統計的機械学習論
5. ロジスティック回帰とROC曲線
〜学習モデルの評価方法〜

夏季集中講義(2019/09/02-13)

@九州工業大学

(飯塚キャンパス 大学院セミナー室 7 F)

このクラスのねらいと達成目標

ねらい

ロジスティック回帰による分類の学習を通して確率を考慮に入れた分類を理解する。

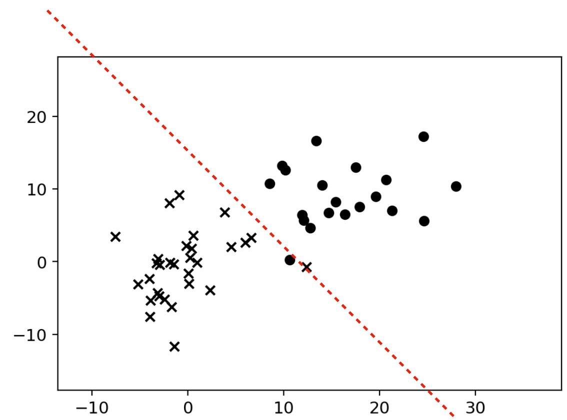
達成目標

- **ロジスティック回帰の考えかたを理解する**
- **IRLS（反復再重み付け最小二乗法）を理解する**
- **ロジスティック回帰が最尤推定法の考え方を踏襲した最適解を求める問題であることを理解する**

クラスの進行

5 ロジスティック回帰とROC曲線

「最尤推定法でパラメータを決定」
確率的な推定が可能になる



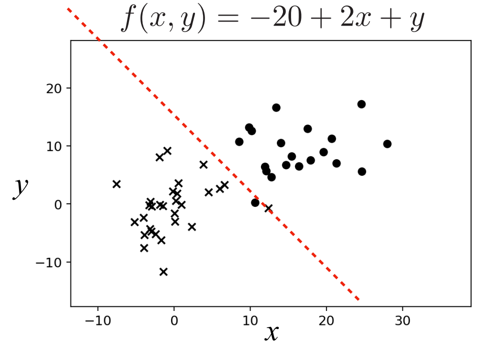
5.1 分類問題への最尤推定法の適用

最尤推定法のふりかえり

トレーニングセットのデータが得られる確率
（尤度関数）がもっとも高くなるように係数を決定した

5.1.1 データ発生率の推定

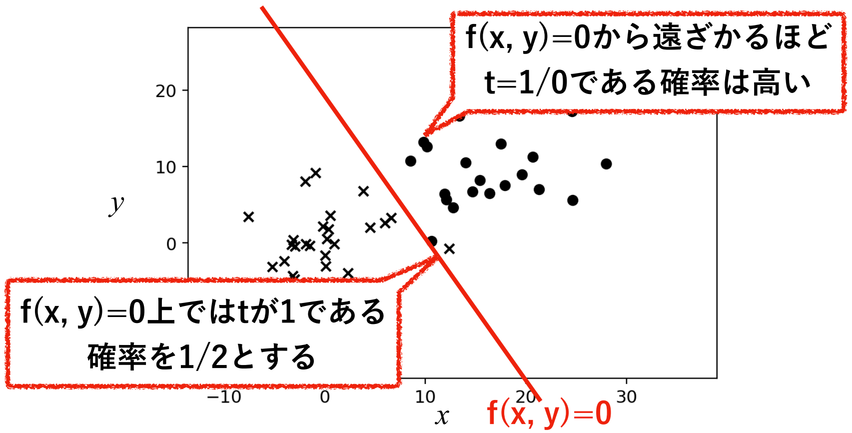
2種類のデータを分類する直線を表す一次関数を定義



$$\boxed{f(x,y) = w_0 + w_1 x + w_2 y}$$

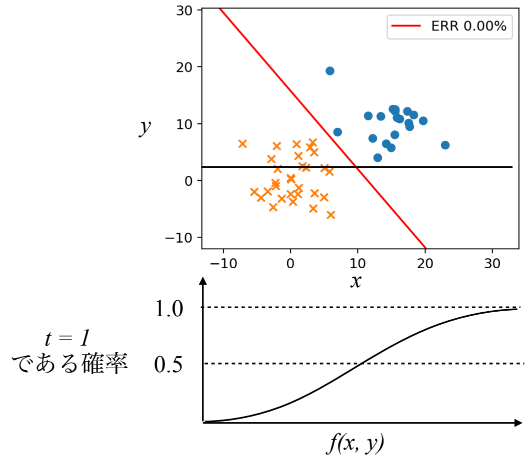
5.1.1 データ発生率の推定

得られるデータの属性が1である確率：P
得られるデータの属性が1である確率：1-P



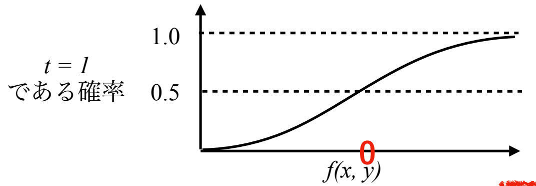
5.1.1 データ発生率の推定

分割線からどれくらい離れているかはf(x, y)で判断できる



5.1.1 データ発生率の推定

t=1である確率とf(x, y)との関係は、ロジスティック関数で表現することができる



一般的なロジスティック関数

ここでのaがf(x, y)

$$\sigma(a) = \frac{1}{1+e^{-a}}$$

5.1.1 データ発生率の推定

t=1のデータが発生する確率： $P(x, y) = \sigma(w_0 + w_1 x + w_2 y)$

t=0のデータが発生する確率： $1 - P$

個々のデータが発生する確率：

$$\begin{cases} t_n = 1 & \to P(x_n, y_n) \\ t_n = 0 & \to 1 - P(x_n, y_n) \end{cases}$$

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

5.1.1 データ発生率の推定

確率をベクトルで表現することができる
（Z_nはn番目がt=1である確率）

$$Z_n = \sigma(\mathbf{w}^{\mathrm{T}}\phi_n)$$

$$\mathbf{w} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix}, \quad \phi_n = \begin{pmatrix} 1 \\ x_n \\ y_n \end{pmatrix}$$

トレーニングセット全体が得られる確率について考えると

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

5.1.2 最尤推定法によるパラメータの決定

式が複雑で式変形だけで係数を求めることはできない

↓

確率が大きくなる方向に係数を修正するアルゴリズムが必要

↓

ニュートン・ラフソン法
（ニュートン法を多次元に拡張したもの）

ニュートン・ラフソン法における係数の更新の仕方

$$\mathbf{w}_{new} = \mathbf{w}_{old} - \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi}\right)\mathbf{\Phi}^{\mathrm{T}}(\mathbf{z} - \mathbf{t})$$

5.1.2 最尤推定法によるパラメータの決定

ニュートン・ラフソン法における係数の更新の仕方

依存　依存

$$\mathbf{w}_{new} = \mathbf{w}_{old} - \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\right)\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{z}-\mathbf{t})$$

w_{old}が与えられたときにzとRを計算してwの式に代入

↓

w_{new}が決まる

↓

w_{new}をw_{old}として次のw_{new}を計算する

↓

確率の値が大きくなり最大値に達する

5.1.2 最尤推定法によるパラメータの決定

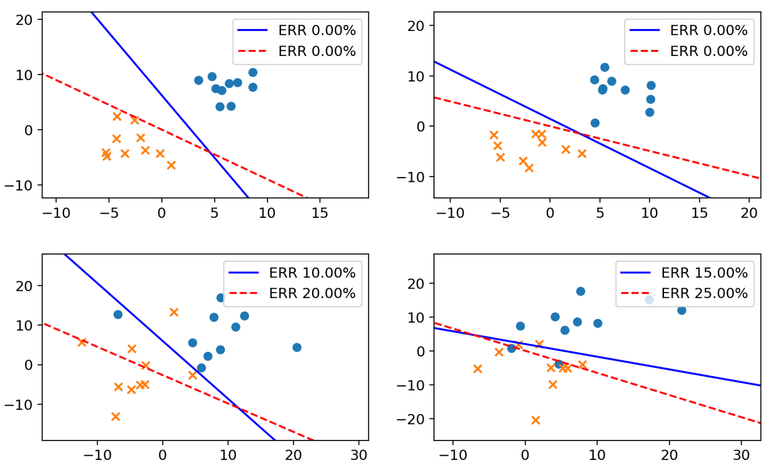
計算終了の条件：

$$\frac{\|\mathbf{w}_{new} - \mathbf{w}_{old}\|^2}{\|\mathbf{w}_{old}\|^2} < 0.001$$

「反復再重み付け最小二乗法」

5.1.3 サンプルコードによる確認

Blue: Logistic Regression, Red: Perceptron



5.2 ROC曲線による学習モデルの評価

ロジスティック回帰ではxy平面上で得られるデータがt=1である確率を考えることで分割線を決定した。

分割線はどのような性質を持っていたか、覚えていますか？

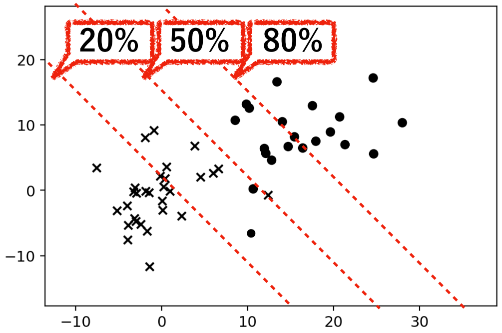
- **モデルが多項式f(x, y)で表現できるとき、f(x, y)=0**
- **分割線上ではt=1もしくは0である確率が0.5**

現実の問題に適用するとき、判定が確率0.5である場所を境界にするのは適切とはいえない場合もある。

ROC(Receiver Operating Characteristics)曲線を使って適切な分割線の設定とアルゴリズムの性能を評価する

5.2.1 ロジスティック回帰の現実問題への適用

例えば重篤な病気の感染／非感染を判定したデータだとしたら？



ロジスティック回帰をそのまま適用するなら50%の確率で感染している人に精密検査を勧告するが、50%未満の人にも勧告するほうがいい

5.2.1 ロジスティック回帰の現実問題への適用

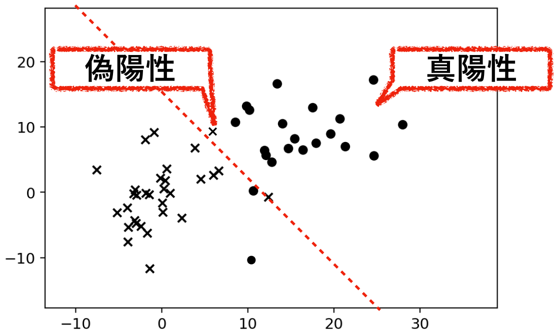
- **発見したい属性を持つデータ→「陽性」**
- **発見したい属性を持たないデータ→「陰性」**

陽性と判断したデータのうち、

- **本当に陽性であるデータ→「真陽性」**
  (True Positive; TP)
- **本当は陰性であるデータ→「偽陽性」**
  (False Positive; FP)

5.2.1 ロジスティック回帰の現実問題への適用

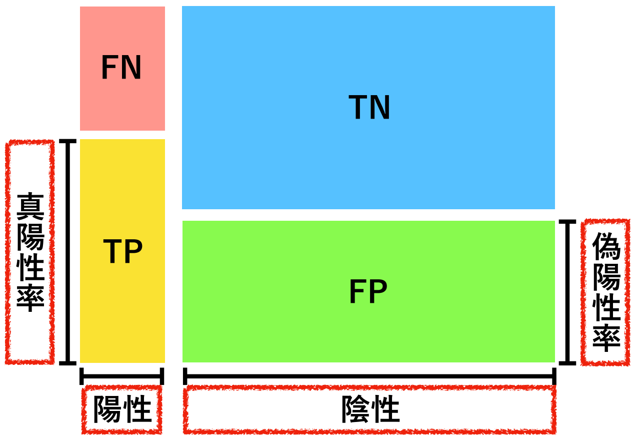
真陽性・偽陽性のデータの位置と分割線との関係



- **真陽性になる確率→「TP率」もしくは「真陽性率」**
- **偽陽性になる確率→「FP率」もしくは「偽陽性率」**

5.2.1 ロジスティック回帰の現実問題への適用

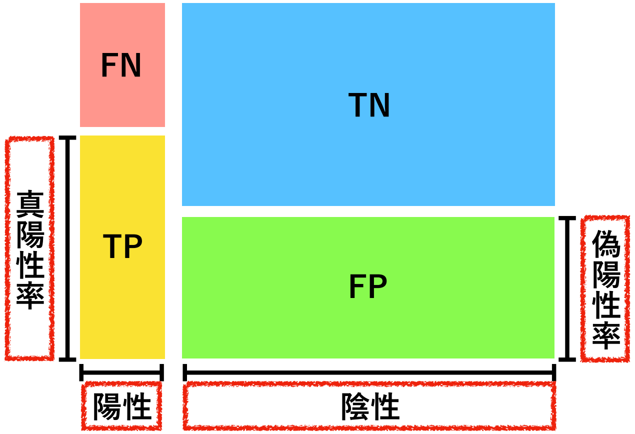
真陽性・偽陽性の内訳を図形で解説



5.2.1 ロジスティック回帰の現実問題への適用

真陽性率はできる高くしたい⇄偽陽性率は低くしたい

（両立しないトレードオフの可能性）



5.2.2 ROC曲線による性能の評価

$f(x,y)$のパラメータは(w_0, w_1, w_2)

$$P(x,y) = \sigma(w_0 + w_1 x + w_2 y)$$

$$\sigma(a) = \frac{1}{1+e^{-a}}$$

	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
…	…	…	…	…
10	10.73	-4.88	0	0.53

5.2.2 ROC曲線による性能の評価

「陽性」と判断する基準をP > 1とするなら、

データはすべて「陰性」と判定する

- **TP率 = 0 → 正しく陽性と判断できたデータは存在しない**
- **FP率 = 0 → 陰性のデータを誤って陽性とすることもない**

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
…	…	…	…	…
10	10.73	-4.88	0	0.53

判定規準

5.2.2 ROC曲線による性能の評価

次に、No.1のデータのみ「陽性」と判断する基準Pならば、

データは１つだけ「陽性」と判定される

- **TP率 = 1/10 →陽性10個のうち陽性と判定したのは1個**
- **FP率 = 0 → 陰性のデータは全て陰性と判定した**

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
…	…	…	…	…
20	-4.06	-15.70	0	0.02

判定規準

5.2.2 ROC曲線による性能の評価

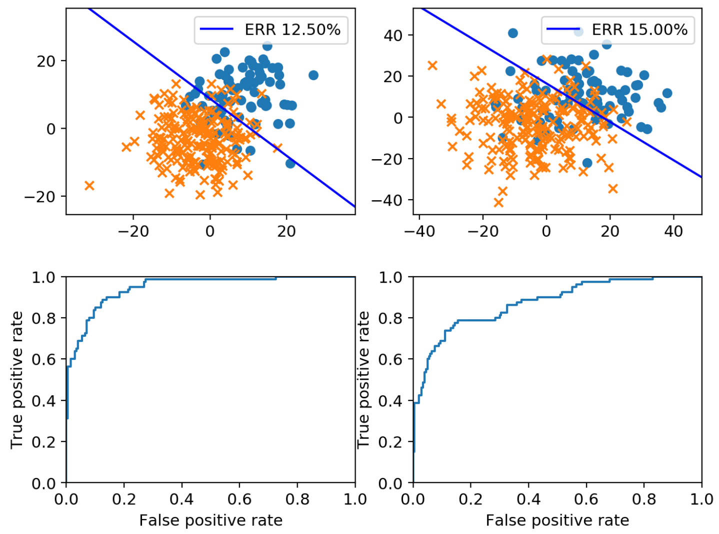
次に、No.1, 2のデータを「陽性」と判断する基準Pならば、
データは2つだけ「陽性」と判定される

- **TP率 = 2/10 →陽性10個のうち陽性と判定したのは2個**
- **FP率 = 0 → 陰性のデータは全て陰性と判定した**

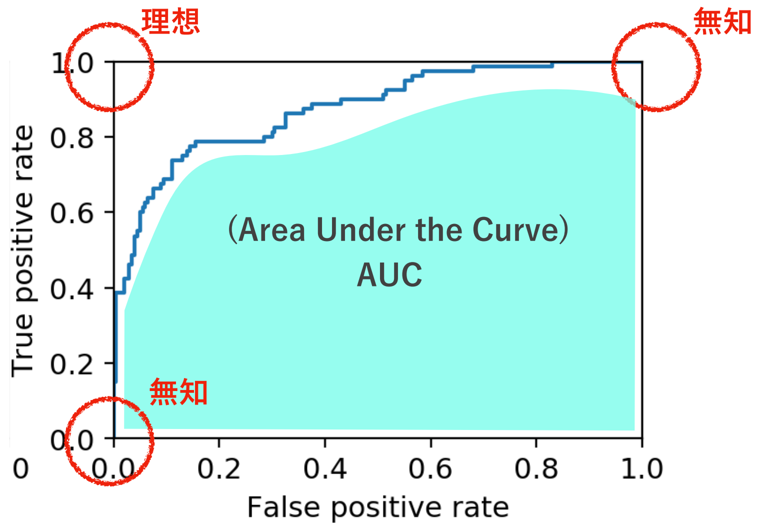
No	x	y	t	p
1	24.43	6.95	1	0.98
2	8.84	11.92	1	0.91
3	18.69	-1.17	1	0.86
4	17.37	-0.07	1	0.86
…	…	…	…	…
20	-4.06	-15.70	0	0.02

判定規準

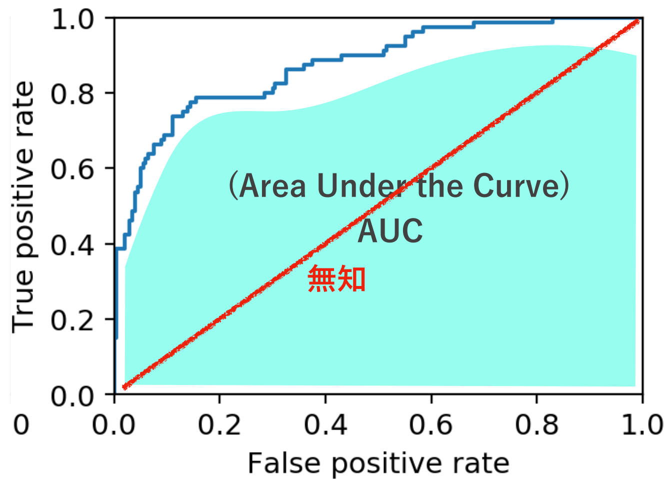
5.2.3 サンプルコードによる確認



5.2.4 「無知」な判定法



5.2.5 他の「無知」な判定法



まとめ

- ロジスティック回帰について理解できた
- ROC曲線を理解できた
- ROC曲線を描くことでモデルを評価できることを学んだ
- 「無知」な判定法とは何かを理解した

ミニッツペーパー

- n番目のデータがt=1(あるいはt=0)である確率を求める式を記述しなさい
  $P_n = P(x_n, y_n)^{t_n}\{1 - P(x_n, y_n)\}^{1-t_n}$
- ロジスティック回帰による分類においても、データが得られる確率の最大値を求めていたはずだったが、いつの間にか誤差を最小にするような係数を求めることに話が落ち着く。それはなぜ何故か説明してください。
  EとPの関係式で説明できる
- ヘッセ行列とはどのような行列か、定義とどんなことでも良いので性質を記述しなさい
  ヘッセ行列とは行列の成分について2階偏微分を行ったもの。
  ヘッセ行列が正定値ならその停留点は最小値である。