toy problem（表計算）
  - アダブースト（Breast Cancer Dataset）

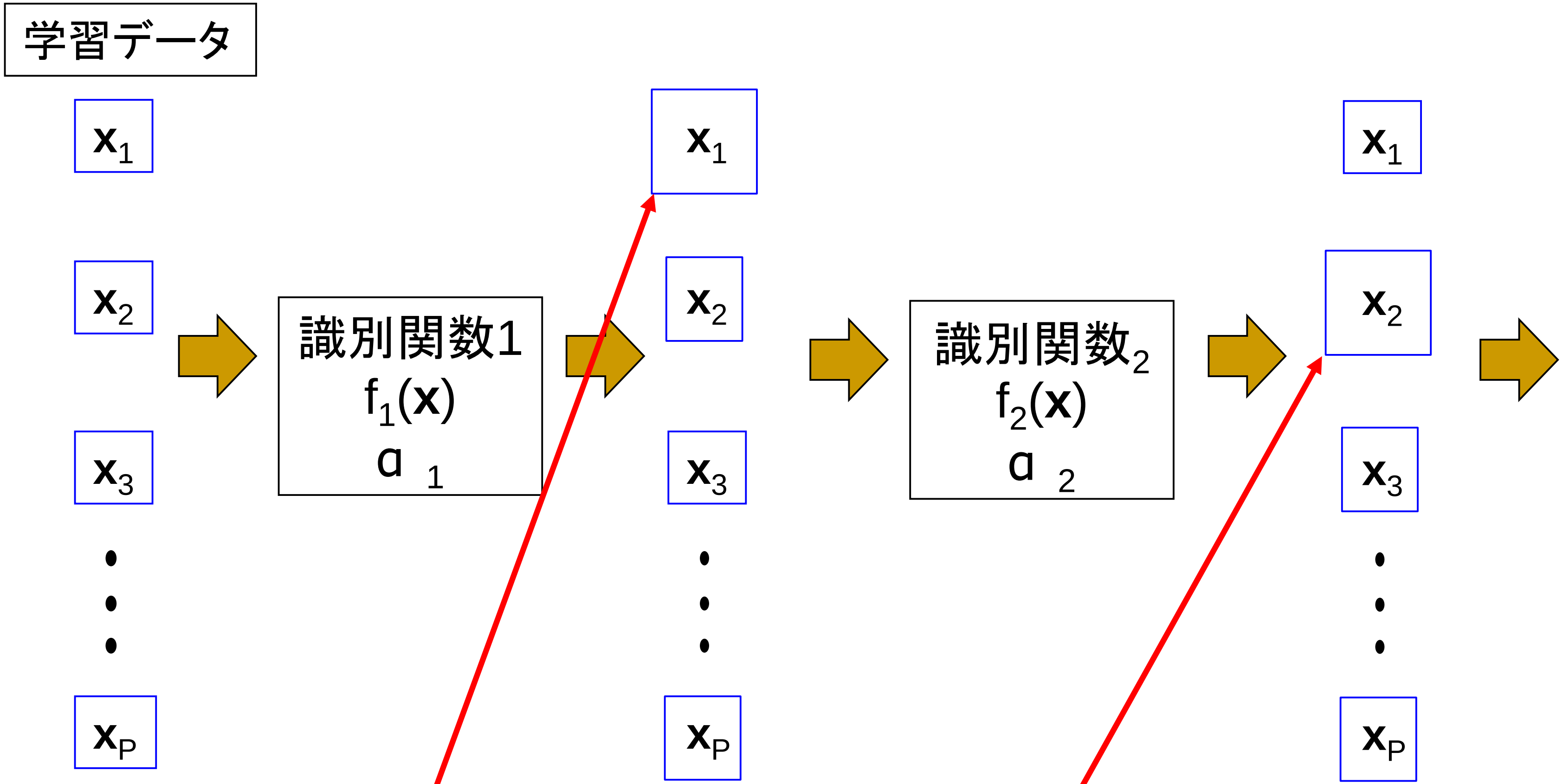
ブースティング

アダブースト

ブースティング①

□ はデータに対する重み（誤認識の場合は大きい）

正しく学習できなかった場合，次の識別関数で訂正

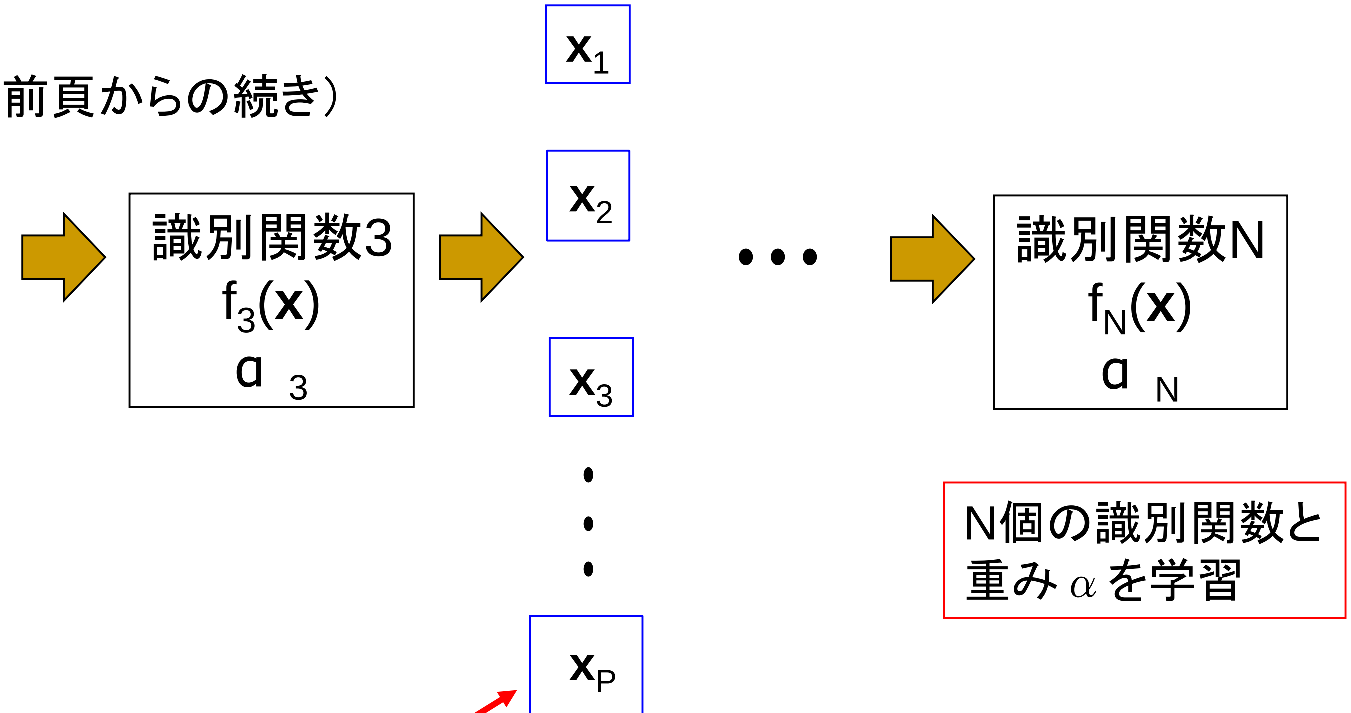


x_1を識別関数1では正しく学習できなかった
→ x_1に対する重みを大きくし，次の識別関数の学習の際，学習できるようにする

x_2を識別関数2では正しく学習できなかった
→ x_2に対する重みを大きくし，次の識別関数の学習の際，学習できるようにする

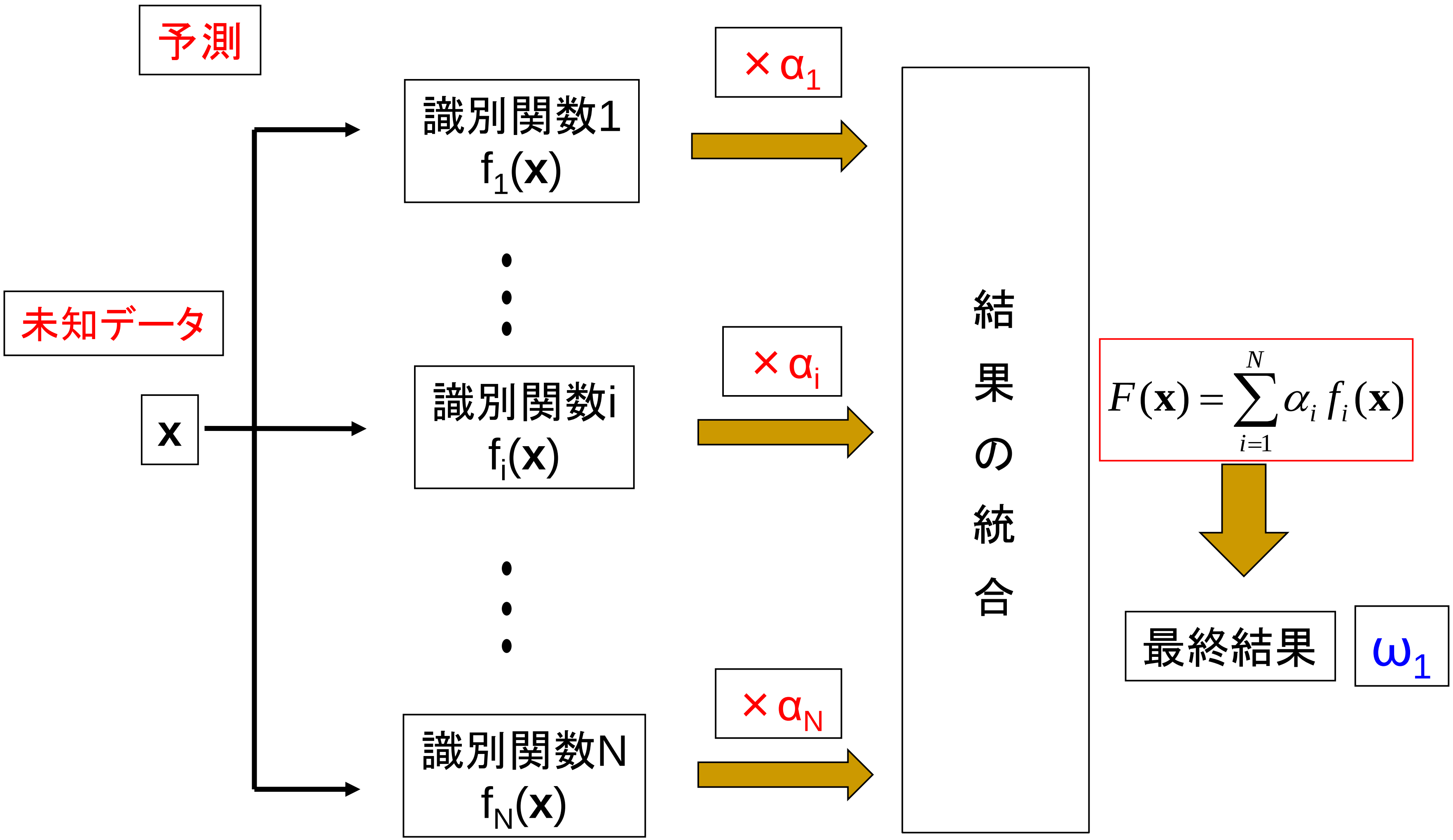
ブースティング②

（前頁からの続き）



x_Pを識別関数3では正しく学習できなかった
→ x_Pに対する重みを大きくし，次の識別関数の学習の際，学習できるようにする

ブースティング③


予測
未知データ
$\mathbf{x}$
識別関数1
$f_1(\mathbf{x})$
識別関数i
$f_i(\mathbf{x})$
識別関数N
$f_N(\mathbf{x})$
$\times\,\alpha_1$
$\times\,\alpha_i$
$\times\,\alpha_N$
結果の統合
$F(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i f_i(\mathbf{x})$
最終結果
ω_1

アダブースト①

- アダブースト（AdaBoost）
  - 二クラス（ω_1, ω_2）分類問題に対する手法*
  - N個の識別関数f_i（i=1,2,・・・,N）をあらかじめ準備

$$f_i(\mathbf{x}) = \mathrm{sgn}(g_i(\mathbf{x})) = \begin{cases} -1 & \text{if } \mathbf{x} \in \omega_1 \\ +1 & \text{if } \mathbf{x} \in \omega_2 \end{cases}$$

識別関数の精度は50%を超えていなければならない

*多クラス分類問題も可能です

アダブースト②

- 最終的な識別関数

$$F(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{N} \alpha_i f_i(\mathbf{x}) \right)$$

  - 使用する識別関数f_iおよび係数α_iを決定する手法

アダブーストのアルゴリズム①

二クラス分類問題

- データ $\mathbf{x}_i$（i=1,2,・・・,P）
- データ$\mathbf{x}_i$に対する正解y_i={-1,+1}
- データの重み $D(\mathbf{x}_i)$ $\sum_{i=1}^{P} D(\mathbf{x}_i)=1, D(\mathbf{x}_i)\geq 0$
- N個の識別関数f_jを準備*

（どのような方法を用いてもよい，ただし精度は50%を越えなければならない）

*あらかじめN個の識別関数を準備しなくても可能です．その場合，次頁の②において
識別関数f_iの学習→誤差ε(f_i)の計算を行ないます．

アダブーストのアルゴリズム②

① データの重みの初期化

$$D(\mathbf{x}_i) = \frac{1}{P}$$

② 識別関数の選択

データ x_i およびデータの重み$D(x_i)$ を用いて識別関数ごとで誤差$\varepsilon(f_j)$を計算

$$\varepsilon(f_j) = \sum_{i=1}^{P} D(\mathbf{x}_i) I(y_i \neq f_j(\mathbf{x}_i))$$

$$\boxed{I(y_i \neq f_j(\mathbf{x}_i)) = \begin{cases} 1 & y_i \neq f_j(\mathbf{x}_i) \\ 0 & y_i = f_j(\mathbf{x}_i) \end{cases}}$$

誤差が最小となる識別関数f_jを選択

正解の場合　→0
間違いの場合→1

アダブーストのアルゴリズム②

③ 選択した識別関数f_jの誤差の計算

$$\varepsilon(f_j) = \sum_{i=1}^{P} D(\mathbf{x}_i) I(y_i \neq f_j(\mathbf{x}_i))$$

$$\boxed{I(y_i \neq f_j(\mathbf{x}_i)) = \begin{cases} 1 & y_i \neq f_j(\mathbf{x}_i) \\ 0 & y_i = f_j(\mathbf{x}_i) \end{cases}}$$

$$\varepsilon(f_j) \leftarrow \varepsilon(f_j) + D(\mathbf{x}_i) \ \text{ if } \ y_i \neq f_j(\mathbf{x}_i)$$

④ 係数の計算

$$\alpha_j \leftarrow \frac{1}{2} \ln\left(\frac{1-\varepsilon(f_j)}{\varepsilon(f_j)}\right)$$

アダブーストのアルゴリズム③

⑤ データの重みの更新

$$D(\mathbf{x}_i) \leftarrow \begin{cases} D(\mathbf{x}_i)e^{-\alpha_j} / Z \ \ \text{if} \ \ f_j(\mathbf{x}_i) = y_i \\ D(\mathbf{x}_i)e^{+\alpha_j} / Z \ \ \text{if} \ \ f_j(\mathbf{x}_i) \neq y_i \end{cases}$$

$$Z = \sum_{i=1}^{P} D(\mathbf{x}_i)\exp(-\alpha y_i f_j(\mathbf{x}_i))$$

Zは$\Sigma D(\mathbf{x}_i)=1$と正規化するための定数

②～⑤をN回繰り返し，N個の識別関数を選択

⑥ 最終的な識別関数を求める

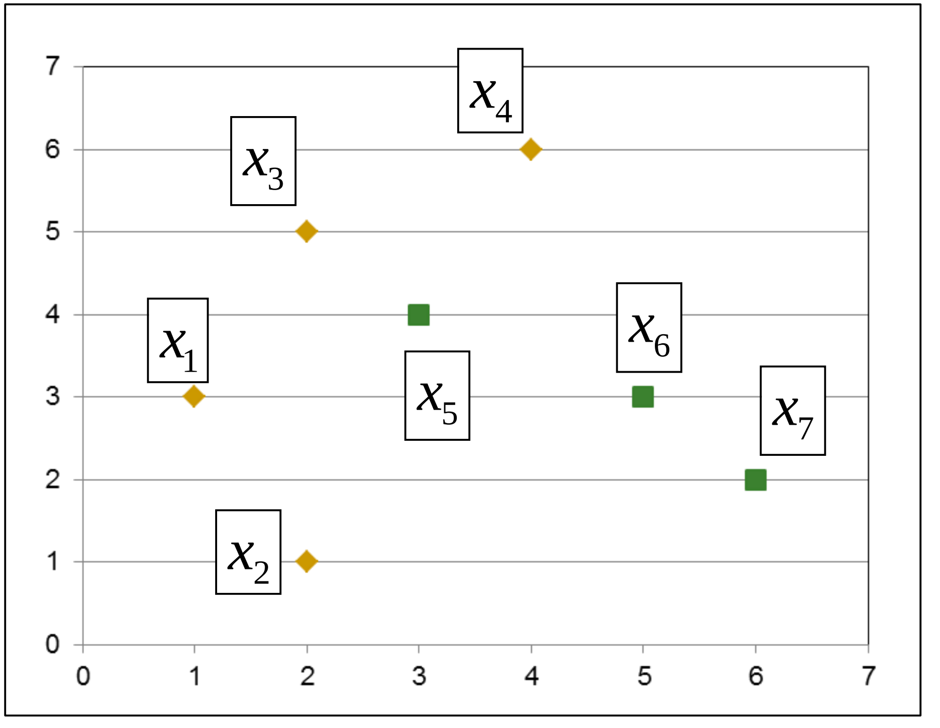
$$F(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i f_i(\mathbf{x}) \right)$$

アダブーストの例（Toy Problem）

表計算でのアダブーストとのアルゴリズムの理解

二クラス分類問題

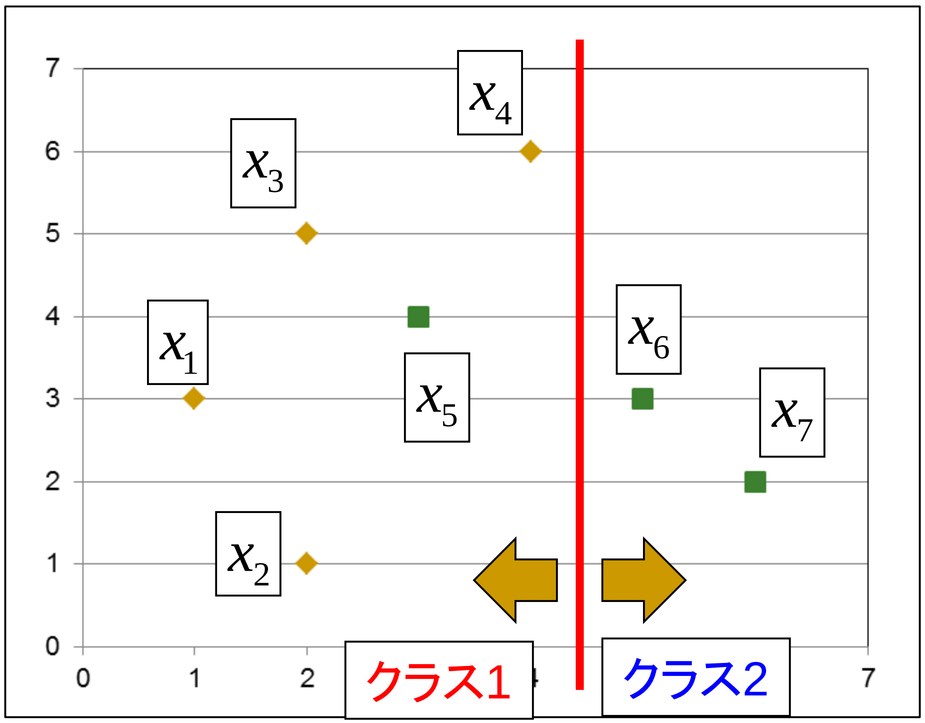
	x_1	x_2	
x_1	1	3	クラス1
x_2	2	1	クラス1
x_3	2	5	クラス1
x_4	4	6	クラス1
x_5	3	4	クラス2
x_6	5	3	クラス2
x_7	6	2	クラス2



使用する識別関数①

識別関数1

$$f_1(\mathbf{x}_i) = \begin{cases} -1 & \text{if } x_1 < 4.5 \\ +1 & \text{if } x_1 > 4.5 \end{cases}$$



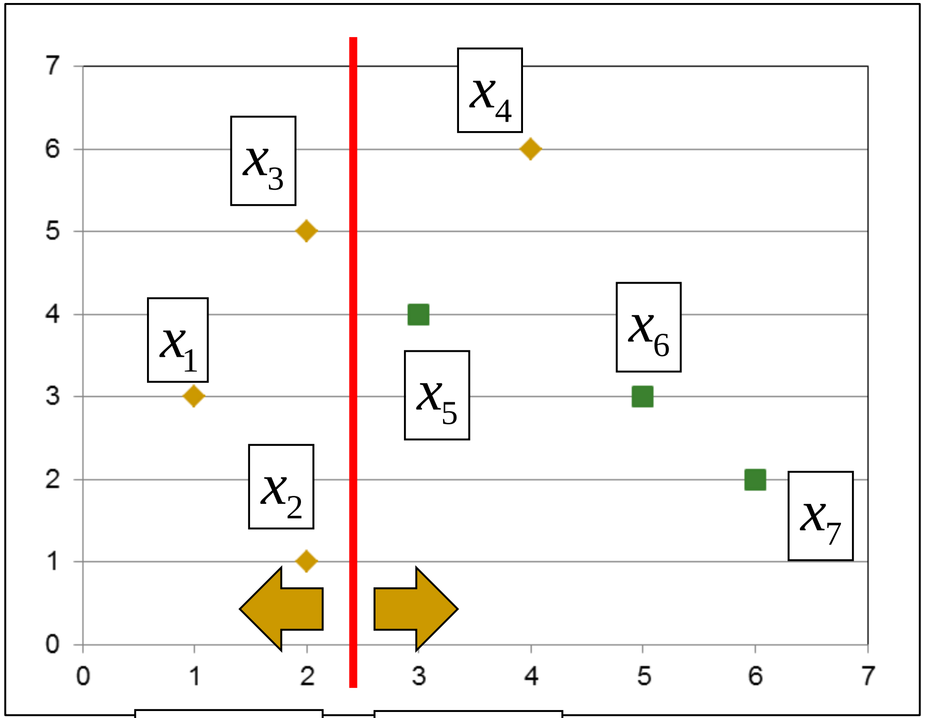
識別関数1による結果

	x_1	x_2	$f_1(x)$
x_1	1	3	-1
x_2	2	1	-1
x_3	2	5	-1
x_4	4	6	-1
x_5	3	4	-1
x_6	5	3	+1
x_7	6	2	+1

使用する識別関数②

識別関数2

$$f_2(\mathbf{x}_i) = \begin{cases} -1 & \text{if} \;\; x_1 < 2.5 \\ +1 & \text{if} \;\; x_1 > 2.5 \end{cases}$$



クラス1 クラス2

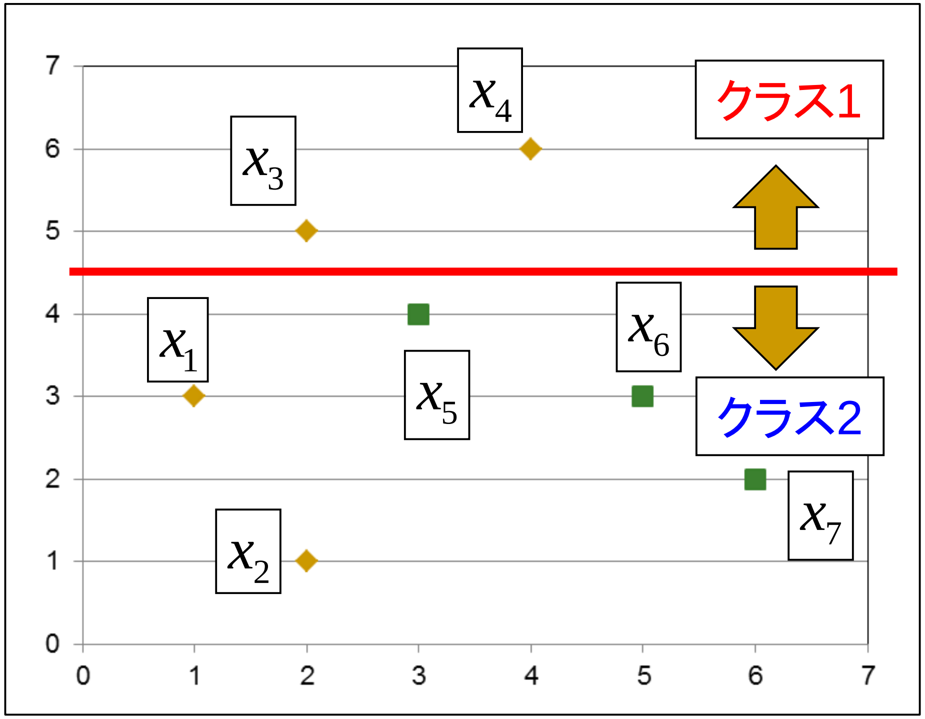
識別関数2による結果

	x_1	x_2	$f_3(x)$
x_1	1	3	-1
x_2	2	1	-1
x_3	2	5	-1
x_4	4	6	+1
x_5	3	4	+1
x_6	5	3	+1
x_7	6	2	+1

使用する識別関数③

識別関数3

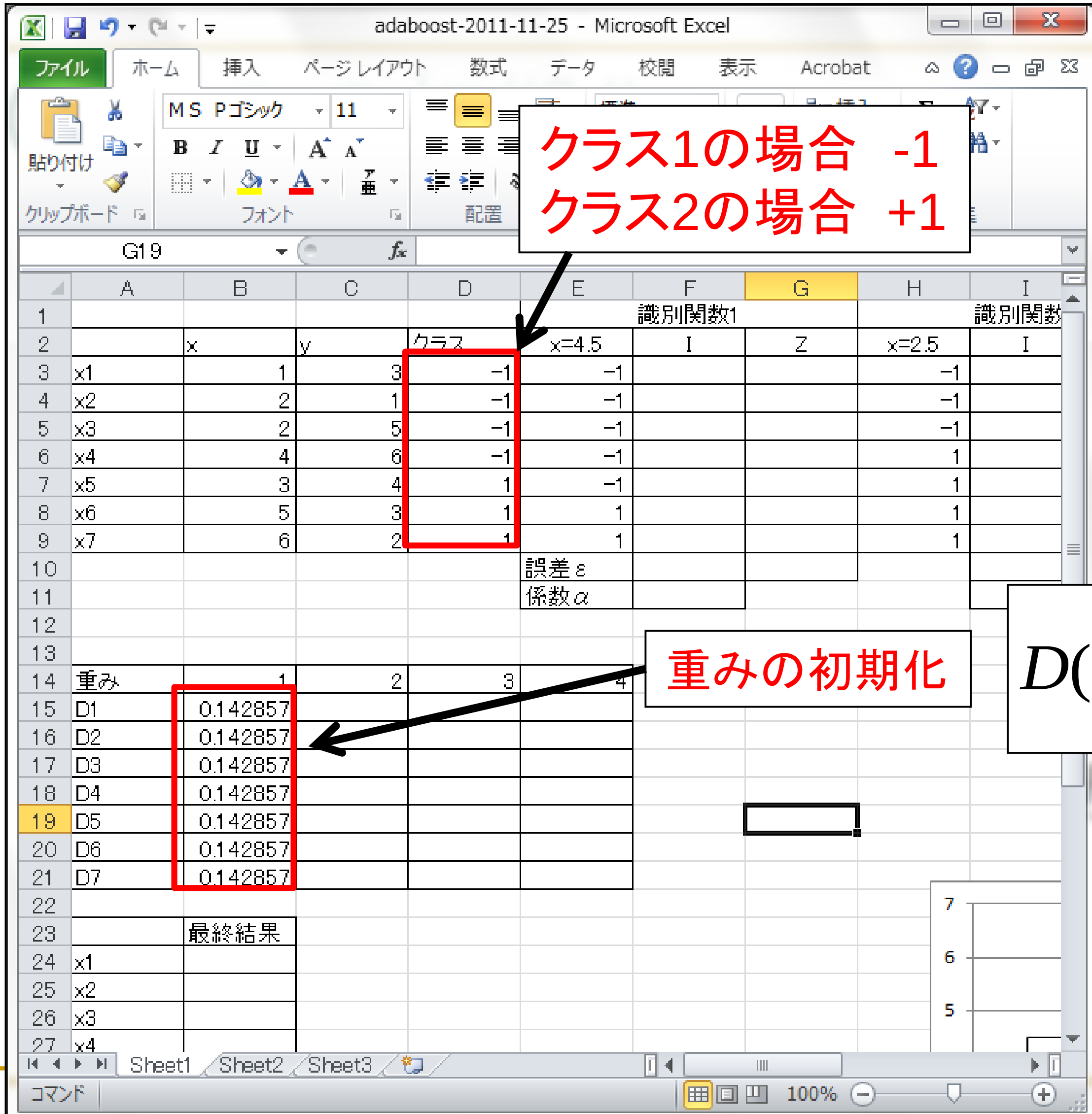
$$f_3(\mathbf{x}_i) = \begin{cases} -1 & \text{if } x_2 < 4.5 \\ +1 & \text{if } x_2 > 4.5 \end{cases}$$



識別関数3による結果

	x_1	x_2	$f_2(x)$
x_1	1	3	+1
x_2	2	1	+1
x_3	2	5	-1
x_4	4	6	-1
x_5	3	4	+1
x_6	5	3	+1
x_7	6	2	+1

① 重みの初期化



adaboost-2011-11-25 - Microsoft Excel

	A	B	C	D	E	F	G	H	I
1						識別関数1			識別関数
2		x	y	クラス	x=4.5	I	Z	x=2.5	I
3	x1	1	3	-1	-1			-1	
4	x2	2	1	-1	-1			-1	
5	x3	2	5	-1	-1			-1	
6	x4	4	6	-1	-1			1	
7	x5	3	4	1	-1			1	
8	x6	5	3	1	1			1	
9	x7	6	2	1	1			1	
10					誤差ε				
11					係数α				
12									
13									
14	重み	1	2	3	4				
15	D1	0.142857							
16	D2	0.142857							
17	D3	0.142857							
18	D4	0.142857							
19	D5	0.142857							
20	D6	0.142857							
21	D7	0.142857							
22									
23		最終結果							
24	x1								
25	x2								
26	x3								
27	x4								

クラス1の場合　-1
クラス2の場合　+1

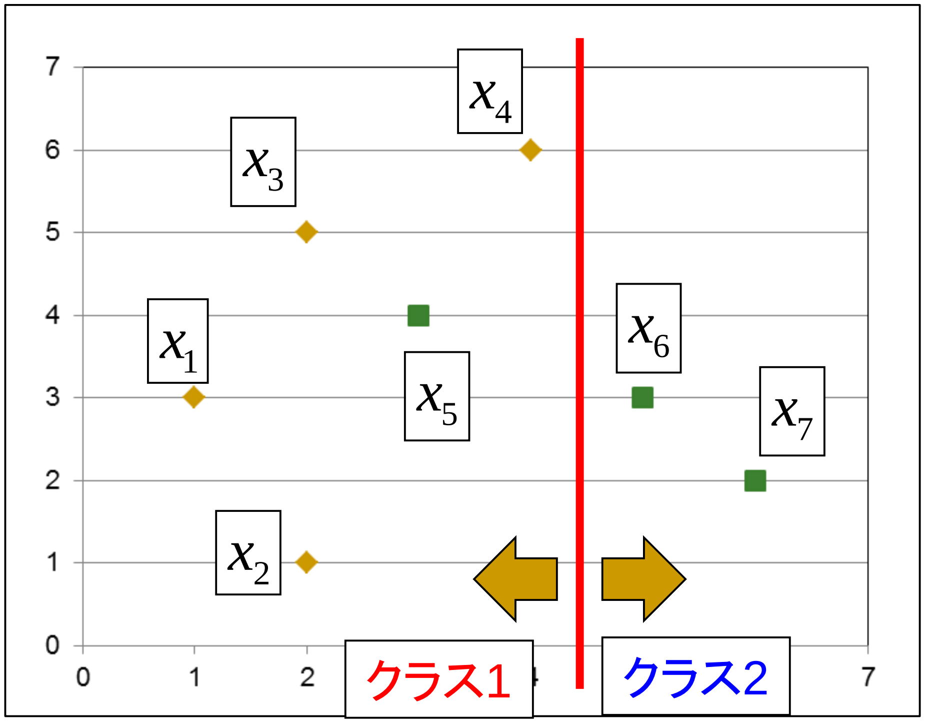
重みの初期化

$$D(\mathbf{x}_i) = \frac{1}{P}$$

② 識別関数の選択*

識別関数1

$$f_1(\mathbf{x}_i) = \begin{cases} -1 & \text{if } x_1 < 4.5 \\ +1 & \text{if } x_1 > 4.5 \end{cases}$$



識別関数1による結果

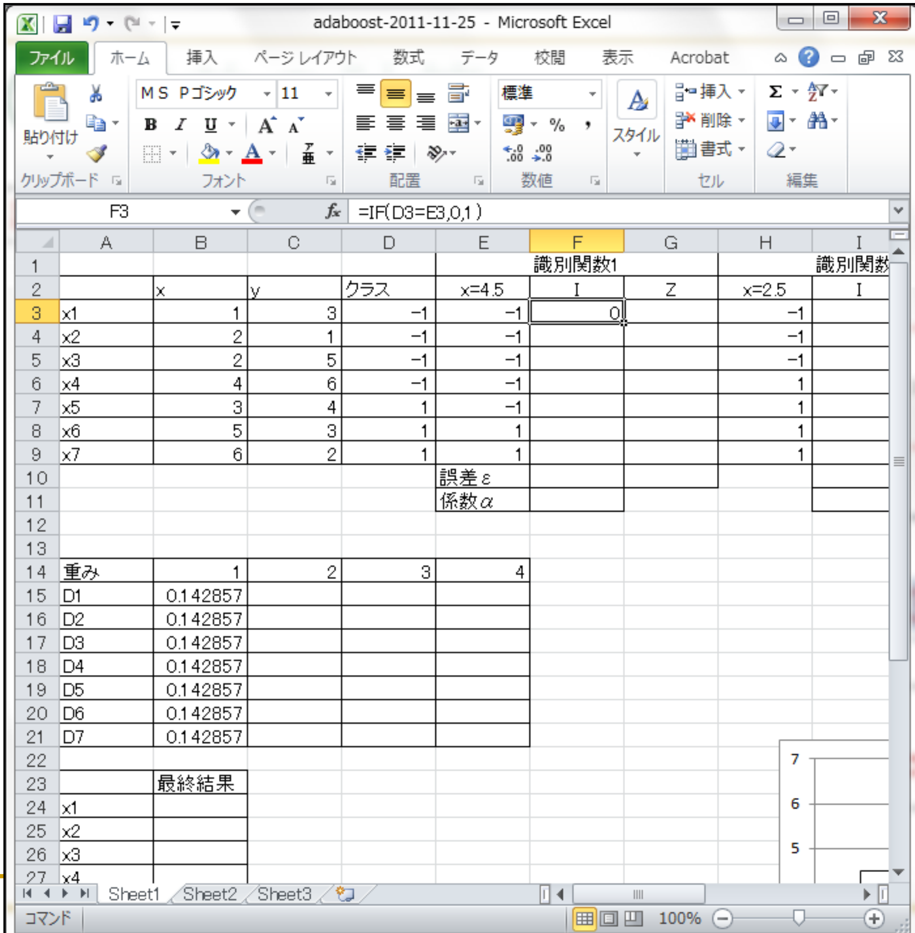
	x_1	x_2	$f_1(x)$
x_1	1	3	-1
x_2	2	1	-1
x_3	2	5	-1
x_4	4	6	-1
x_5	3	4	-1
x_6	5	3	+1
x_7	6	2	+1

*3個の識別関数の中から$\varepsilon(f_i)$が最小の識別関数を選択する，この場合，$\varepsilon(f_1)=0.142$，$\varepsilon(f_2)=0.142$，$\varepsilon(f_3)=0.285$より識別関数1とする

③ 誤差の計算

識別関数1の誤差

$$\varepsilon(f_j) = \sum_{i=1}^{P} D(\mathbf{x}_i) I(y_i \neq f_j(\mathbf{x}_i))$$



	A	B	C	D	E	F	G	H	I
1						識別関数1			識別関数
2		x	y	クラス	x=4.5	I	Z	x=2.5	I
3	x1	1	3	-1	-1	0		-1	
4	x2	2	1	-1	-1			-1	
5	x3	2	5	-1	-1			-1	
6	x4	4	6	-1	-1			1	
7	x5	3	4	1	-1			1	
8	x6	5	3	1	1			1	
9	x7	6	2	1	1			1	
10					誤差ε				
11					係数α				
12									
13									
14	重み	1	2	3	4				
15	D1	0.142857							
16	D2	0.142857							
17	D3	0.142857							
18	D4	0.142857							
19	D5	0.142857							
20	D6	0.142857							
21	D7	0.142857							
22									
23		最終結果							
24	x1								
25	x2								
26	x3								
27	x4								

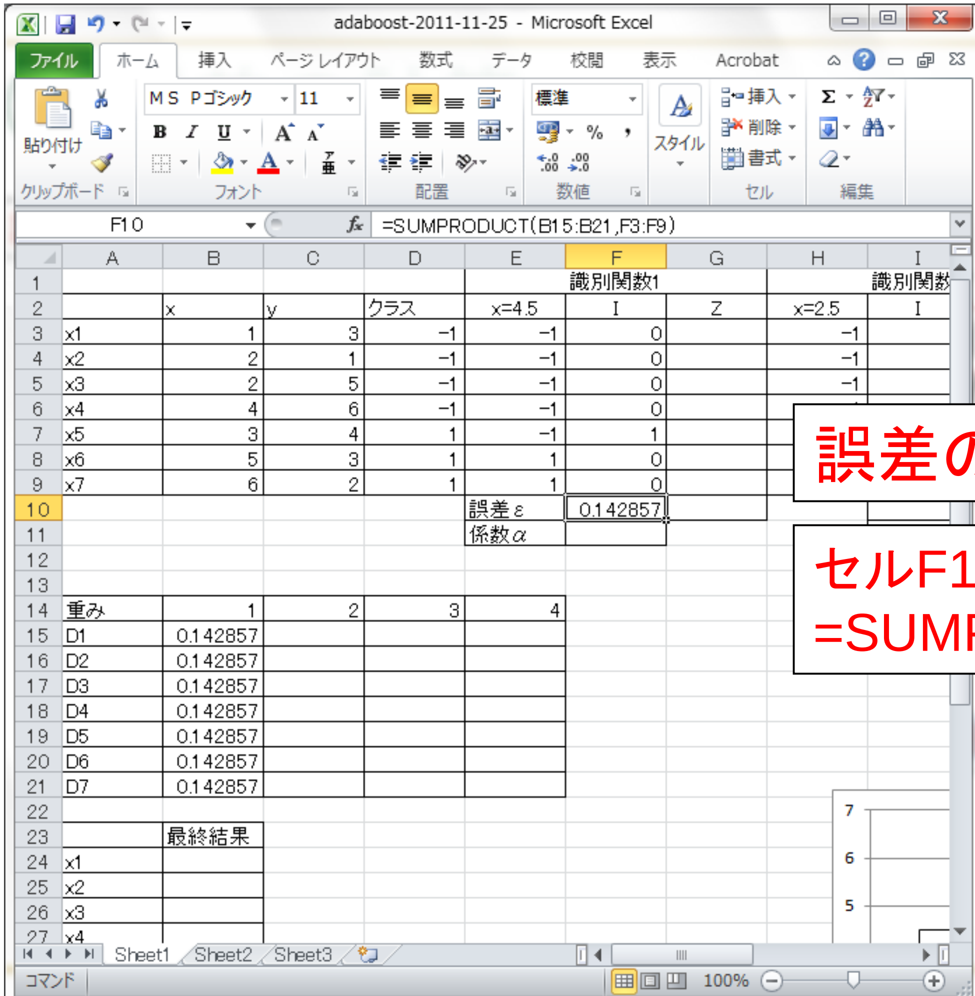
$$I(y_i \neq f_j(\mathbf{x}_i)) = \begin{cases} 1 & y_i \neq f_j(\mathbf{x}_i) \\ 0 & y_i = f_j(\mathbf{x}_i) \end{cases}$$

セルF3
=IF(D3=E3,0,1)

セルF3
セルF4～F9にコピー

③ 誤差の計算

識別関数1の誤差



	A	B	C	D	E	F	G	H	I
1						識別関数1			識別関数
2		x	y	クラス	x=4.5	I	Z	x=2.5	I
3	x1	1	3	-1	-1	0		-1	
4	x2	2	1	-1	-1	0		-1	
5	x3	2	5	-1	-1	0		-1	
6	x4	4	6	-1	-1	0			
7	x5	3	4	1	-1	1			
8	x6	5	3	1	1	0			
9	x7	6	2	1	1	0			
10					誤差ε	0.142857			
11					係数α				
12									
13									
14	重み	1	2	3	4				
15	D1	0.142857							
16	D2	0.142857							
17	D3	0.142857							
18	D4	0.142857							
19	D5	0.142857							
20	D6	0.142857							
21	D7	0.142857							
22									
23		最終結果							
24	x1								
25	x2								
26	x3								
27	x4								

$$\varepsilon(f_j) = \sum_{i=1}^{P} \boxed{D(\mathbf{x}_i)} \boxed{I(y_i \neq f_j(\mathbf{x}_i))}$$

セルB15～B21

セルF3～F9

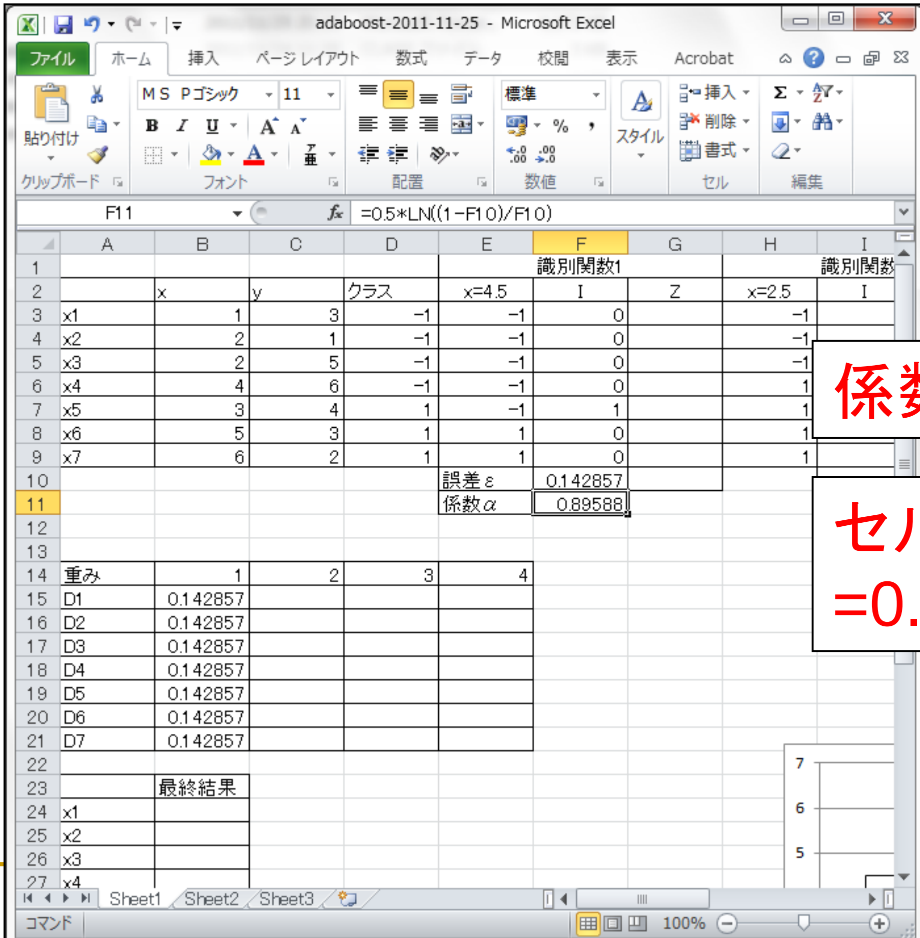
誤差の計算

セルF10
=SUMPRODUCT(B15:B21,F3:F9)

④ 係数の計算

識別関数1の係数

$$\alpha_j \leftarrow \frac{1}{2}\ln\left(\frac{1-\varepsilon(f_j)}{\varepsilon(f_j)}\right)$$



	A	B	C	D	E	F	G	H	I
1						識別関数1			識別関数
2		x	y	クラス	x=4.5	I	Z	x=2.5	I
3	x1	1	3	-1	-1	0		-1	
4	x2	2	1	-1	-1	0		-1	
5	x3	2	5	-1	-1	0		-1	
6	x4	4	6	-1	-1	0		1	
7	x5	3	4	1	-1	1		1	
8	x6	5	3	1	1	0		1	
9	x7	6	2	1	1	0		1	
10					誤差ε	0.142857			
11					係数α	0.89588			
12									
13									
14	重み	1	2	3	4				
15	D1	0.142857							
16	D2	0.142857							
17	D3	0.142857							
18	D4	0.142857							
19	D5	0.142857							
20	D6	0.142857							
21	D7	0.142857							
22									
23		最終結果							
24	x1								
25	x2								
26	x3								
27	x4								

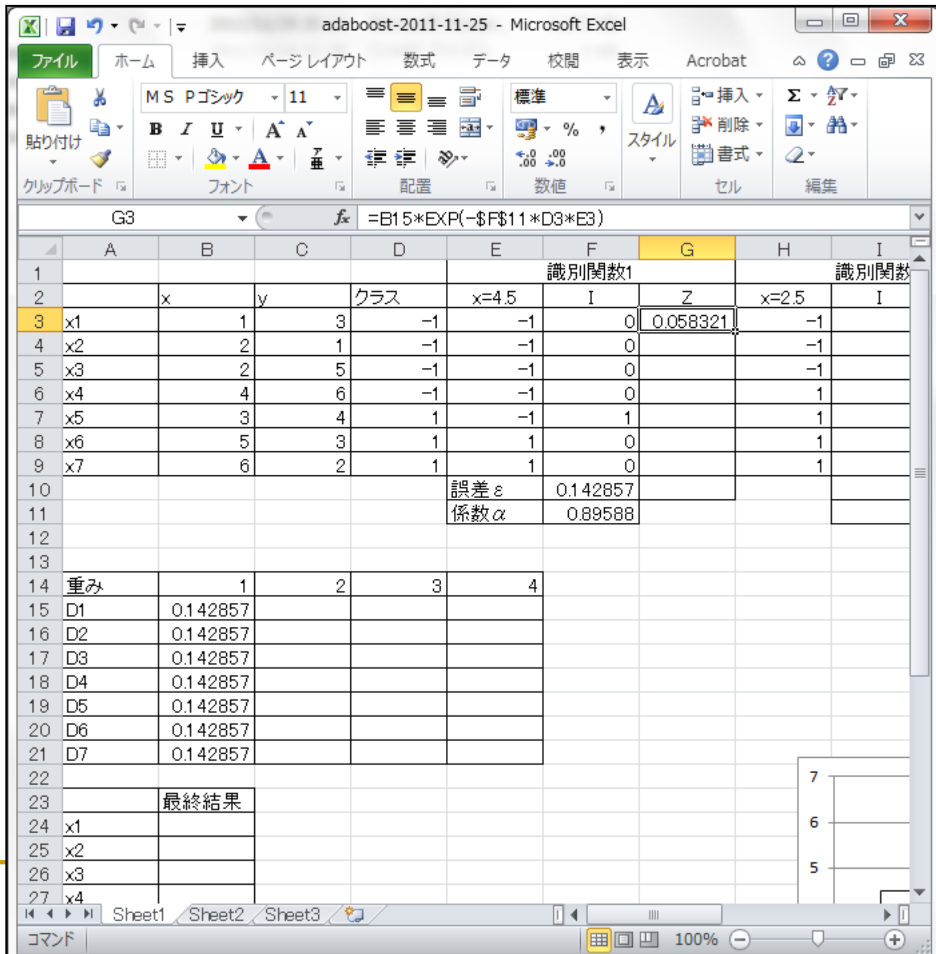
係数の計算

セルF11
=0.5*LN((1-F10)/F10)

⑤データの重みの更新

$$D(\mathbf{x}_i) \leftarrow \begin{cases} D(\mathbf{x}_i)e^{-\alpha_j}/Z & \text{if } f_j(\mathbf{x}_i)=y_i \\ D(\mathbf{x}_i)e^{+\alpha_j}/Z & \text{if } f_j(\mathbf{x}_i)\neq y_i \end{cases}$$

$$Z=\sum_{i=1}^{P}\boxed{D(\mathbf{x}_i)\exp(-\alpha y_i f_j(\mathbf{x}_i))}$$



	A	B	C	D	E	F	G	H	I
1						識別関数1			識別関数
2		x	y	クラス	x=4.5	I	Z	x=2.5	I
3	x1	1	3	-1	-1	0	0.058321	-1	
4	x2	2	1	-1	-1	0		-1	
5	x3	2	5	-1	-1	0		-1	
6	x4	4	6	-1	-1	0		1	
7	x5	3	4	1	-1	1		1	
8	x6	5	3	1	1	0		1	
9	x7	6	2	1	1	0		1	
10					誤差ε	0.142857			
11					係数α	0.89588			
12									
13									
14	重み	1	2	3	4				
15	D1	0.142857							
16	D2	0.142857							
17	D3	0.142857							
18	D4	0.142857							
19	D5	0.142857							
20	D6	0.142857							
21	D7	0.142857							
22									
23		最終結果							
24	x1								
25	x2								
26	x3								
27	x4								

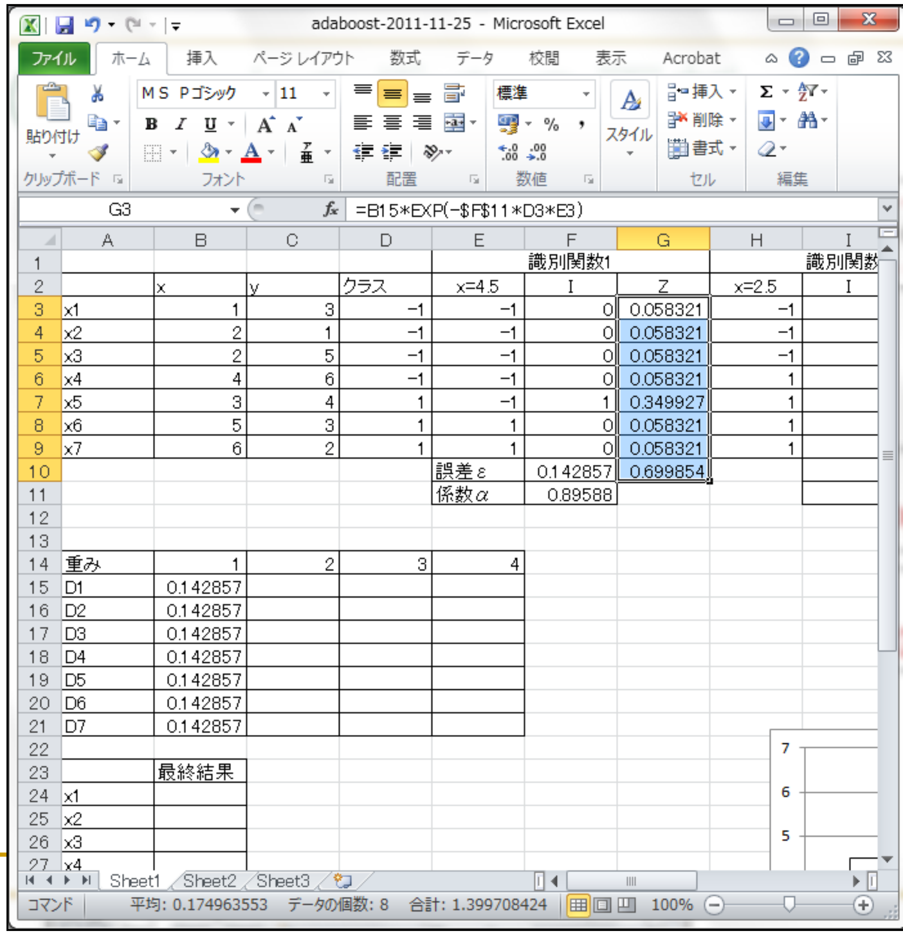
セルG3
=B15*EXP(-$F$11*D3*E3)

セルG3
セルG4～G9にコピー

⑤データの重みの更新

$$D(\mathbf{x}_i) \leftarrow \begin{cases} D(\mathbf{x}_i)e^{-\alpha_j} / Z & \text{if} \ \ f_j(\mathbf{x}_i) = y_i \\ D(\mathbf{x}_i)e^{+\alpha_j} / Z & \text{if} \ \ f_j(\mathbf{x}_i) \neq y_i \end{cases}$$

$$Z = \sum_{i=1}^{P} D(\mathbf{x}_i)\exp(-\alpha y_i f_j(\mathbf{x}_i))$$



	A	B	C	D	E	F	G	H	I
1						識別関数1			識別関数
2		x	y	クラス	x=4.5	I	Z	x=2.5	I
3	x1	1	3	-1	-1	0	0.058321	-1	
4	x2	2	1	-1	-1	0	0.058321	-1	
5	x3	2	5	-1	-1	0	0.058321	-1	
6	x4	4	6	-1	-1	0	0.058321	1	
7	x5	3	4	1	-1	1	0.349927	1	
8	x6	5	3	1	1	0	0.058321	1	
9	x7	6	2	1	1	0	0.058321	1	
10					誤差ε	0.142857	0.699854		
11					係数α	0.89588			
12									
13									
14	重み	1	2	3	4				
15	D1	0.142857							
16	D2	0.142857							
17	D3	0.142857							
18	D4	0.142857							
19	D5	0.142857							
20	D6	0.142857							
21	D7	0.142857							
22									
23		最終結果							
24	x1								
25	x2								
26	x3								
27	x4								

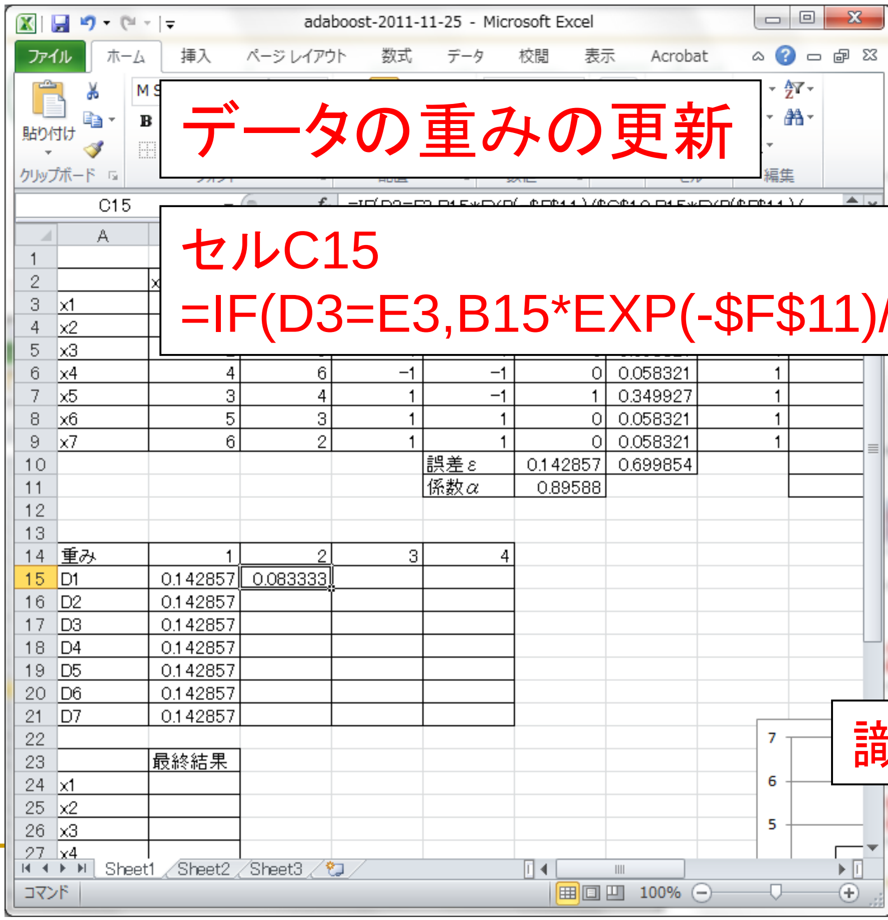
Zの計算

セルG10
=SUM(G3:G9)

⑤データの重みの更新

$$D(\mathbf{x}_i) \leftarrow \begin{cases} D(\mathbf{x}_i)e^{-\alpha_j}/Z & \text{if } \ f_j(\mathbf{x}_i)=y_i \\ D(\mathbf{x}_i)e^{+\alpha_j}/Z & \text{if } \ f_j(\mathbf{x}_i)\neq y_i \end{cases}$$

$$Z=\sum_{i=1}^{P} D(\mathbf{x}_i)\exp(-\alpha y_i f_j(\mathbf{x}_i))$$



データの重みの更新

セルC15
=IF(D3=E3,B15*EXP(-$F$11)/$G$10,B15*EXP($F$11)/$G$10)

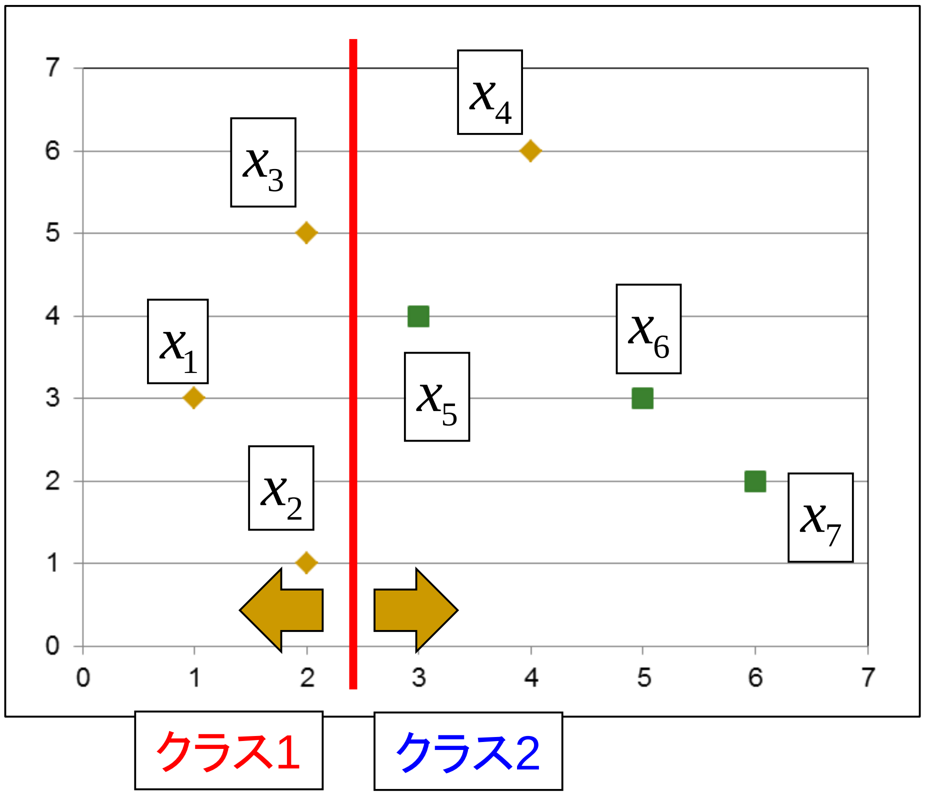
セルC15
セルC16～C21にコピー

識別関数2について②～⑤を繰り返す

②（繰り返し）識別関数の構築（選択）*

識別関数2

$$f_2(\mathbf{x}_i) = \begin{cases} -1 & \text{if } x_1 < 2.5 \\ +1 & \text{if } x_1 > 2.5 \end{cases}$$



識別関数2による結果

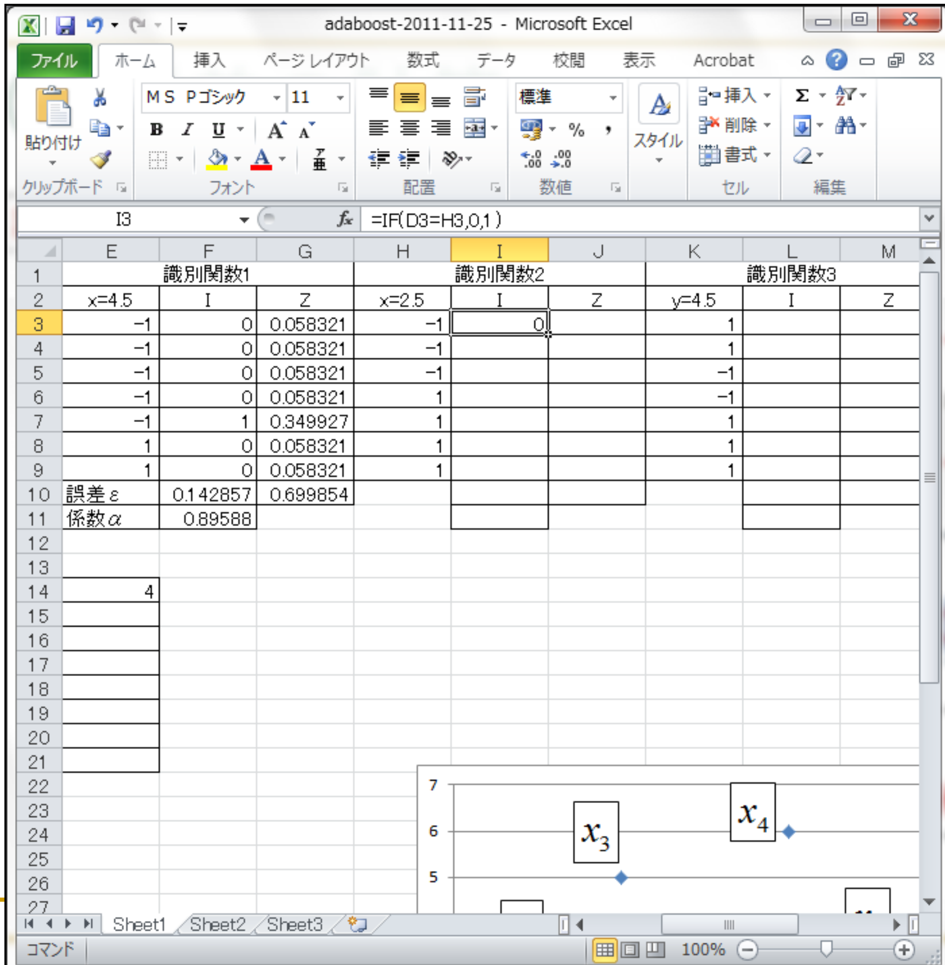
	x_1	x_2	$f_3(x)$
x_1	1	3	-1
x_2	2	1	-1
x_3	2	5	-1
x_4	4	6	+1
x_5	3	4	+1
x_6	5	3	+1
x_7	6	2	+1

*残り2個の識別関数の中から$\varepsilon(f_i)$が最小の識別関数を選択する，この場合，$\varepsilon(f_2)=0.083$，$\varepsilon(f_3)=0.167$より識別関数2とする

③ 誤差の計算

識別関数2の誤差

$$\varepsilon(f_j) = \sum_{i=1}^{P} D(\mathbf{x}_i) I(y_i \neq f_j(\mathbf{x}_i))$$



	E	F	G	H	I	J	K	L	M
1	識別関数1			識別関数2			識別関数3		
2	x=4.5	I	Z	x=2.5	I	Z	y=4.5	I	Z
3	−1	0	0.058321	−1	0		1		
4	−1	0	0.058321	−1			1		
5	−1	0	0.058321	−1			−1		
6	−1	0	0.058321	1			−1		
7	−1	1	0.349927	1			1		
8	1	0	0.058321	1			1		
9	1	0	0.058321	1			1		
10	誤差ε	0.142857	0.699854						
11	係数α	0.89588							
14	4								

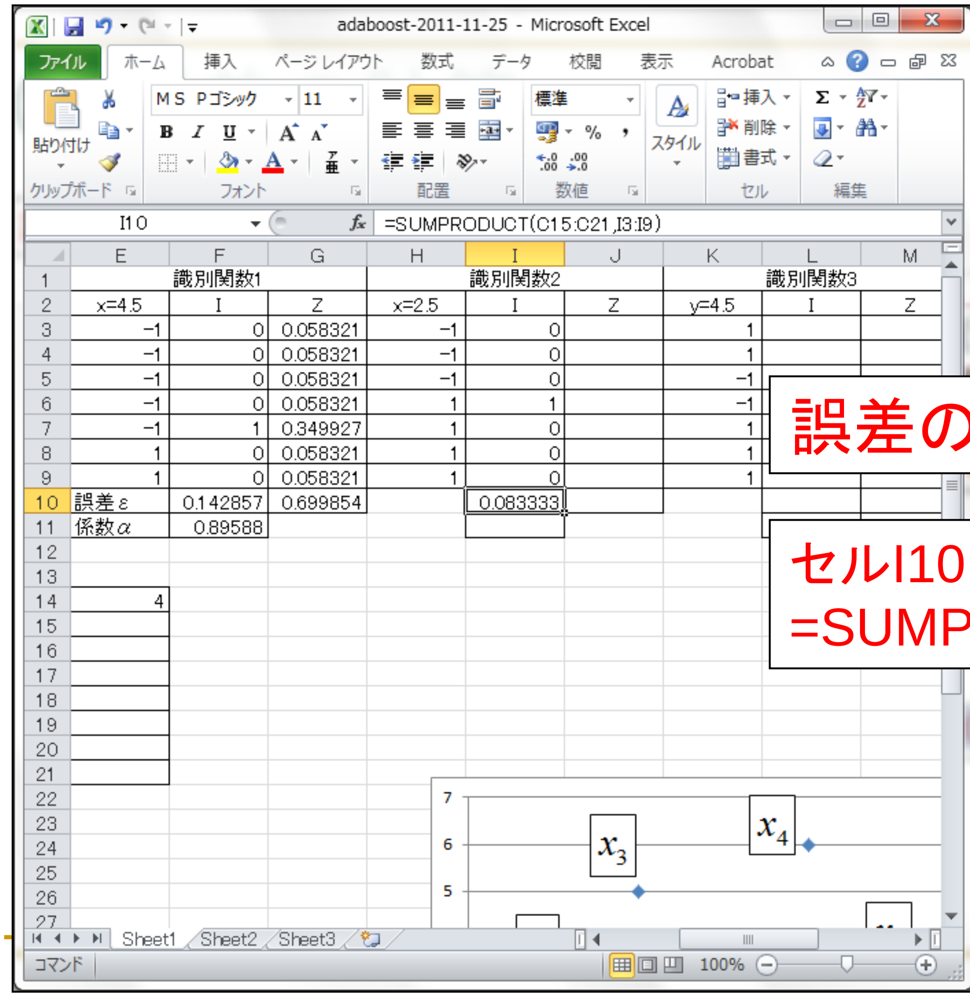
$$I(y_i \neq f_j(\mathbf{x}_i)) = \begin{cases} 1 & y_i \neq f_j(\mathbf{x}_i) \\ 0 & y_i = f_j(\mathbf{x}_i) \end{cases}$$

セルI3
=IF(D3=H3,0,1)

セルI3
セルI4～I9にコピー

③ 誤差の計算

識別関数2の誤差



	E	F	G	H	I	J	K	L	M
1		識別関数1			識別関数2			識別関数3	
2	x=4.5	I	Z	x=2.5	I	Z	y=4.5	I	Z
3	−1	0	0.058321	−1	0		1		
4	−1	0	0.058321	−1	0		1		
5	−1	0	0.058321	−1	0		−1		
6	−1	0	0.058321	1	1		−1		
7	−1	1	0.349927	1	0		1		
8	1	0	0.058321	1	0		1		
9	1	0	0.058321	1	0		1		
10	誤差ε	0.142857	0.699854		0.083333				
11	係数α	0.89588							
14	4								

$$\varepsilon(f_j) = \sum_{i=1}^{P} \boxed{D(\mathbf{x}_i)}\,\boxed{I(y_i \neq f_j(\mathbf{x}_i))}$$

セルC15～C21

セルI3～I9

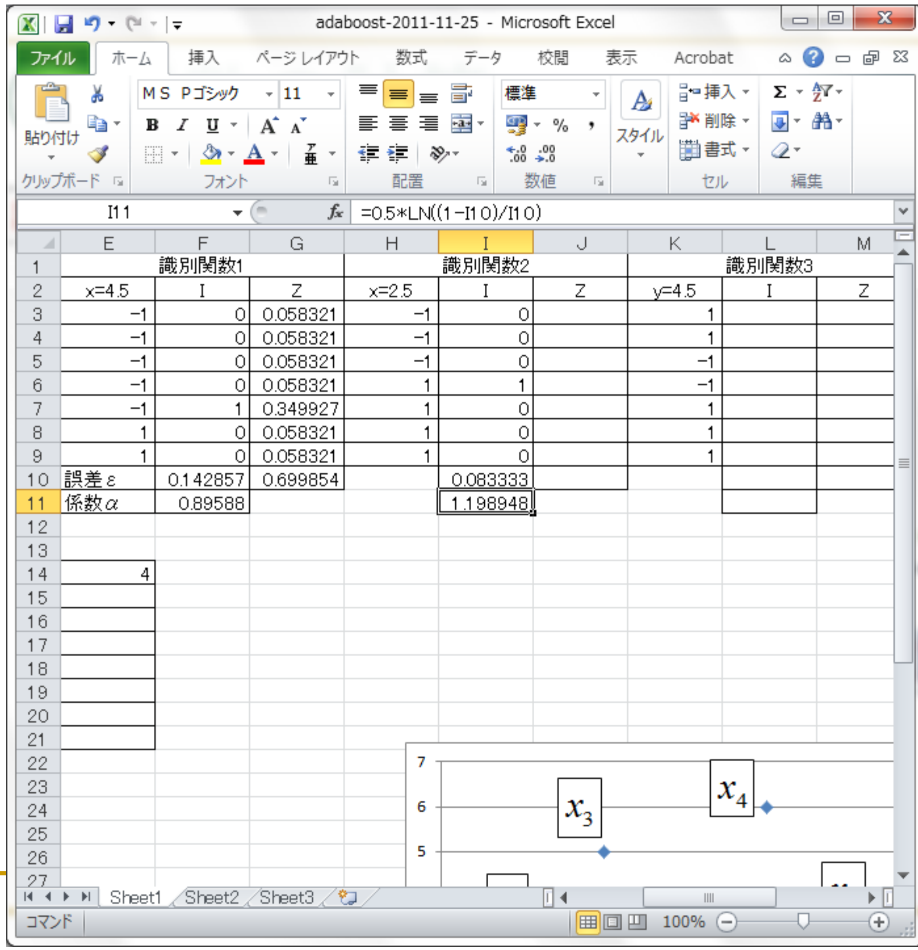
誤差の計算

セルI10
=SUMPRODUCT(C15:C21,I3:I9)

④ 係数の計算

識別関数2の係数

$$\alpha_j \leftarrow \frac{1}{2}\ln\left(\frac{1-\varepsilon(f_j)}{\varepsilon(f_j)}\right)$$



	E	F	G	H	I	J	K	L	M
1		識別関数1			識別関数2			識別関数3	
2	x=4.5	I	Z	x=2.5	I	Z	y=4.5	I	Z
3	-1	0	0.058321	-1	0		1		
4	-1	0	0.058321	-1	0		1		
5	-1	0	0.058321	-1	0		-1		
6	-1	0	0.058321	1	1		-1		
7	-1	1	0.349927	1	0		1		
8	1	0	0.058321	1	0		1		
9	1	0	0.058321	1	0		1		
10	誤差ε	0.142857	0.699854		0.083333				
11	係数α	0.89588			1.198948				
14	4								

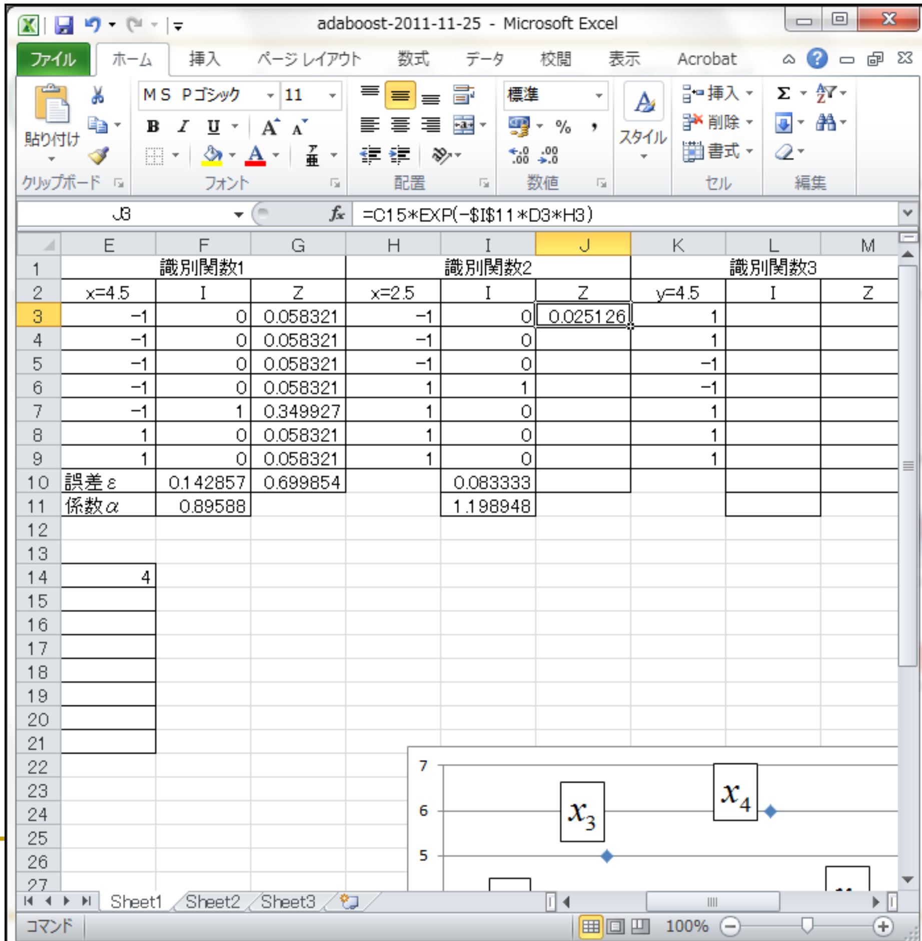
係数の計算

セルI11
=0.5*LN((1-I10)/I10)

⑤データの重みの更新

$$D(\mathbf{x}_i) \leftarrow \begin{cases} D(\mathbf{x}_i)e^{-\alpha_j}/Z & \text{if} \quad f_j(\mathbf{x}_i) = y_i \\ D(\mathbf{x}_i)e^{+\alpha_j}/Z & \text{if} \quad f_j(\mathbf{x}_i) \neq y_i \end{cases}$$

$$Z = \sum_{i=1}^{P} \boxed{D(\mathbf{x}_i)\exp(-\alpha y_i f_j(\mathbf{x}_i))}$$



	E	F	G	H	I	J	K	L	M
1		識別関数1			識別関数2			識別関数3	
2	x=4.5	I	Z	x=2.5	I	Z	y=4.5	I	Z
3	-1	0	0.058321	-1	0	0.025126	1		
4	-1	0	0.058321	-1	0		1		
5	-1	0	0.058321	-1	0		-1		
6	-1	0	0.058321	1	1		-1		
7	-1	1	0.349927	1	0		1		
8	1	0	0.058321	1	0		1		
9	1	0	0.058321	1	0		1		
10	誤差ε	0.142857	0.699854		0.083333				
11	係数α	0.89588			1.198948				
14	4								

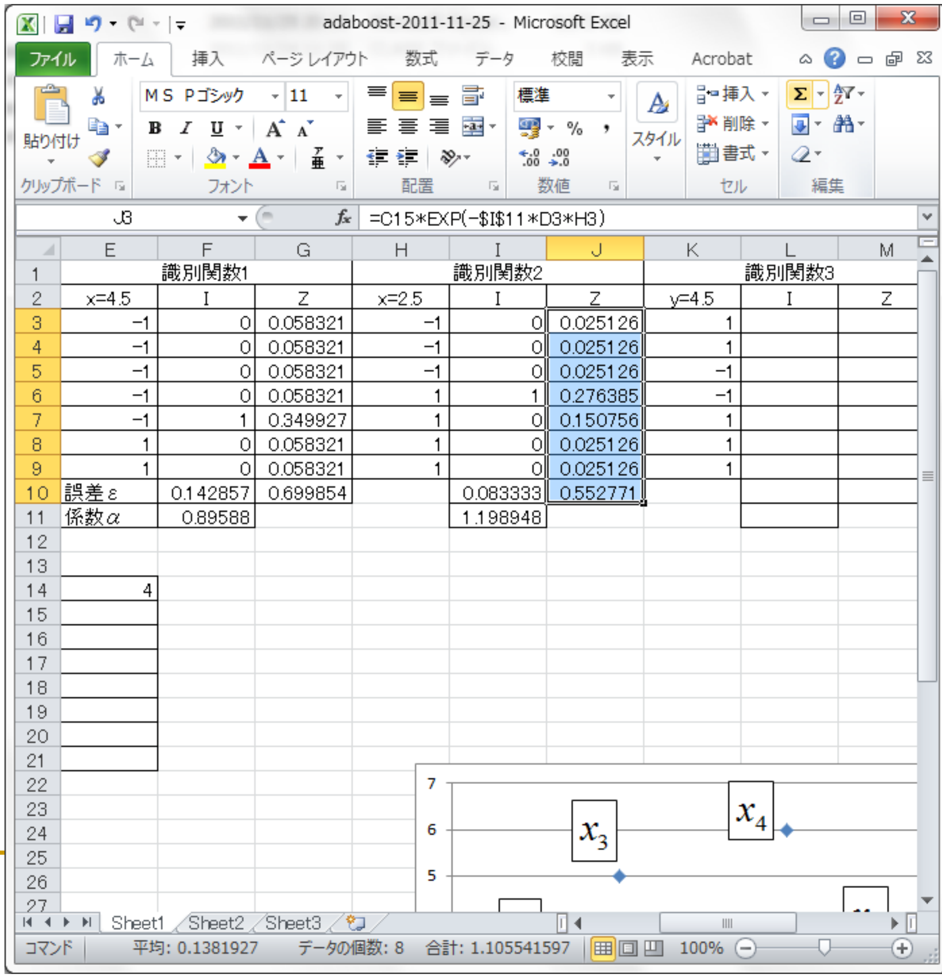
セルJ3
=C15*EXP(-$I$11*D3*H3)

セルJ3
セルJ4～J9にコピー

⑤データの重みの更新

$$D(\mathbf{x}_i) \leftarrow \begin{cases} D(\mathbf{x}_i)e^{-\alpha_j} / Z & \text{if} \;\; f_j(\mathbf{x}_i) = y_i \\ D(\mathbf{x}_i)e^{+\alpha_j} / Z & \text{if} \;\; f_j(\mathbf{x}_i) \neq y_i \end{cases}$$

$$Z = \sum_{i=1}^{P} D(\mathbf{x}_i)\exp(-\alpha y_i f_j(\mathbf{x}_i))$$



	E	F	G	H	I	J	K	L	M
1		識別関数1			識別関数2			識別関数3	
2	x=4.5	I	Z	x=2.5	I	Z	y=4.5	I	Z
3	−1	0	0.058321	−1	0	0.025126	1		
4	−1	0	0.058321	−1	0	0.025126	1		
5	−1	0	0.058321	−1	0	0.025126	−1		
6	−1	0	0.058321	1	1	0.276385	−1		
7	−1	1	0.349927	1	0	0.150756	1		
8	1	0	0.058321	1	0	0.025126	1		
9	1	0	0.058321	1	0	0.025126	1		
10	誤差ε	0.142857	0.699854		0.083333	0.552771			
11	係数α	0.89588			1.198948				
14	4								

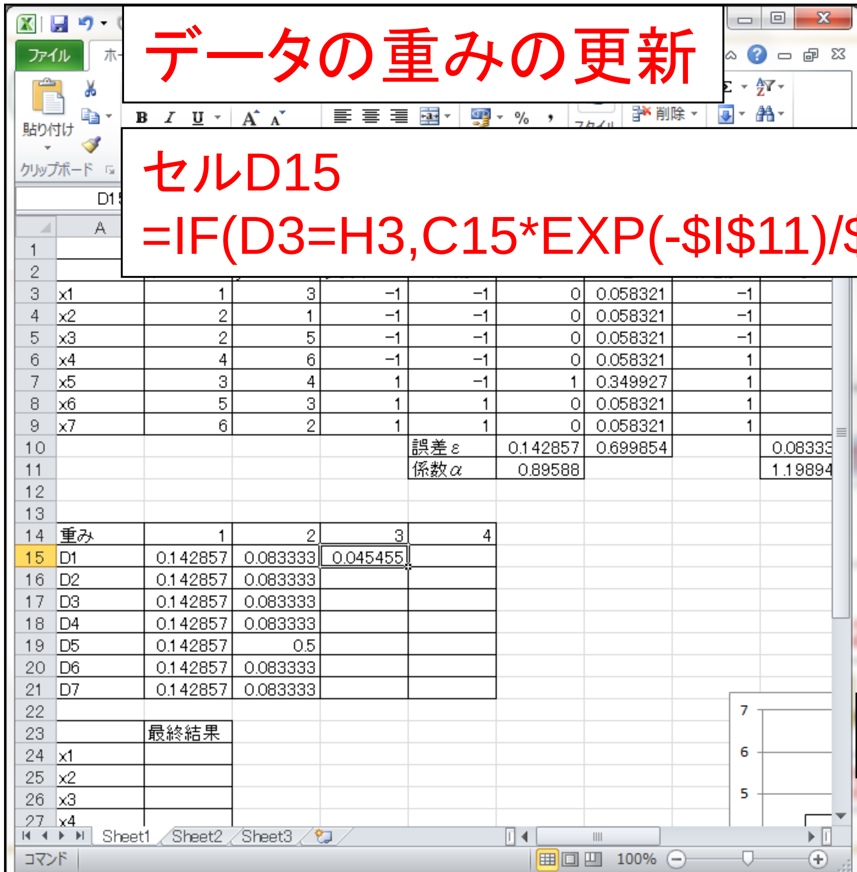
Zの計算

セルJ10
=SUM(J3:J9)

⑤データの重みの更新

$$D(\mathbf{x}_i) \leftarrow \begin{cases} D(\mathbf{x}_i) e^{-\alpha_j} / Z & \text{if} \quad f_j(\mathbf{x}_i) = y_i \\ D(\mathbf{x}_i) e^{+\alpha_j} / Z & \text{if} \quad f_j(\mathbf{x}_i) \neq y_i \end{cases}$$

$$Z = \sum_{i=1}^{P} D(\mathbf{x}_i) \exp(-\alpha y_i f_j(\mathbf{x}_i))$$



	A	B	C	D	E	F	G	H	I
3	x1	1	3	-1	-1	0	0.058321	-1	
4	x2	2	1	-1	-1	0	0.058321	-1	
5	x3	2	5	-1	-1	0	0.058321	-1	
6	x4	4	6	-1	-1	0	0.058321	1	
7	x5	3	4	1	-1	1	0.349927	1	
8	x6	5	3	1	1	0	0.058321	1	
9	x7	6	2	1	1	0	0.058321	1	
10					誤差ε	0.142857	0.699854		0.08333
11					係数α	0.89588			1.19894
14	重み	1	2	3	4				
15	D1	0.142857	0.083333	0.045455					
16	D2	0.142857	0.083333						
17	D3	0.142857	0.083333						
18	D4	0.142857	0.083333						
19	D5	0.142857	0.5						
20	D6	0.142857	0.083333						
21	D7	0.142857	0.083333						
23		最終結果							
24	x1								
25	x2								
26	x3								
27	x4								

データの重みの更新

セルD15
=IF(D3=H3,C15*EXP(-$I$11)/$J$10,C15*EXP($I$11)/$J$10)

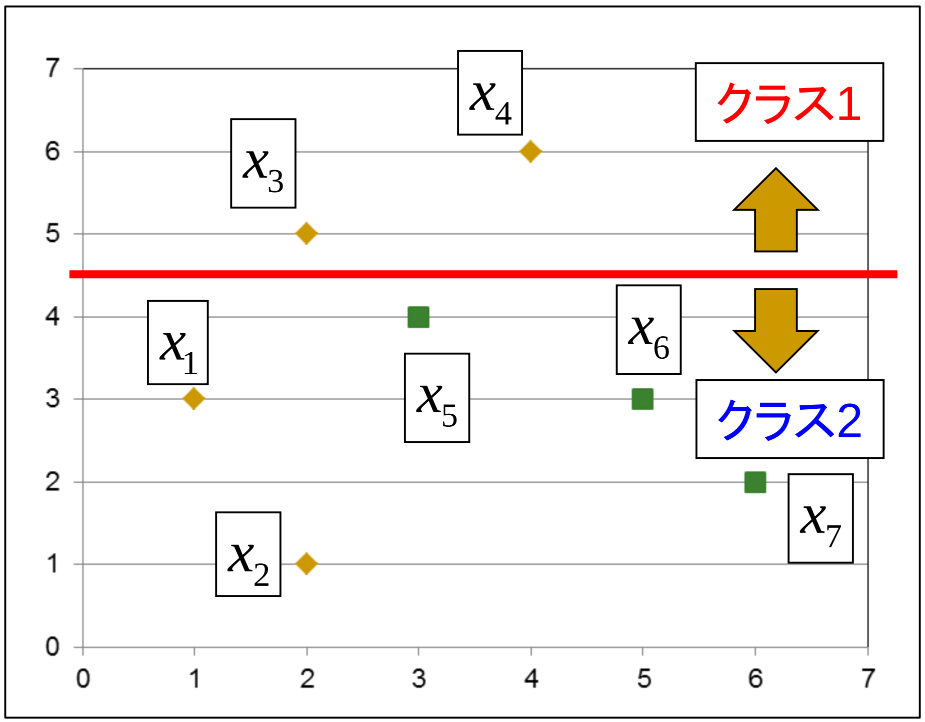
セルD15
セルD16～D21にコピー

識別関数3について②～⑤を繰り返す

② (さらに繰り返し) 識別関数の選択*

識別関数3

$$f_3(\mathbf{x}_i) = \begin{cases} -1 & \text{if } \ x_2 < 4.5 \\ +1 & \text{if } \ x_2 > 4.5 \end{cases}$$



識別関数3による結果

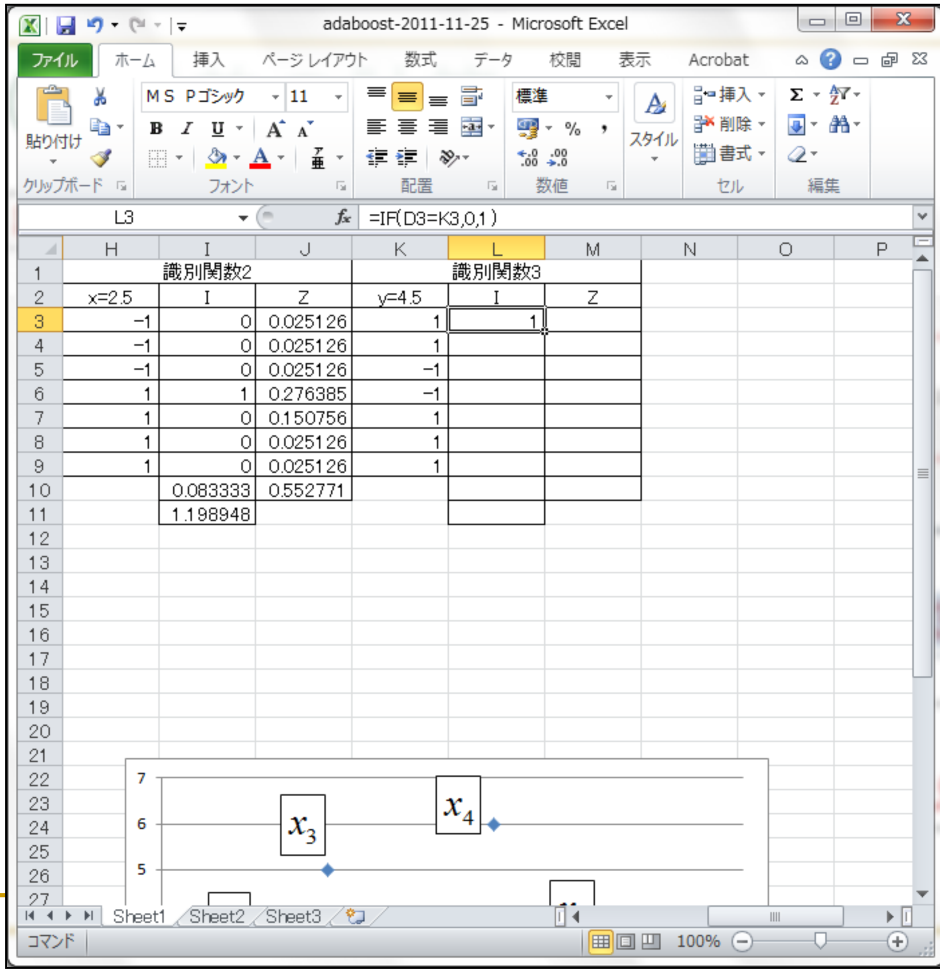
	x_1	x_2	$f_2(x)$
x_1	1	3	+1
x_2	2	1	+1
x_3	2	5	-1
x_4	4	6	-1
x_5	3	4	+1
x_6	5	3	+1
x_7	6	2	+1

*最後に残っている識別関数3を選択する

③ 誤差の計算

識別関数3の誤差

$$\varepsilon(f_j) = \sum_{i=1}^{P} D(\mathbf{x}_i) I(y_i \neq f_j(\mathbf{x}_i))$$



	H	I	J	K	L	M
1	識別関数2			識別関数3		
2	x=2.5	I	Z	y=4.5	I	Z
3	-1	0	0.025126	1	1	
4	-1	0	0.025126	1		
5	-1	0	0.025126	-1		
6	1	1	0.276385	-1		
7	1	0	0.150756	1		
8	1	0	0.025126	1		
9	1	0	0.025126	1		
10		0.083333	0.552771			
11		1.198948				

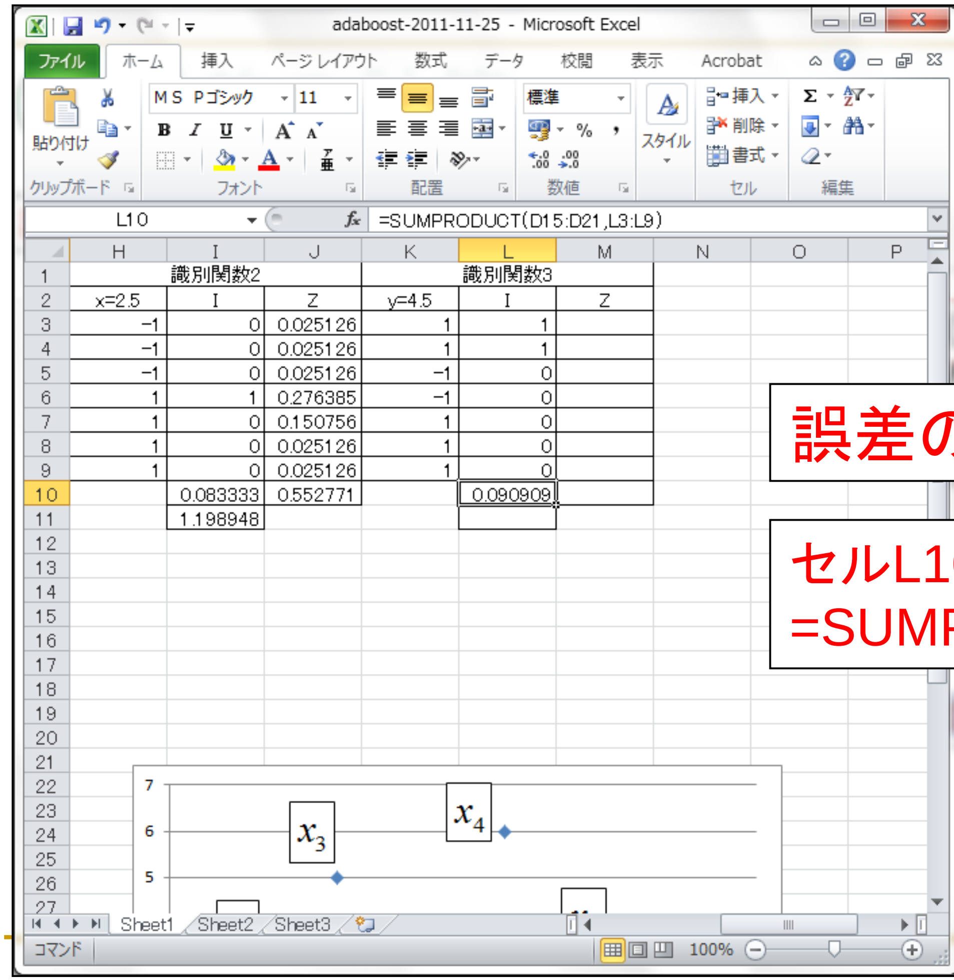
$$I(y_i \neq f_j(\mathbf{x}_i)) = \begin{cases} 1 & y_i \neq f_j(\mathbf{x}_i) \\ 0 & y_i = f_j(\mathbf{x}_i) \end{cases}$$

セルL3
=IF(D3=K3,0,1)

セルL3
セルL4～L9にコピー

③ 誤差の計算

識別関数3の誤差


adaboost-2011-11-25 - Microsoft Excel

L10 =SUMPRODUCT(D15:D21,L3:L9)

	H	I	J	K	L	M
1	識別関数2			識別関数3		
2	x=2.5	I	Z	y=4.5	I	Z
3	−1	0	0.025126	1	1	
4	−1	0	0.025126	1	1	
5	−1	0	0.025126	−1	0	
6	1	1	0.276385	−1	0	
7	1	0	0.150756	1	0	
8	1	0	0.025126	1	0	
9	1	0	0.025126	1	0	
10		0.083333	0.552771		0.090909	
11		1.198948				

$$\varepsilon(f_j) = \sum_{i=1}^{P} \boxed{D(\mathbf{x}_i)}\,\boxed{I(y_i \neq f_j(\mathbf{x}_i))}$$

セルD15～D21 セルL3～L9

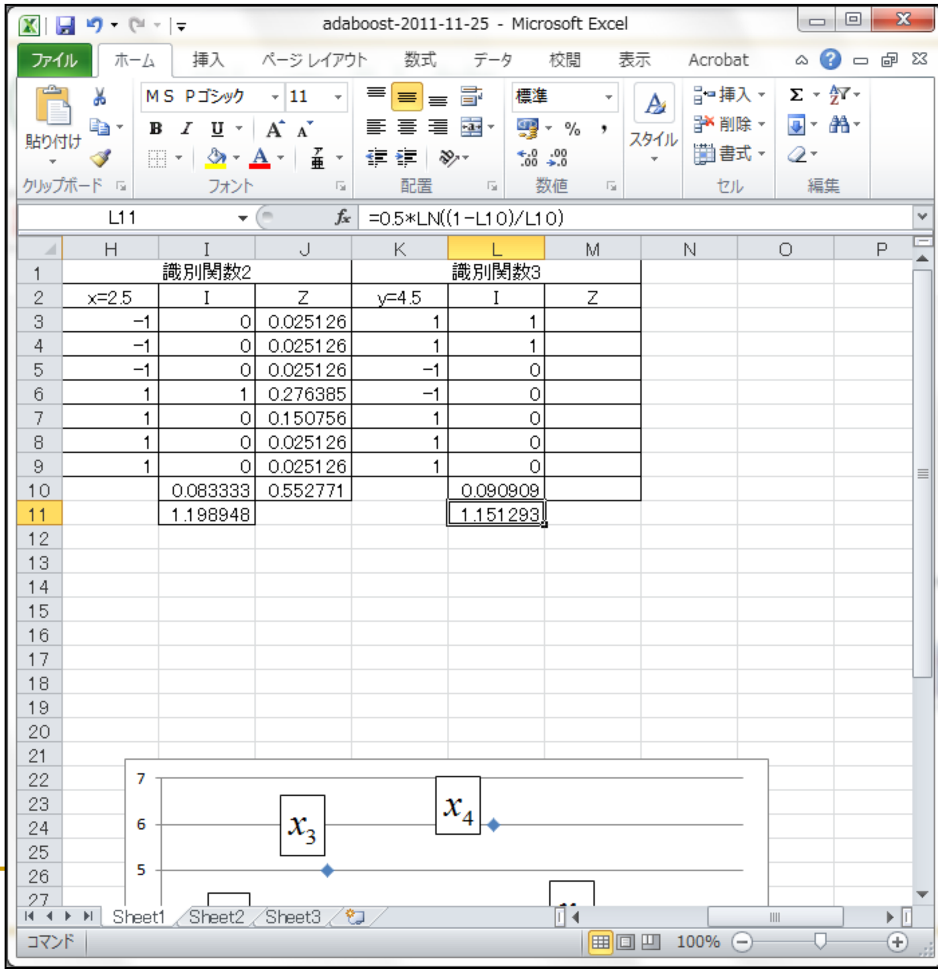
誤差の計算

セルL10
=SUMPRODUCT(D15:D21,L3:L9)

④ 係数の計算

識別関数2の係数

$$\alpha_j \leftarrow \frac{1}{2}\ln\left(\frac{1-\varepsilon(f_j)}{\varepsilon(f_j)}\right)$$



	H	I	J	K	L	M
1	識別関数2			識別関数3		
2	x=2.5	I	Z	y=4.5	I	Z
3	-1	0	0.025126	1	1	
4	-1	0	0.025126	1	1	
5	-1	0	0.025126	-1	0	
6	1	1	0.276385	-1	0	
7	1	0	0.150756	1	0	
8	1	0	0.025126	1	0	
9	1	0	0.025126	1	0	
10		0.083333	0.552771		0.090909	
11		1.198948			1.151293	

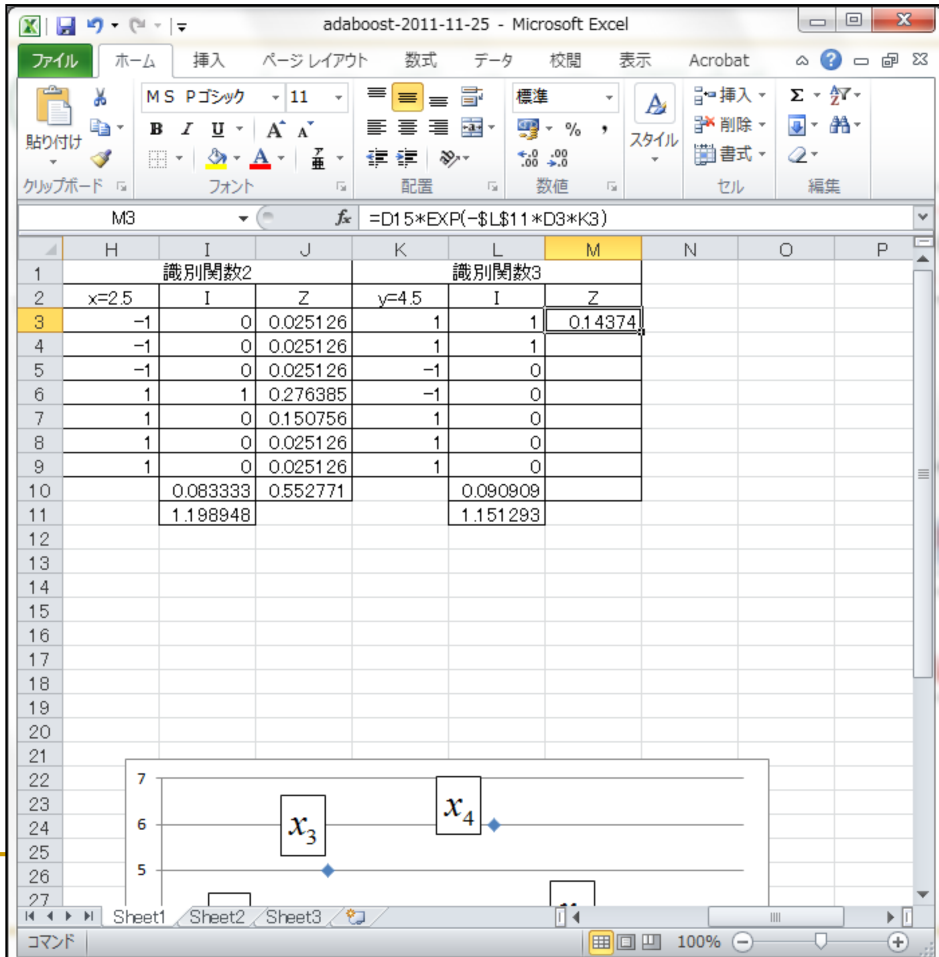
係数の計算

セルL11
=0.5*LN((1-L10)/L10)

⑤データの重みの更新

$$D(\mathbf{x}_i) \leftarrow \begin{cases} D(\mathbf{x}_i)e^{-\alpha_j} / Z \ \ \text{if} \ \ f_j(\mathbf{x}_i) = y_i \\ D(\mathbf{x}_i)e^{+\alpha_j} / Z \ \ \text{if} \ \ f_j(\mathbf{x}_i) \neq y_i \end{cases}$$

$$Z = \sum_{i=1}^{P} \boxed{D(\mathbf{x}_i)\exp(-\alpha y_i f_j(\mathbf{x}_i))}$$



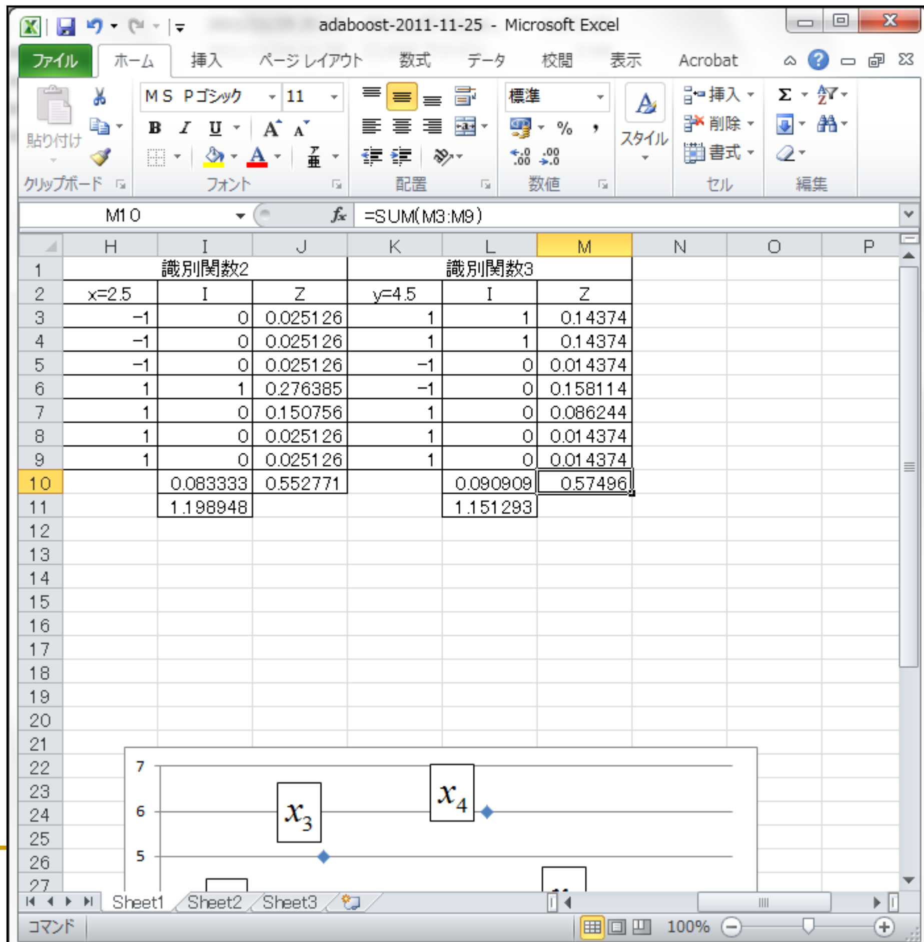
セルM3
=D15*EXP(-$L$11*D3*K3)

セルM3
セルM4～M9にコピー

⑤データの重みの更新

$$D(\mathbf{x}_i) \leftarrow \begin{cases} D(\mathbf{x}_i) e^{-\alpha_j} / Z & \text{if} \quad f_j(\mathbf{x}_i) = y_i \\ D(\mathbf{x}_i) e^{+\alpha_j} / Z & \text{if} \quad f_j(\mathbf{x}_i) \neq y_i \end{cases}$$

$$Z = \sum_{i=1}^{P} D(\mathbf{x}_i) \exp(-\alpha y_i f_j(\mathbf{x}_i))$$



	H	I	J	K	L	M
1	識別関数2			識別関数3		
2	x=2.5	I	Z	y=4.5	I	Z
3	-1	0	0.025126	1	1	0.14374
4	-1	0	0.025126	1	1	0.14374
5	-1	0	0.025126	-1	0	0.014374
6	1	1	0.276385	-1	0	0.158114
7	1	0	0.150756	1	0	0.086244
8	1	0	0.025126	1	0	0.014374
9	1	0	0.025126	1	0	0.014374
10		0.083333	0.552771		0.090909	0.57496
11		1.198948			1.151293	

誤差の計算

セルM10
=SUM(M3:M9)

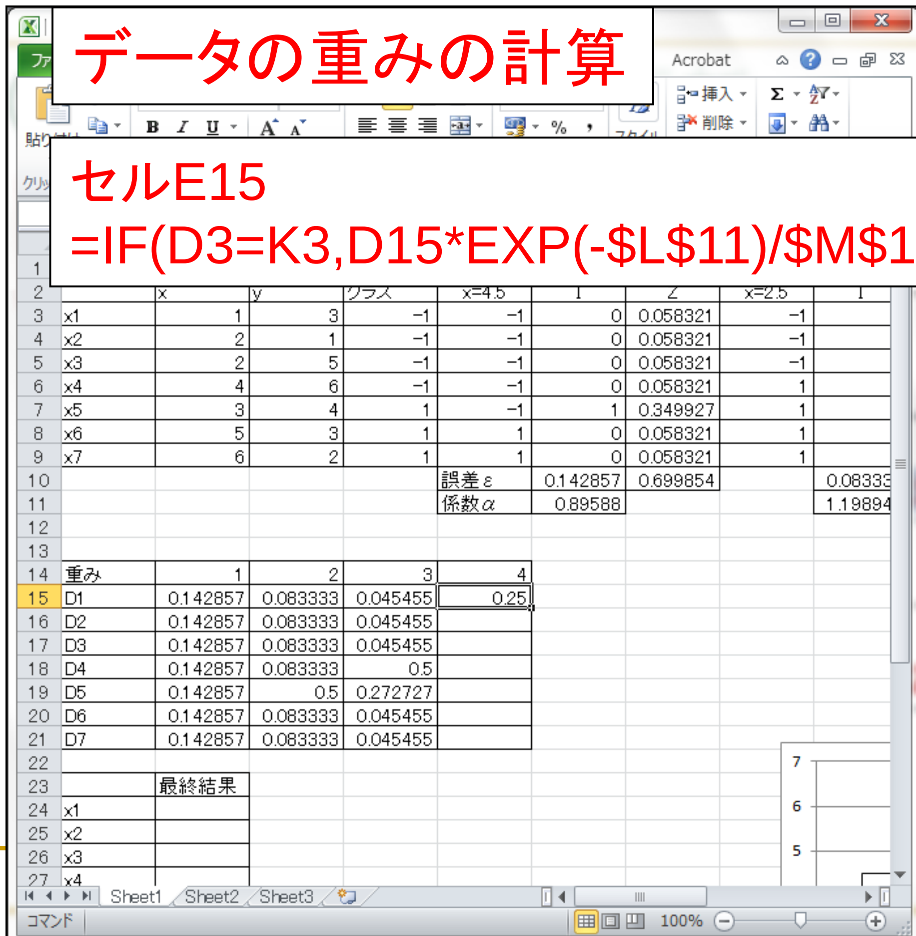
⑤データの重みの更新

$$D(\mathbf{x}_i) \leftarrow \begin{cases} D(\mathbf{x}_i)e^{-\alpha_j}/Z & \text{if} \quad f_j(\mathbf{x}_i) = y_i \\ D(\mathbf{x}_i)e^{+\alpha_j}/Z & \text{if} \quad f_j(\mathbf{x}_i) \neq y_i \end{cases}$$

$$Z = \sum_{i=1}^{P} D(\mathbf{x}_i)\exp(-\alpha y_i f_j(\mathbf{x}_i))$$

データの重みの計算

セルE15
=IF(D3=K3,D15*EXP(-$L$11)/$M$10,D15*EXP($L$11)/$M$10)

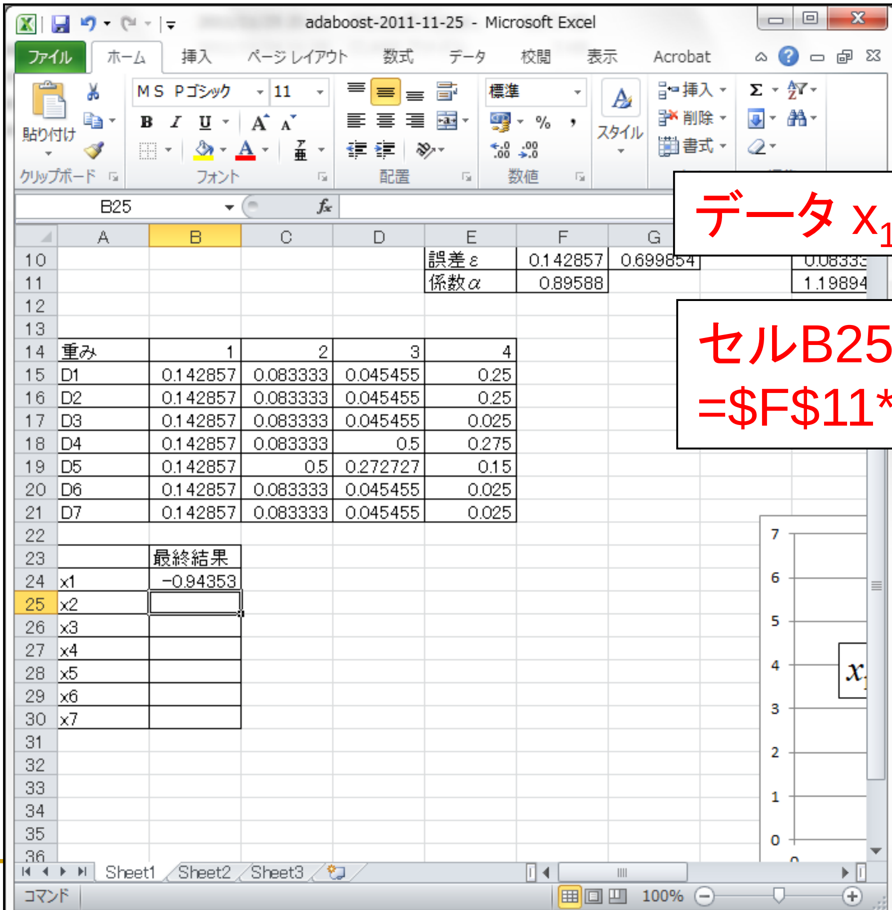


	A	B	C	D	E	F	G	H	I
2		x	y	クラス	x=4.5	I	Z	x=2.5	I
3	x1	1	3	−1	−1	0	0.058321	−1	
4	x2	2	1	−1	−1	0	0.058321	−1	
5	x3	2	5	−1	−1	0	0.058321	−1	
6	x4	4	6	−1	−1	0	0.058321	1	
7	x5	3	4	1	−1	1	0.349927	1	
8	x6	5	3	1	1	0	0.058321	1	
9	x7	6	2	1	1	0	0.058321	1	
10					誤差ε	0.142857	0.699854		0.08333
11					係数α	0.89588			1.19894
12									
13									
14	重み	1	2	3	4				
15	D1	0.142857	0.083333	0.045455	0.25				
16	D2	0.142857	0.083333	0.045455					
17	D3	0.142857	0.083333	0.045455					
18	D4	0.142857	0.083333	0.5					
19	D5	0.142857	0.5	0.272727					
20	D6	0.142857	0.083333	0.045455					
21	D7	0.142857	0.083333	0.045455					
22									
23		最終結果							
24	x1								
25	x2								
26	x3								
27	x4								

セルE15
セルE16～E21にコピー

⑥最終的な識別関数を求める

$$F(\mathbf{x}) = \mathrm{sgn}\left(\boxed{\sum_{i=1}^{n} \alpha_i f_i(\mathbf{x})} \right)$$



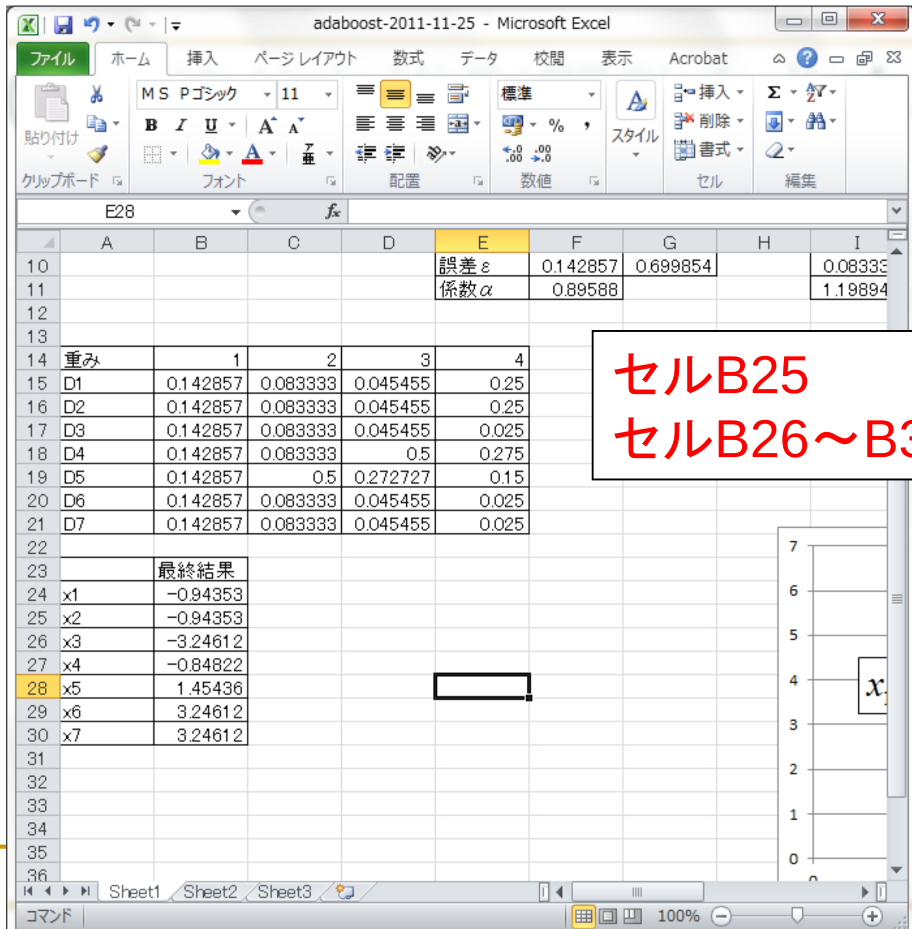
	A	B	C	D	E	F	G
10					誤差ε	0.142857	0.699854
11					係数α	0.89588	
12							
13							
14	重み	1	2	3	4		
15	D1	0.142857	0.083333	0.045455	0.25		
16	D2	0.142857	0.083333	0.045455	0.25		
17	D3	0.142857	0.083333	0.045455	0.025		
18	D4	0.142857	0.083333	0.5	0.275		
19	D5	0.142857	0.5	0.272727	0.15		
20	D6	0.142857	0.083333	0.045455	0.025		
21	D7	0.142857	0.083333	0.045455	0.025		
22							
23		最終結果					
24	x1	-0.94353					
25	x2						
26	x3						
27	x4						
28	x5						
29	x6						
30	x7						

データ x_1の判定

セルB25
=$F$11*E3+$I$11*H3+$L$11*K3

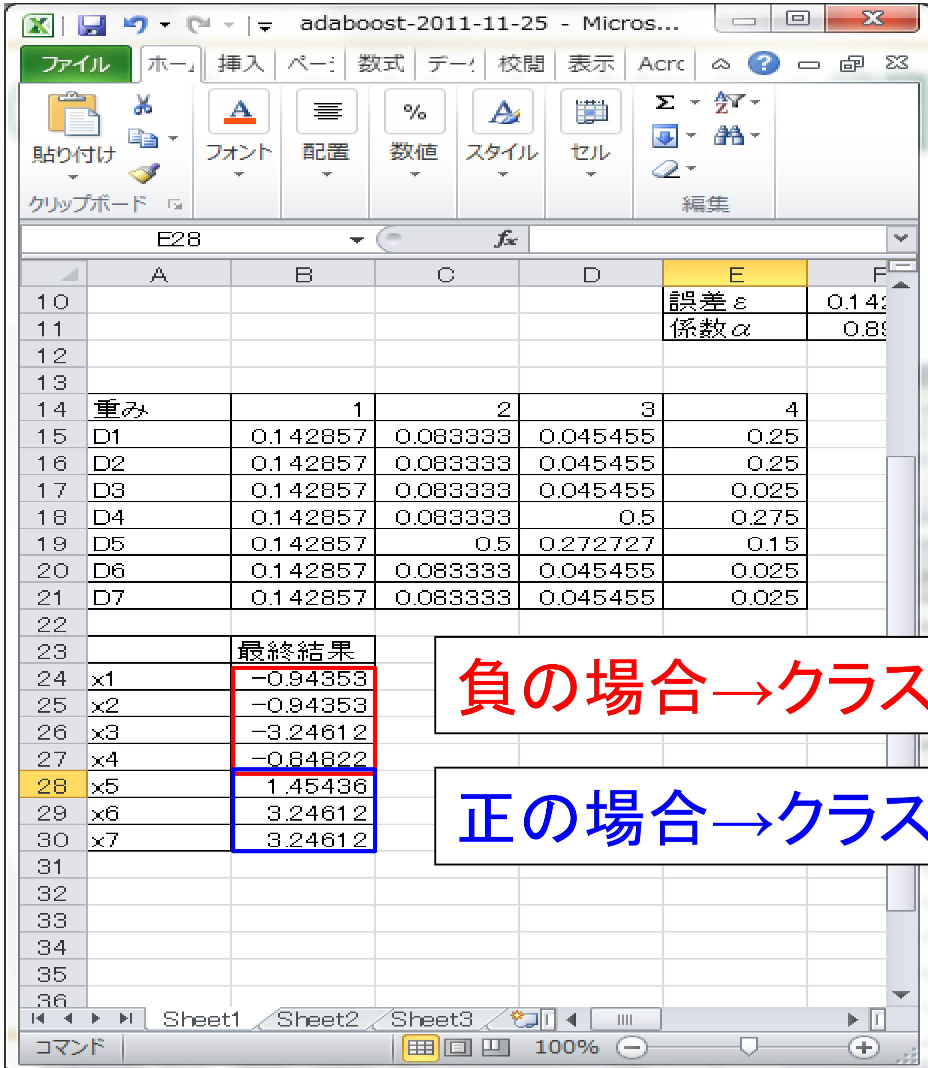
⑥最終的な識別関数を求める

$$F(\mathbf{x}) = \mathrm{sgn}\left(\boxed{\sum_{i=1}^{n} \alpha_i f_i(\mathbf{x})}\right) = \mathrm{sgn}\big(0.856 \times f_1(\mathbf{x}) + 1.199 \times f_2(\mathbf{x}) + 1.151 \times f_3(\mathbf{x})\big)$$



セルB25
セルB26～B30にコピー

最終結果



行	A	B	C	D	E	F
10					誤差ε	0.14
11					係数α	0.8
14	重み	1	2	3	4	
15	D1	0.142857	0.083333	0.045455	0.25	
16	D2	0.142857	0.083333	0.045455	0.25	
17	D3	0.142857	0.083333	0.045455	0.025	
18	D4	0.142857	0.083333	0.5	0.275	
19	D5	0.142857	0.5	0.272727	0.15	
20	D6	0.142857	0.083333	0.045455	0.025	
21	D7	0.142857	0.083333	0.045455	0.025	
23		最終結果				
24	x1	−0.94353				
25	x2	−0.94353				
26	x3	−3.24612				
27	x4	−0.84822				
28	x5	1.45436				
29	x6	3.24612				
30	x7	3.24612				

負の場合→クラス1

正の場合→クラス2

アダブーストの手順

N個の識別関数

データの重み

$$D(\mathbf{x}_i) = \frac{1}{P}$$

$\mathbf{x}_i, D(\mathbf{x}_i)$ → 識別関数f_1 → ε_1, α_1

重みの更新

$\mathbf{x}_i, D(\mathbf{x}_i)$ → 識別関数f_2 → ε_2, α_2

重みの更新

$\mathbf{x}_i, D(\mathbf{x}_i)$ → 識別関数f_3 → ε_3, α_3

$\mathbf{x}_i, D(\mathbf{x}_i)$ → 識別関数f_n → ε_N, α_N

最終的な識別関数

$$F(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{N} \alpha_i f_i(\mathbf{x}) \right)$$

識別関数f_jを選択

→ 係数α_jを決定

データの重みの更新

正しく識別できた場合

$$D(\mathbf{x}_i) \leftarrow D(\mathbf{x}_i) e^{-\alpha_j} / Z \ \text{ if } \ f_j(\mathbf{x}_i) = y_i$$

⇒ データの重みを小さくする

正しく識別できなかった場合

$$D(\mathbf{x}_i) \leftarrow D(\mathbf{x}_i) e^{+\alpha_j} / Z \ \text{ if } \ f_j(\mathbf{x}_i) \neq y_i$$

⇒ データの重みを大きくする

⇓

次回は $\varepsilon(f_i)$を最小にする識別関数f_iを選択する

⇕

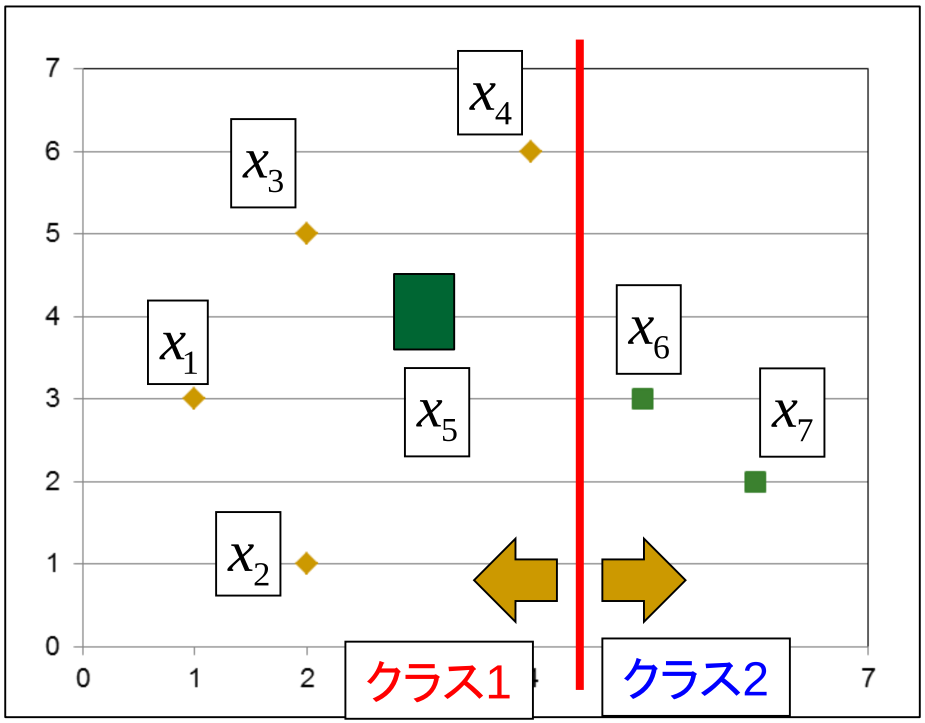
$$\varepsilon(f_j) = \sum_{i=1}^{P} D(\mathbf{x}_i) I(y_i \neq f_j(\mathbf{x}_i))$$

次回は，今回識別できなかったデータを識別しやすい識別関数を選択する

データの重みの更新

識別関数1

$$f_1(\mathbf{x}_i) = \begin{cases} -1 & \text{if } x_1 < 4.5 \\ +1 & \text{if } x_1 > 4.5 \end{cases}$$



次回は，x_5を間違えずに識別できる識別関数を選択

データの重みの大きさに比例して図示

係数の更新

$$\alpha_j \leftarrow \frac{1}{2}\ln\left(\frac{1-\varepsilon(f_j)}{\varepsilon(f_j)}\right)$$

誤差が小さい→係数は大きい

誤差が大きい→係数は小さい

$$F(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{N} \alpha_i f_i(\mathbf{x})\right)$$

誤差が小さい→係数は大きく，影響も大きい

誤差が大きい→係数は小さく，影響は少ない

アダブーストの損失関数

損失関数

$$\boxed{l(yF(\mathbf{x}))}$$

$$\boxed{l(a) = \exp(-a)}$$

教師信号　識別関数

正解の場合

y	F(x)	yF(x)	損失関数
+1	+1	+1	小さい
−1	−1	+1	小さい
+1	−1	−1	大きい
−1	+1	−1	大きい

不正解の場合

$$F(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i f_i(\mathbf{x})\right)$$

識別関数

$$l(yF(\mathbf{x})) = l\left(y\,\mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i f_i(\mathbf{x})\right)\right)$$

データxの損失関数

$$L = \sum_{j=1}^{P} l(y_j F(\mathbf{x}_j)) = \sum_{j=1}^{P} l\left(y_j\,\mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i f_i(\mathbf{x}_j)\right)\right)$$

全データの損失関数

アダブーストのプログラム

Breast Cancer Dataset

アダブーストのプログラム（Cancer_AdaBoost.py）

- 乳がんの分類問題
  - Breast cancer dataset

用途	クラス分類
データ数	569
特徴量	30
目的変数	2

クラス名	データ数
malignant	212
benign	357

Cancer_AdaBoost.py

```
import numpy as np
from sklearn import datasets
from sklearn.ensemble import AdaBoostClassifier
from sklearn import tree
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, accuracy_score,
confusion_matrix

# データのロード
cancer = datasets.load_breast_cancer()

# 種類( malignant, benign )
name = cancer.target_names
label = cancer.target
```

アダブーストのために必要

弱分類器（決定木，ベイズ決定則）のために必要

breast cancerデータセットの読み込み

data 特徴量

大きさ（569,30）

特徴量

```
feature_names = cancer.feature_names
data = cancer.data
```

label : 正解ラベル
malignant → 0
benign → 1

個数
569

学習データ, テストデータ

ホールドアウト法

```
train_data, test_data, train_label, test_label = train_test_split(data, label,
test_size=0.5, random_state=None)
```

弱分類器 : 決定木（default）

```
model =
AdaBoostClassifier(base_estimator=tree.DecisionTreeClassifier(max_depth=1),
n_estimators=5)
```

base_estimator : 弱分類器
defaultは決定木

n_estimators:
弱分類器の個数

tree.DecisionTreeClassifier
決定木（深さは1）

弱分類器 : ベイズ決定則の場合

```
#wc = GaussianNB()
#model = AdaBoostClassifier(base_estimator=wc,n_estimators=5)
```

ベイズ決定則を弱分類器として用いる場合

```
# 重みの表示
print( "¥n [ 重み ]" )
for i in range(model.n_estimators):
    print( "{0} : {1:6.4f}".format( i , model.  estimator_weights_[i] ) )

# 最終予測結果
predict = model.predict( test_data )

# 弱分類器の予測確率
predict_wc = model.staged_predict_proba( test_data )

result = np.zeros( (model.n_estimators,len(test_label),2) )
for i , x in enumerate(predict_wc):
    result[i] = x
```

estimator_weights_
弱分類器ごとの重み

staged_predict_proba
弱分類器ごとの予測確率
返り値：（弱分類器の個数，データ数，クラス数）

staged_predict
弱分類器ごとの予測結果
返り値：（弱分類器の個数，データ数）

resultに弱分類器ごとの予測確率を代入

```
print( "¥n [ 予測結果 ]" )
#for i in range(len(test_label)):
for i in range(10):
    F = np.zeros( 2 )
    for j in range(model.n_estimators):
        print( "{} : ".format( j ) , end="" )
        for k in range(2):
            F[ k ] += model.estimator_weights_[j] * result[j][i][k]
            print( "{0:6.4f} ".format( result[j][i][k] ) , end="")
        print()
    print( " ------------ " )
    print( " {0:6.4f} {1:6.4f} -> {2} [ {3} ]¥n".format( F[0] , F[1] , np.argmax( F ) ,
test_label[i] ) )
```

10個のデータのみ表示

$$F(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i f_i(\mathbf{x})$$

弱分類器の重み

弱分類器ごとの予測確率

np.argmax（配列）
配列中，最大値の要素番号を返す

クラス1のF

クラス2のF

正解ラベル

```
print( " [ 予測結果 ]" )
print( classification_report(test_label, predict) )

print( "¥n [ 正解率 ]" )
print( accuracy_score(test_label, predict) )

print( "¥n [ 混同行列 ]" )
print( confusion_matrix(test_label, predict) )
```

accuracy
precision
recall
F値

accuracyの表示

混同行列の表示

AdaBoostClassifier

```
from sklearn.ensemble import AdaBoostClassifier

AdaBoostClassifier(base_estimator=tree.DecisionTreeClassifier
(max_depth=1),n_estimators=5)
```

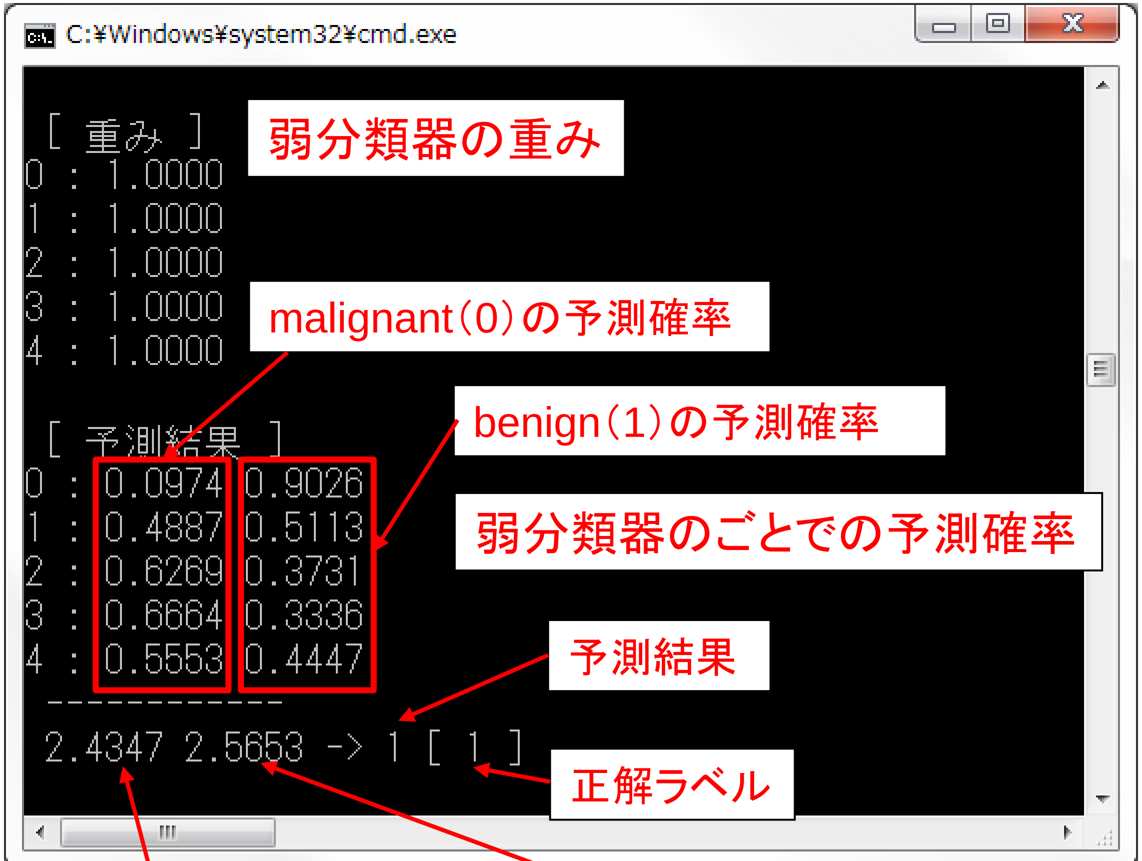
n_estimators:
弱分類器の個数

base_estimator：弱分類器
defaultは決定木（深さは1）

```
wc = GaussianNB()
model = AdaBoostClassifier(base_estimator=wc,n_estimators=5)
```

弱分類器に単純ベイズ（GaussianNB）

実行結果①



5個の弱分類器による
malignant（0）の重み付け予測確率

5個の弱分類器による
benign（1）の重み付け予測確率

実行結果②

C:¥Windows¥system32¥cmd.exe

```
[ 予測結果 ]
              precision    recall  f1-score   support

           0       0.91      0.91      0.91       105
           1       0.95      0.94      0.95       180

    accuracy                           0.93       285
   macro avg       0.93      0.93      0.93       285
weighted avg       0.93      0.93      0.93       285


[ 正解率 ]
0.9333333333333333

[ 混同行列 ]
[[ 96   9]
 [ 10 170]]
```

accuracy
precision
recall
F値

accuracyの表示

混同行列の表示

参考文献

- 加藤直樹他: データマイニングとその応用, 朝倉書店, 2009
- 平井有三: はじめてのパターン認識, 森北出版株式会社, 2012
- 後藤正幸他: 入門 パターン認識と機械学習, コロナ社, 2014
- 株式会社システム計画研究所編: Pythonによる機械学習入門, オーム社, 2016
- 竹村彰通他: 機械学習, 朝倉書店, 2017
- 荒木雅弘: 機械学習入門, 森北出版株式会社, 2018

参考文献

- AdaBoostClassifier
- https://scikit-learn.org/stable/modules/generated/
sklearn.ensemble.AdaBoostClassifier.html

- AdaBoostRegressor
- https://scikit-learn.org/stable/modules/generated/
sklearn.ensemble.AdaBoostRegressor.html#sklearn.ensemble.AdaBoostRegressor
  - 回帰も可能