ロジスティック回帰
と
ROC曲線

2017/5/19

14T4071T 渡部 勇樹

ロジスティック回帰

- 分類アルゴリズムの一つ

- 最尤推定法

最尤推定法

- 確率を利用する
- 未知のデータの属性を推定する際に、
  「このデータはt=1である」
  といった単純な推定ではなく、
  「t=1である確率は70%」
  というように、確率的な推定ができる。

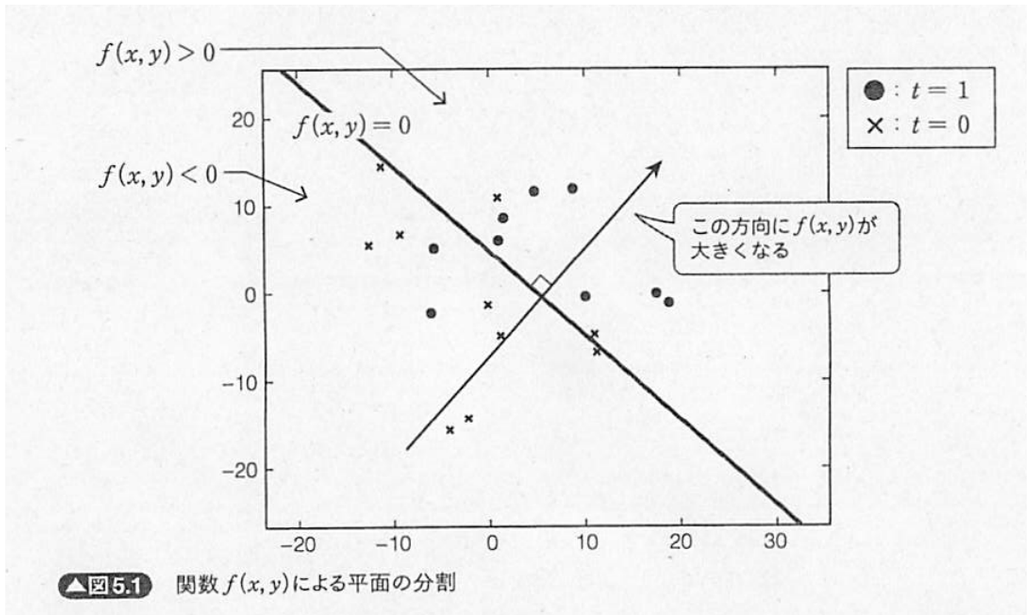
最尤推定法

1. あるデータが得られる確率を設定
2. トレーニングセットとして与えられるデータが得られる確率（尤度関数）を計算
3. 尤度関数が最大になるという条件から、最初偽停止た確率の式に含まれるパラメーターを決定

データの発生確率の設定

1. パーセプトロンと同様に、２種類のデータを分類する直線を表す一次関数を定義

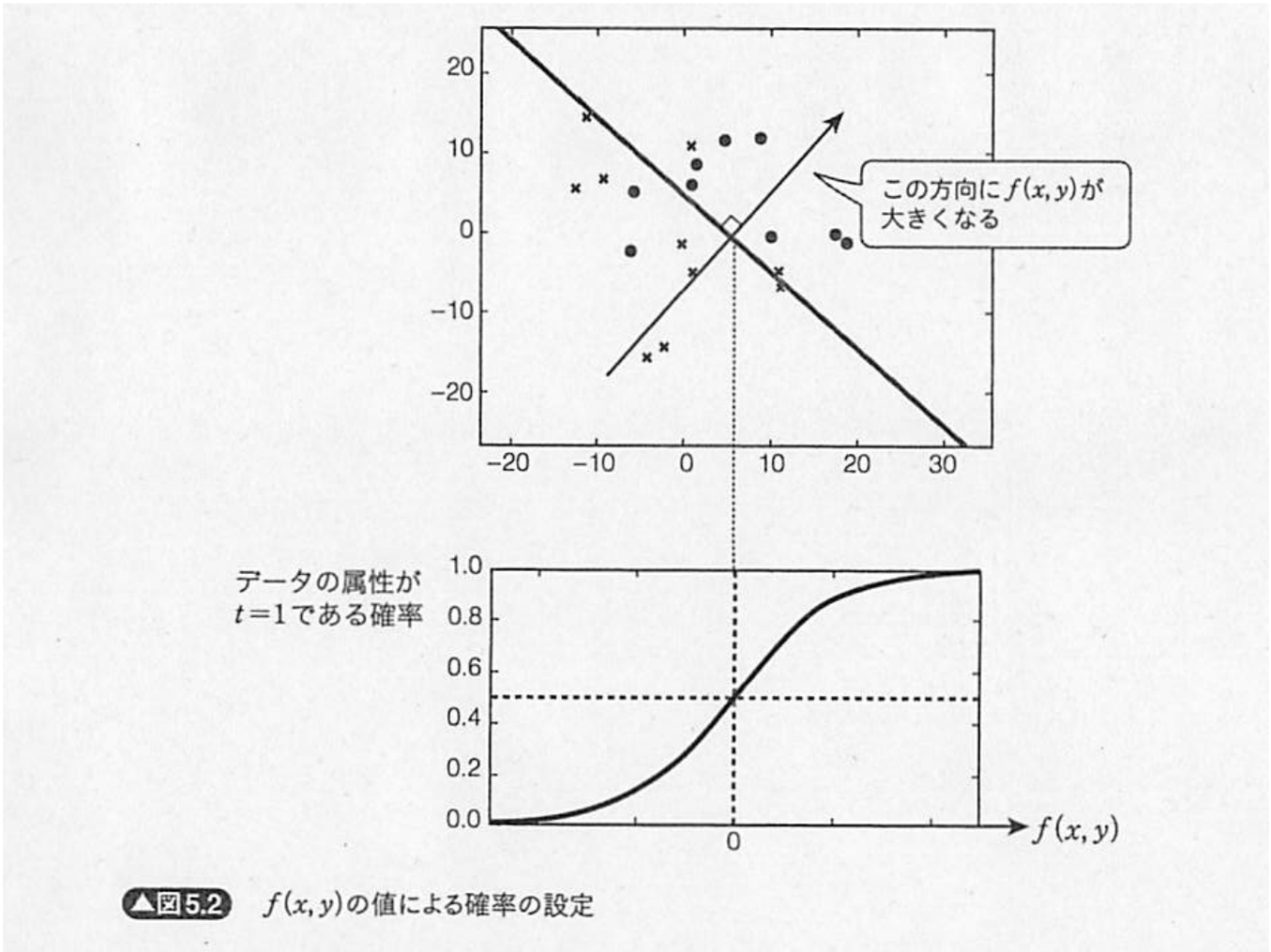
$$f(x,y) = w_0 + w_1 x + w_2 y$$



▲図 5.1 関数$f(x,y)$による平面の分割

2. (x,y)平面上の任意の点において、そこで得られたデータの属性がt=1である確率を考える

データの発生確率の設定

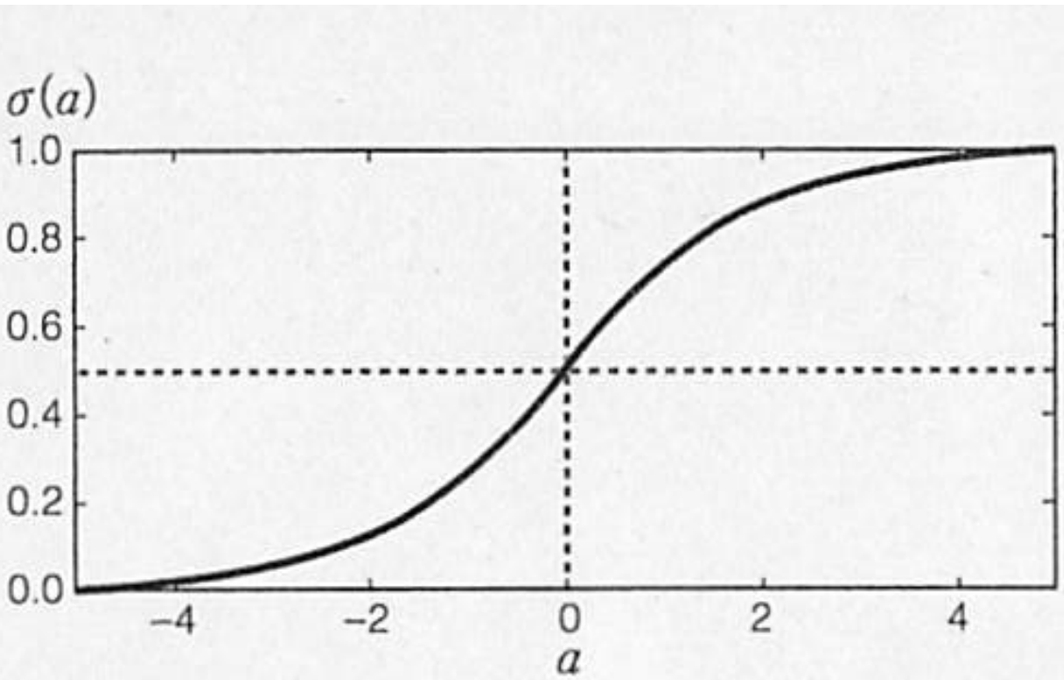


▲図5.2 $f(x,y)$の値による確率の設定

0から１に滑らかに変化するグラフ

ロジスティック関数のグラフ

ロジスティック関数：　$\sigma(a)=\dfrac{1}{1+e^{-a}}$



▲図 5.3　ロジスティック関数のグラフ

データの属性がt=1である確率

<u>点(x,y)において</u>

- データの属性がt=1である確率

$$P(x,y)=\sigma(w_0+w_1x+w_2y)$$

- データの属性がt=0である確率

$$1-P(x,y)$$

式変形

$$P_n = P(x_n, y_n)^{tn}\{1 - P(x_n, y_n)\}^{1-tn}$$

$$z_n = \sigma(W^T \phi_n)$$

n番目のデータの属性がt=1である確率

$$P_n = z_n{}^{tn}(1 - z_n)^{1-tn}$$

$$P = \prod_{n=1}^{N} P_n = \prod_{n=1}^{N} z_n{}^{t_n}(1 - z_n)^{1-t_n}$$

最尤推定法による
パラメーターの決定

$$w_{new} = w_{old} - (\phi^T R \phi)^{-1} \phi^T (z - t)$$

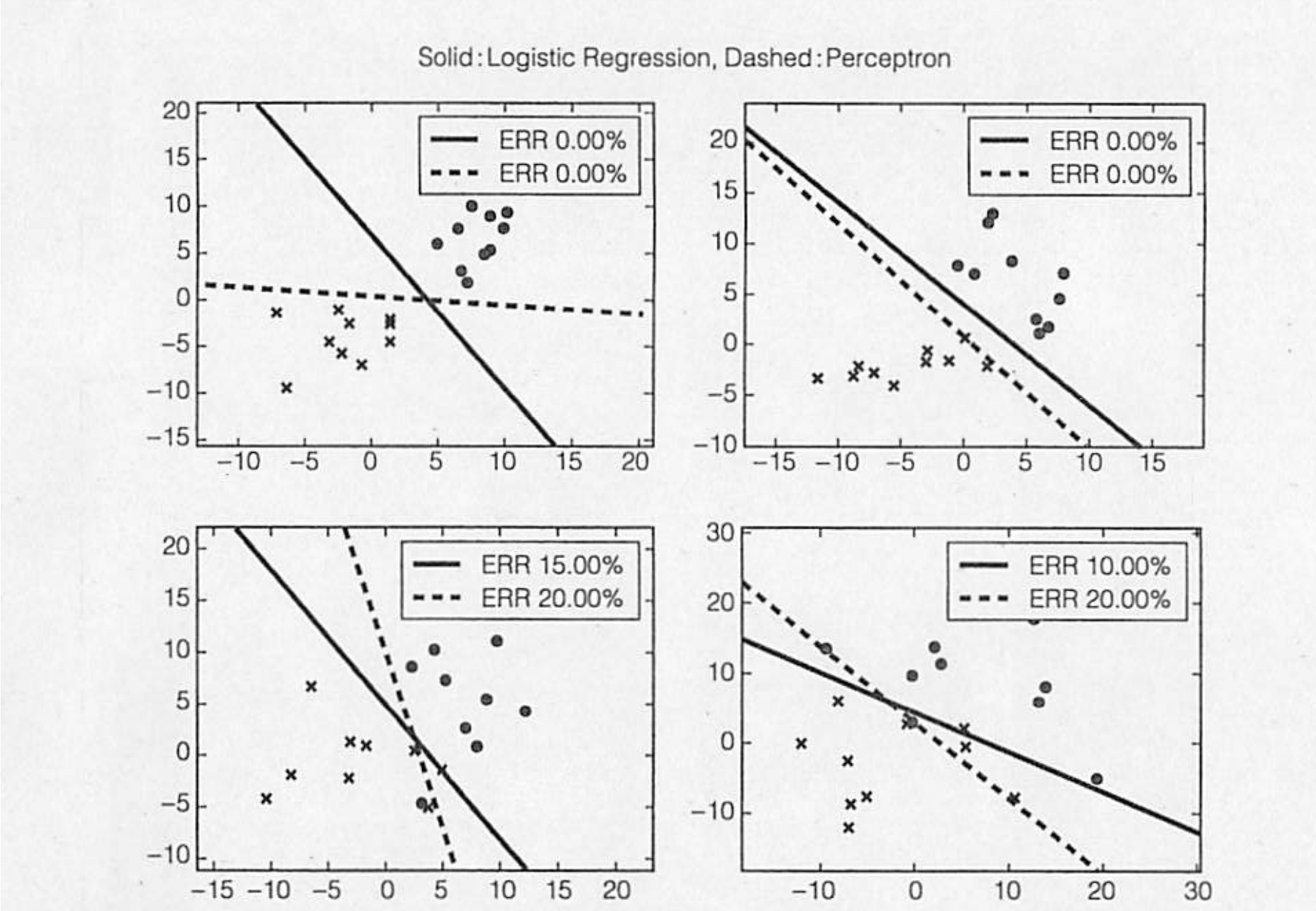
t：トレーニングセットの各データの属性値 t_n を並べたベクトル

ϕ：各データの座標を表すベクトル Φ_n を横ベクトルにして
並べたN×３行列

z： z_n を並べたベクトル

R：$z_n(1 - z_n)$ を対角成分とする対角行列

ロジスティック回帰とパーセプトロン



▲図5.5 ロジスティック回帰とパーセプトロンの比較

ROC曲線

- どのような確率を境界にするのが良いかを判断
- 機械学習に使用したアルゴリズムの良し悪しを判断
- 試験の点数から合格（T）か不合格（F）かを予測したい時
- 検査値から病気（T）か健康（F）か判断したい時

与えられた値から、真（T）か偽（F）を判定

現実の問題

・１次検査を受けた人に精密検査を進める場合

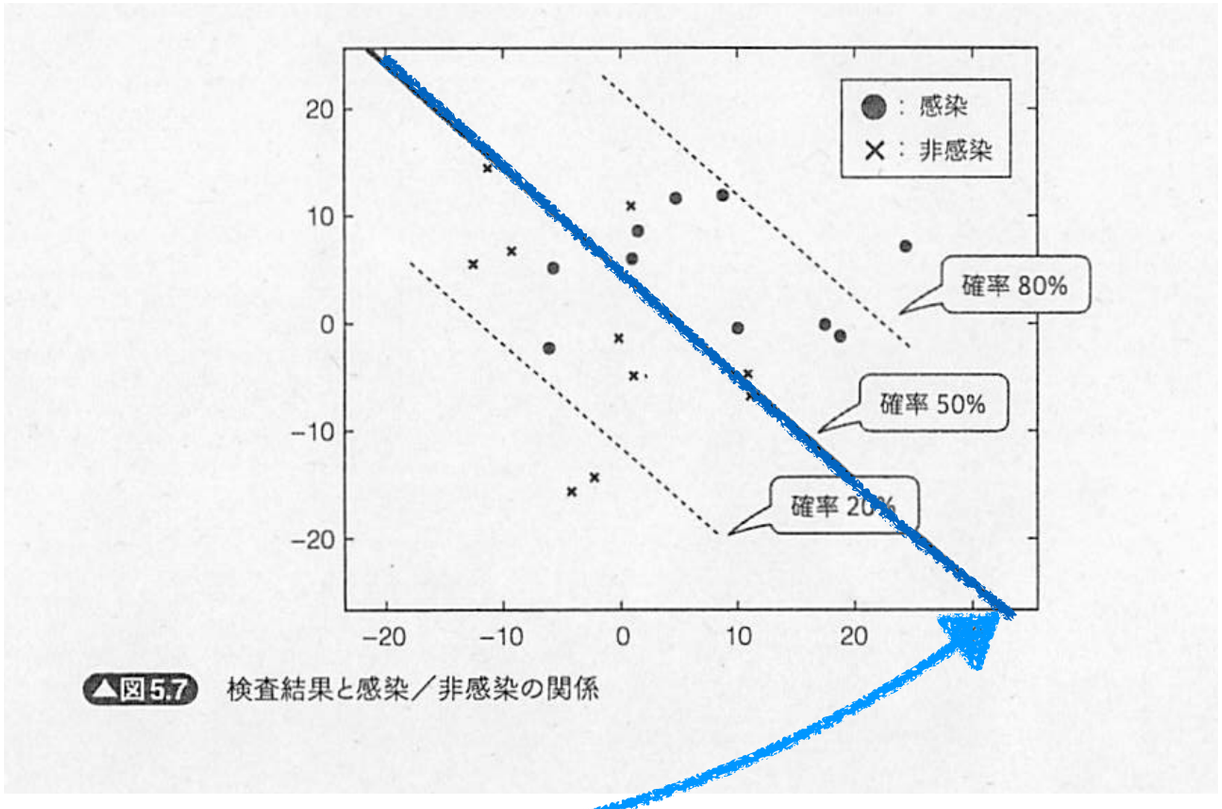
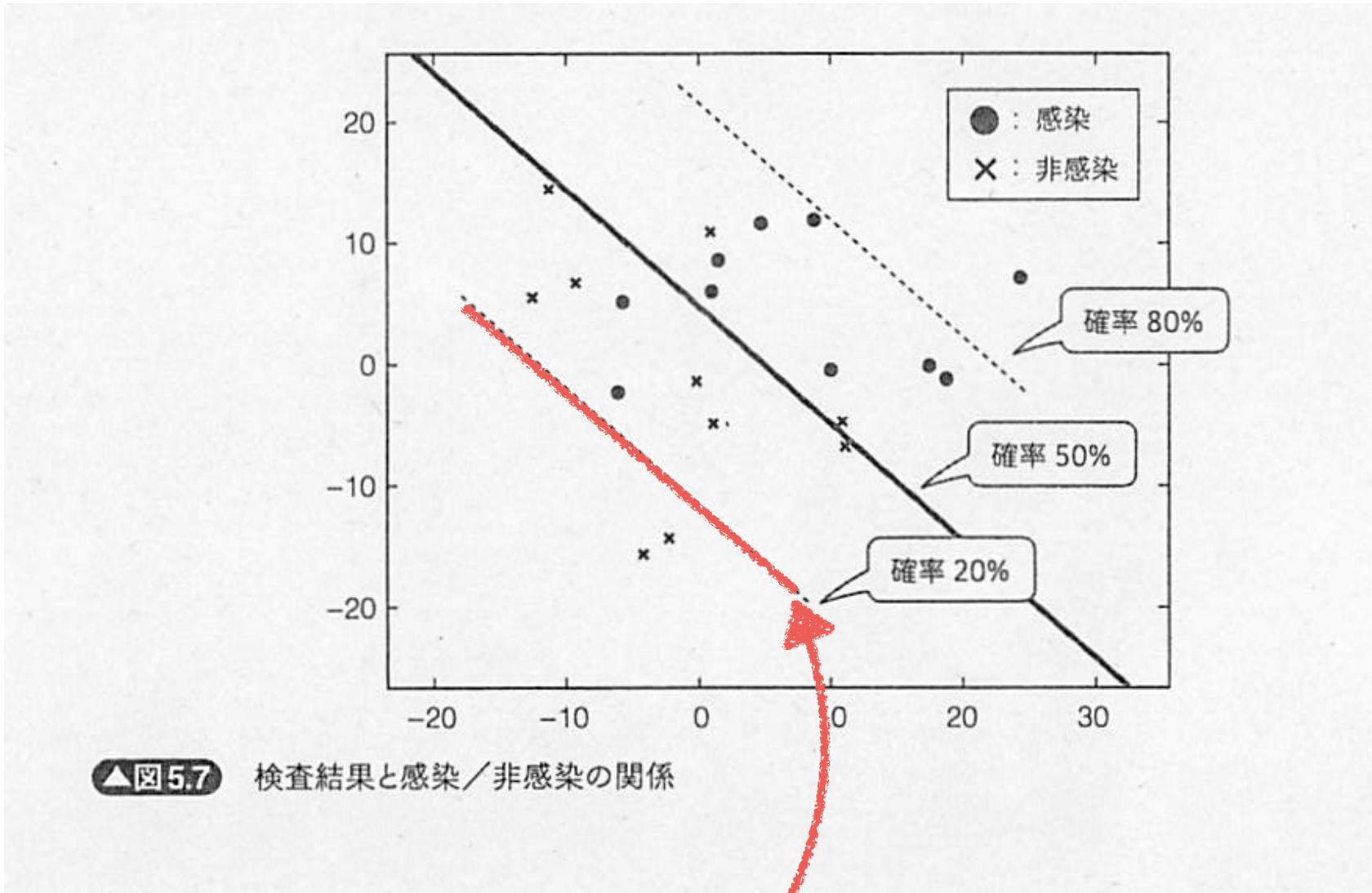


図5.7 検査結果と感染／非感染の関係

ロジスティック回帰：ウイルスに感染している確率50%

判断基準として適切ではない

現実の問題



▲図5.7 検査結果と感染／非感染の関係

精密検査を勧めるのに適している

陽性・陰性

- 陽性：発見したい属性を持つデータ
  （例）t=1

- 陰性：陽性でないデータ
  （例）t=0

陽性・陰性

- 実際に陽性→真陽性（TP：True Positive）
  判断が正しかった

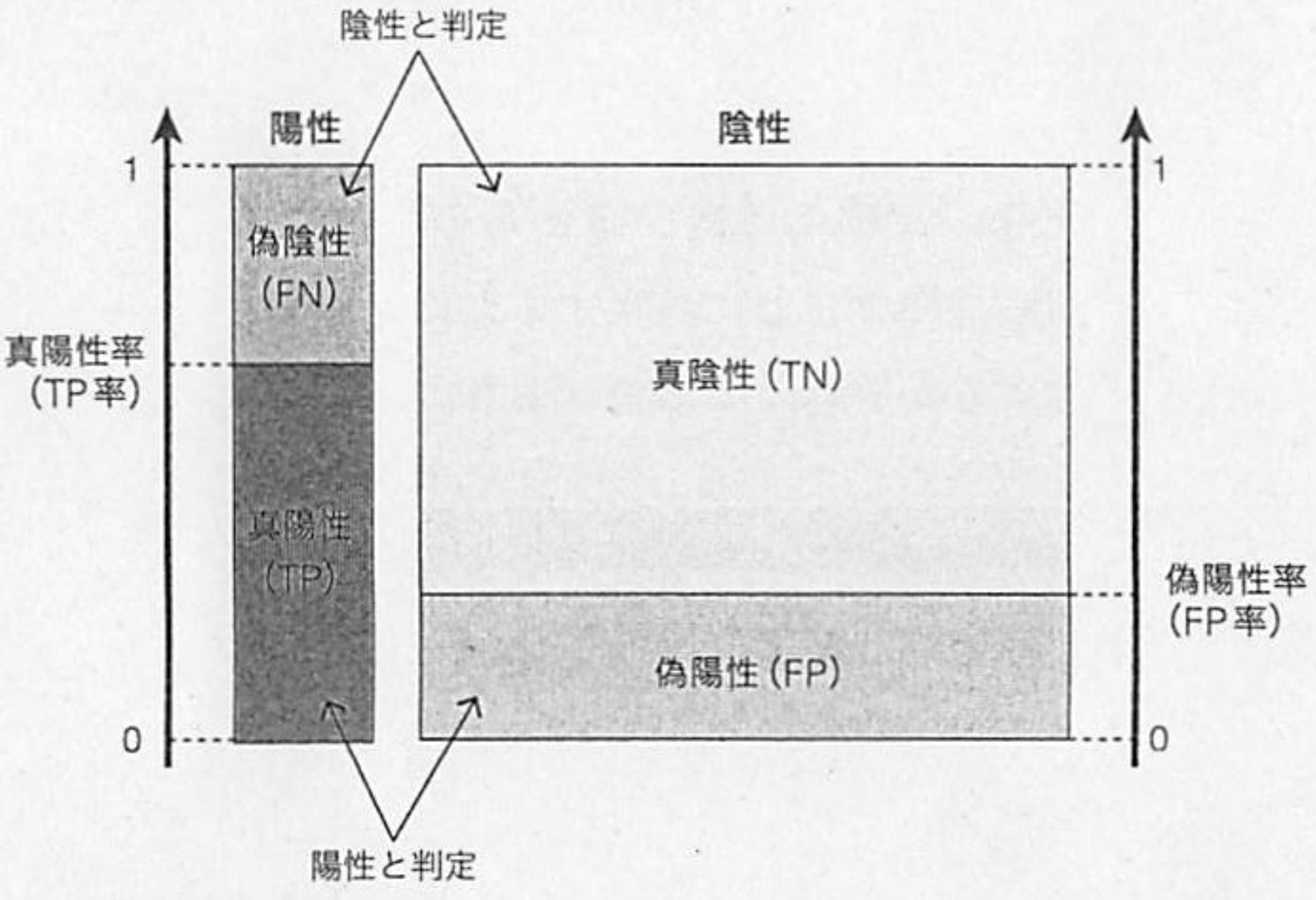
- 実際は陰性→偽陽性（FP：False Positive）
  判断が間違い

真陽性・偽陽性

- 真陽性率：実際の陽性のデータ全体の中で
  （TP率）　真陽性なるデータの割合
  （正しく判定できたものの割合）

- 偽陽性率：実際に陰性のデータ全体の中で
  （FP率）　偽陽性なるデータの割合
  （間違えて判定されたものの割合）

真陽性率



▲図5.8 真陽性率と偽陽性率の定義

現実の問題

医者の立場から見ると

- 真陽性率はできるだけ高くしたい
- 偽陽性率はできるだけ低くしたい

真陽性率と偽陽性率のトレードオフを考えながら
「判定ライン」を設定する必要がある

↓

ROCを用いて真陽性率と偽陽性率の関係を分析

ROCによる性能評価

- トレーニングセットについてロジスティック回帰を適用して、
  関数f(x,y)のパラメーター (w_0, w_1, w_2)
  を具体的に決定

- パラメーターを $P(x,y)=\sigma(w_0+w_1x+w_2y)$ に代入すると、
  座標（x,y）のデータが属性t=1を持つ確率の
  計算式P(x,y)が決まる

- 計算式からトレーニングセットそれぞれのデータについて、
  確率を計算した後に、確率の大きい順にデータを並び替える

確率順に並べたデータ

No.	x	y	t	P
1	24.43	6.95	1	0.98
2	8.84	11.92	1	0.91
3	18.69	-1.17	1	0.86
4	17.37	-0.07	1	0.86
5	4.77	11.66	1	0.85
6	0.83	10.74	0	0.73
7	1.57	8.51	1	0.69
8	10.07	-0.53	1	0.66
9	0.99	6.04	1	0.58
10	10.73	-4.88	0	0.53
11	11.16	-6.77	0	0.47
12	-11.21	14.64	0	0.46
13	-5.67	5.05	1	0.31
14	-0.06	-1.47	0	0.28
15	-9.25	6.74	0	0.26
16	1.05	-4.86	0	0.21
17	-12.35	5.61	0	0.16
18	-6.12	-2.41	1	0.12
19	-2.17	-14.40	0	0.04
20	-4.06	-15.70	0	0.02

▲表5.1 トレーニングセットを確率順にならべたデータ

真陽性率・偽陽性率

- P>1 → すべてのデータが陰性

  確率Pが１を超えるようなデータは存在しない

  → 真陽性率 0　かつ　偽陽性率 0

- P>0.95 → No,1のデータが陽性と判定

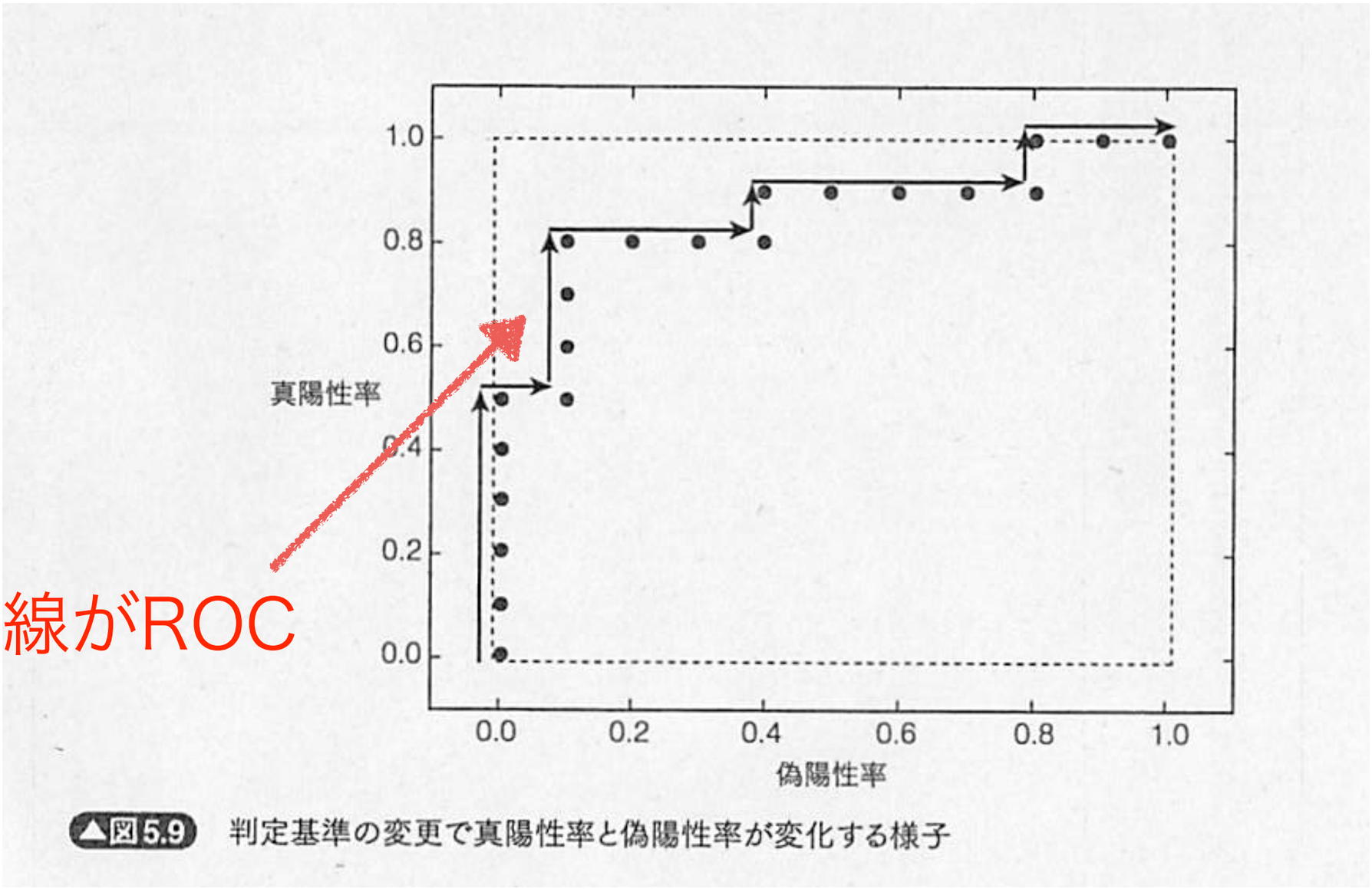
  →真陽性率 1/10　かつ偽陽性率 0

- P>0.90 →No,1とNo,2のデータが陽性と判定

  →真陽性率 2/10　かつ偽陽性率 0

判定基準の場所を一段ずつ下げながら、それぞれの率を計算

真陽性率・偽陽性率

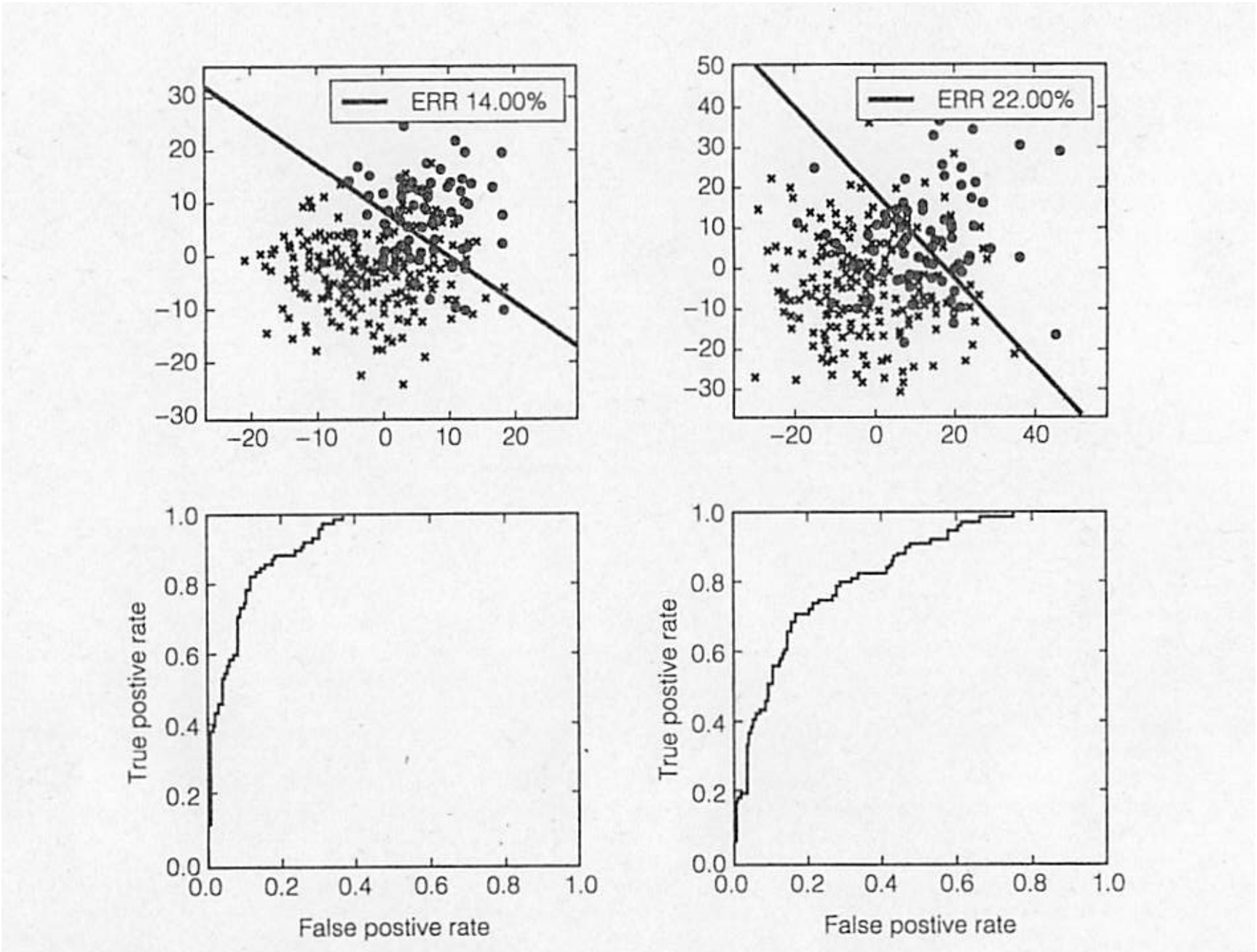


▲図5.9 判定基準の変更で真陽性率と偽陽性率が変化する様子

この線がROC

判定基準の選択に有効

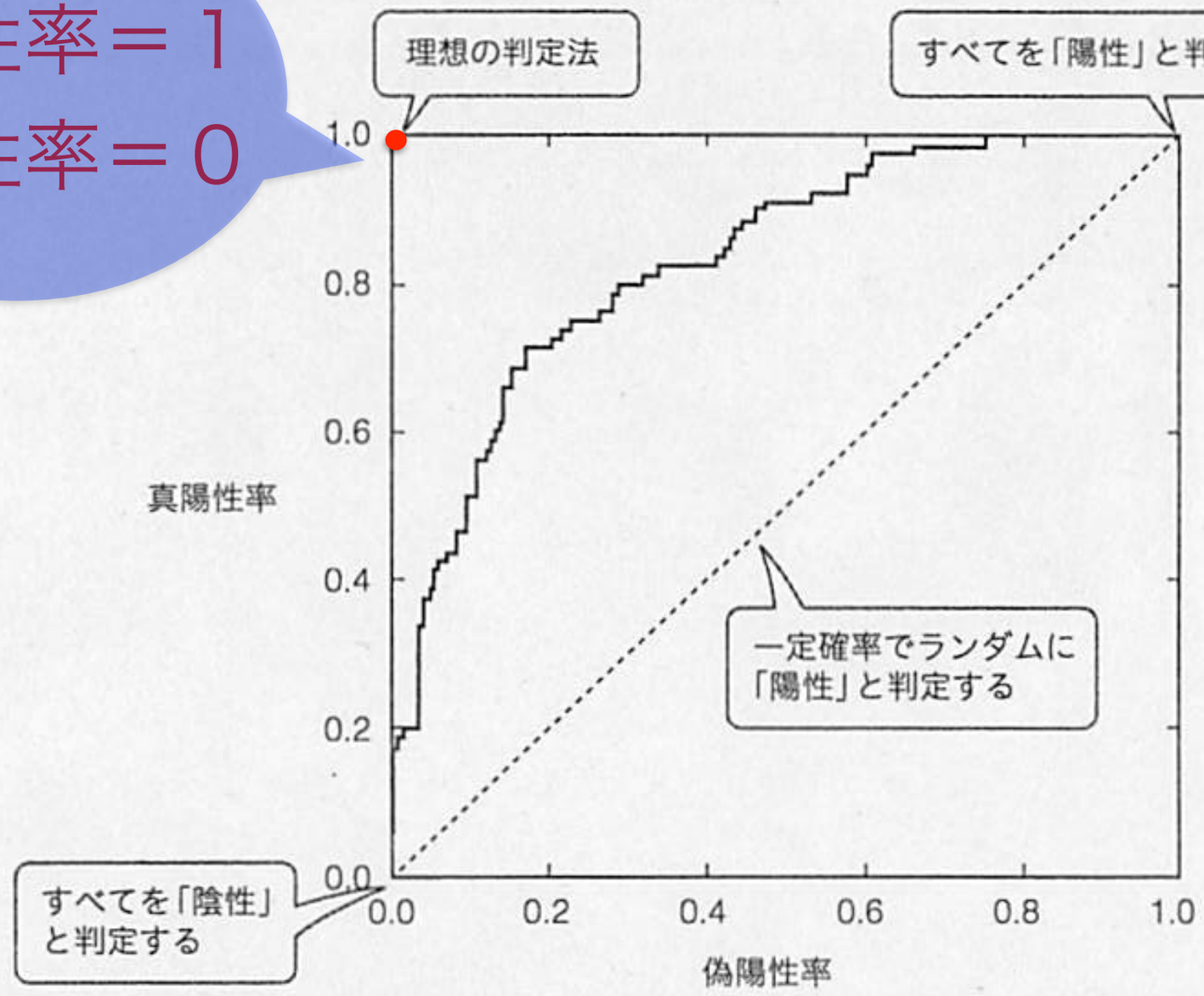
ROC曲線



▲図5.10　2種類のROC曲線の比較

理想の判定法

真陽性率＝１
偽陽性率＝０



▲図 5.11　特別な判定法に対応する部分

- ROC曲線
  https://oku.edu.mie-u.ac.jp/~okumura/stat/ROC.html

- 教科書