L05 ロジスティック回帰

樋口さぶろお

龍谷大学 先端理工学部 数理・情報科学課程

理論物理学特論 L05(2021-10-26 Tue)

最終更新: Time-stamp: "2021-10-26 Tue 10:58 JST hig"

今日の目標

- 機械学習でいう回帰と分類の違いを説明できる
- ロジスティック回帰と, 線形回帰, 母比率の推定の関係を説明できる
- statsmodels でロジスティック回帰できる

ここまで来たよ

4 重回帰

4 ロジスティック回帰
- 分類問題
- ロジスティック回帰の統計モデル
- ロジスティック回帰の実行
- GLM 一般化線形モデル

回帰と分類 (機械学習で言う)

線形回帰の学習データ

	説明変数 x	目的変数 Y
1	10.3	21.1
2	12.1	22.9
⋮		
n	19.8	41.2

機械学習の言葉では
教師あり, 回帰=目的変数が連続値

Y には小さい (?) 誤差を含む.

これから扱う学習データ

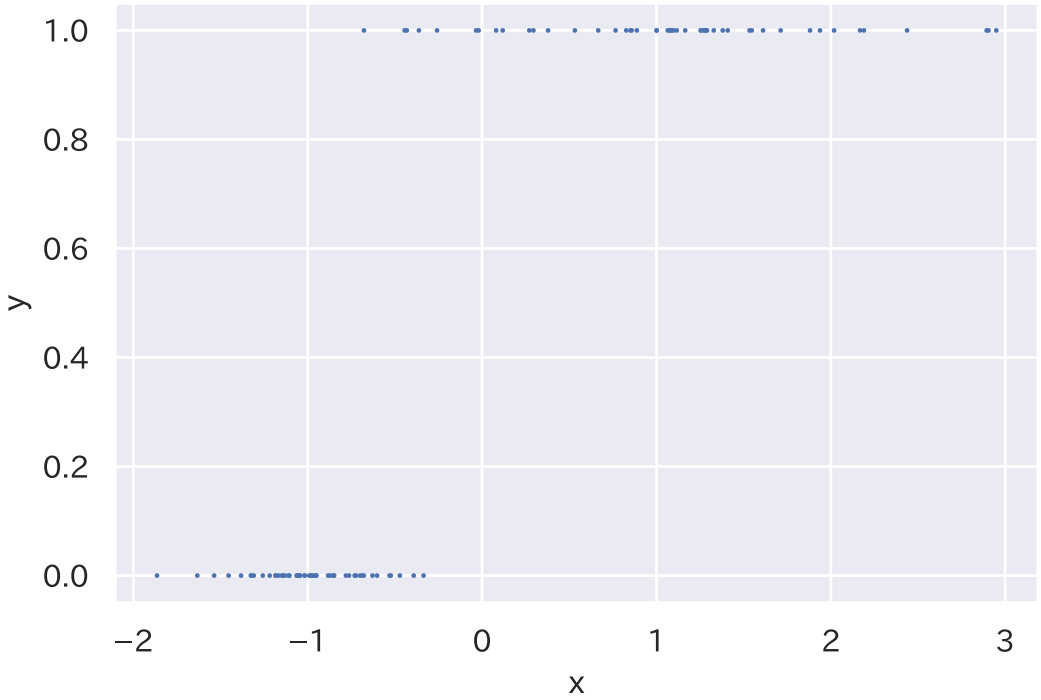
	説明変数 x	目的変数 Y
1	10.3	A
2	12.1	B
⋮		
n	19.8	A

機械学習の言葉では
教師あり, 分類=目的変数がカテゴリ変数

これから扱う学習データ

X 連続値

Y 間隔に意味のない有限個 $\sim$ 離散型, 今日は A=0, B=1 に限定)



ここまで来たよ

ロジスティック回帰

ロジスティック回帰

データ (x_i, y_i) $(i = 1, \ldots, n)$ x は連続値, $y = \mathrm{A}$ xor B から

$$P(Y = \mathrm{B}) = \frac{\mathrm{e}^{\,{}^{\mathrm{t}}\boldsymbol{\beta}\cdot\boldsymbol{x}}}{1 + \mathrm{e}^{\,{}^{\mathrm{t}}\boldsymbol{\beta}\cdot\boldsymbol{x}}}, \quad P(Y = \mathrm{A}) = 1 - P(Y = \mathrm{B}),$$

$$\boldsymbol{\beta} = \begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix}, \boldsymbol{x} = \begin{bmatrix}1 \\ x_1\end{bmatrix}, \ {}^{\mathrm{t}}\boldsymbol{\beta}\cdot\boldsymbol{x} = \beta_0 + \beta_1 \cdot x_1$$

という説明がもっともになるように, β_0, β_1 を推定すること. それを使って, $\boldsymbol{x}$ が与えられたときの $P(Y = \mathrm{B})$ や Y を推定すること.

回帰っていうけど分類じゃん? by 機械学習の人

$\beta_0 + \beta_1 \cdot x_1$ ってあたりが回帰? 0 と 1 の間の連続値 (確率) だから回帰?

by 統計学の人

$\beta_0 + \beta_1 \cdot x_1 + \epsilon$, $\epsilon \sim \mathrm{N}(0, \sigma^2)$ ではない. 推定するもの (左辺) がすでに確率

ロジスティック回帰はもっとも 1:母平均値 対 母比率

	説明変数なし	説明変数あり
母平均値	$Y \sim \mathrm{N}(\mu, \sigma^2)$ の母平均値 μ の推定 $\overline{Y} = \frac{1}{n}(Y_1 + \cdots)$	線形回帰 $Y = \beta_0 + \beta_1 x + \epsilon$ β_j:連立 1 次方程式を解いて
カテゴリ変数の確率	$Y \sim \mathrm{B}(1, p)$ の母比率 p の推定 $\hat{p} = \frac{k}{n}$	ロジスティック回帰 $P(Y=1) = \frac{\mathrm{e}^{{}^{\mathrm{t}}\boldsymbol{\beta}\cdot\boldsymbol{x}}}{1+\mathrm{e}^{{}^{\mathrm{t}}\boldsymbol{\beta}\cdot\boldsymbol{x}}}$ β_j:ニュートン法で

ロジスティック回帰はもっとも 2:1 次関数 対 ロジスティック関数

$$p = P(Y = 1) = \beta_0 + \beta_1 \cdot x_1$$

じゃだめなの?

$\rightsquigarrow$ $p = \beta_0 + \beta_1 x$ は $-\infty < y < +\infty$. 確率のとる範囲 $0 \leq p \leq 1$ に収まらない…

区間 $(-\infty, \infty)$ を $(0, 1)$ に写す写像ある?

一例 (標準) シグモイド関数. sigmoid function $y = \frac{\mathrm{e}^x}{1+\mathrm{e}^x} = \frac{1}{1+\mathrm{e}^{-x}}$

x	$-\infty$		0		$+\infty$
y'	$+$	$+$	$\frac{1}{4}$	$+$	$+$
y	0	↗	$\frac{1}{2}$	↘	$+1$

標準シグモイド関数を移動拡大縮小したもの ロジスティック関数 (logistic function)

x_1	$-\infty$		$-\beta_0/\beta_1$	0		$+\infty$
p'	$+$		$\frac{\beta_1}{4}$	$\frac{\mathrm{e}^{\beta_0}}{(1+\mathrm{e}^{\beta_0})^2}$		$+$
$p=\frac{\mathrm{e}^{{}^{\mathrm{t}}\boldsymbol{\beta}\cdot\boldsymbol{x}}}{1+\mathrm{e}^{{}^{\mathrm{t}}\boldsymbol{\beta}\cdot\boldsymbol{x}}}$	0		$\frac{1}{2}$	$\frac{\mathrm{e}^{\beta_0}}{1+\mathrm{e}^{\beta_0}}$		$+1$

$\beta_1(x-(-\frac{\beta_0}{\beta_1}))$

β_1 横拡大縮小

β_0 横移動

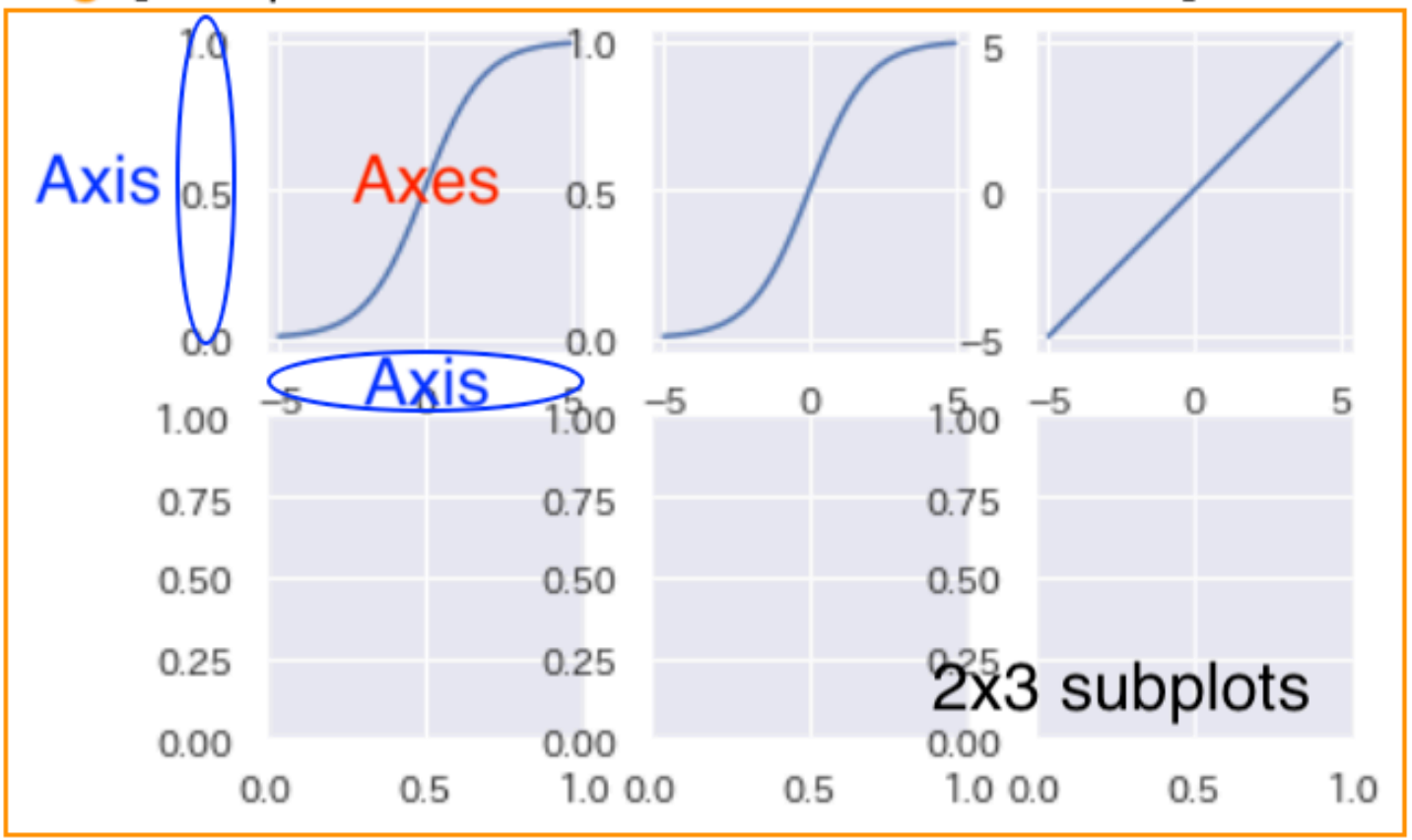
Matplotlib でのグラフ描画

```
[18]  1  import matplotlib.pyplot as plt #低レベルグラフ描画
      2  fig, ax=plt.subplots(2,3)
      3  ax[0][0].plot(x,y1)
      4  ax[0][1].plot(x,y1)
      5  ax[0][2].plot(x,x)
```

✓ 1 秒

Figure

[<matplotlib.lines.Line2D at 0x7fe5b2837e10>]



ここまで来たよ

4 重回帰

4 ロジスティック回帰
- 分類問題
- ロジスティック回帰の統計モデル
- ロジスティック回帰の実行
- GLM 一般化線形モデル

ロジスティック回帰の回帰係数の推定方法

尤度 (2 変数関数)

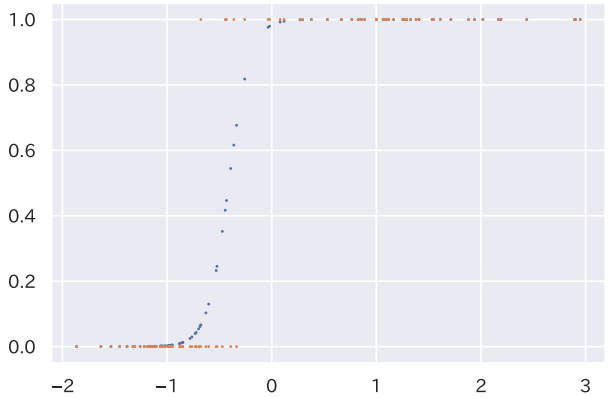
$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p(y_i|x_i)$$

を β_0, β_1 について最大化.

$\rightsquigarrow \frac{\partial L}{\partial \beta_0}(\beta_0, \beta_1) = \frac{\partial L}{\partial \beta_1}(\beta_0, \beta_1) = 0$

線形回帰と違って, 連立 1 次方程式ではすまない.

$\rightsquigarrow$ ニュートン法で数値的に解く　　数値計算法



ニュートン法は反復法 (iteration) だった

import statsmodels.api.formula as smf

- glm in statsmodels

```
[ ]  1 import statsmodels.api as sm
     2 import statsmodels.formula.api as smf
```

```
[121]  1 result_g=smf.glm(formula='y~x',data=dfb, family=sm.families.Binomial(sm.families.links.logit())).fit()
```

```
[122]  1 result_g.summary()
```

Generalized Linear Model Regression Results

モデル

Dep. Variable:	y	No. Observations:	100
Model:	GLM	Df Residuals:	98
Model Family:	Binomial	Df Model:	1
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-8.7065
Date:	Fri, 22 Oct 2021	Deviance:	17.413
Time:	08:45:09	Pearson chi2:	22.8
No. Iterations:	10		
Covariance Type:	nonrobust		

あてはまり

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	4.0625	1.843	2.204	0.028	0.450	7.675
x	9.9026	3.610	2.743	0.006	2.828	16.977

回帰係数

scikit-learn.LogisticRegression では, 変換せずに直接指定できる

ここまで来たよ

4 重回帰

4 ロジスティック回帰
- 分類問題
- ロジスティック回帰の統計モデル
- ロジスティック回帰の実行
- GLM 一般化線形モデル

一般化線形モデル

誤差構造	リンク関数	説明変数の個数は p
$Y \sim \mathrm{N}(\mu, \sigma^2)$	$\mathrm{id}(\mu) =$	${}^{\mathrm{t}}\boldsymbol{\beta} \cdot \boldsymbol{x}$
$Y \sim \mathrm{B}(1, p)$	$\mathrm{logit}(p) =$	${}^{\mathrm{t}}\boldsymbol{\beta} \cdot \boldsymbol{x}$

恒等関数 identity function $\mathrm{id}(\mu) = \mu$

ロジット関数 logit function $\mathrm{logit}(p) = \log \frac{p}{1-p}$.

$\log \frac{p}{1-p} = {}^{\mathrm{t}}\boldsymbol{\beta} \cdot \boldsymbol{x}$ を p について解くと, $p = \frac{\mathrm{e}^{\,{}^{\mathrm{t}}\boldsymbol{\beta}\cdot\boldsymbol{x}}}{1+\mathrm{e}^{\,{}^{\mathrm{t}}\boldsymbol{\beta}\cdot\boldsymbol{x}}}$.

$\frac{p}{1-p}$ オッズ比.