

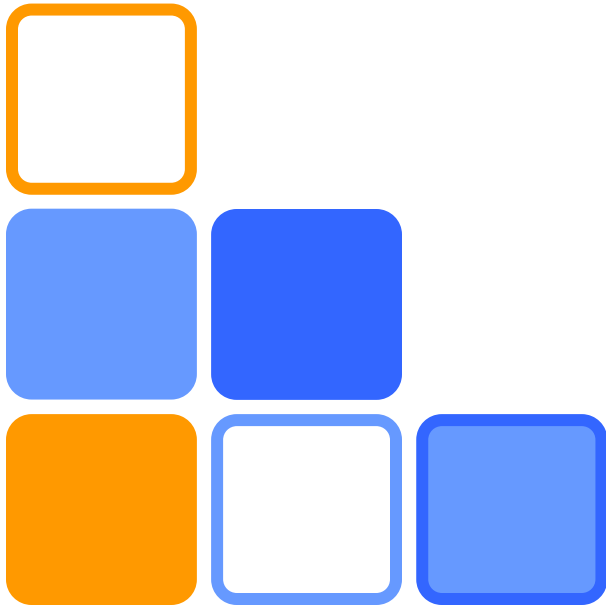
薬学情報処理演習 第6回

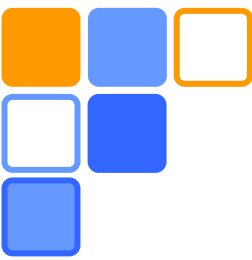
機械学習(1)

多項式を用いた曲線による
フィッティング

奥菌 透

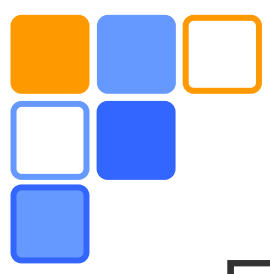
コロイド・高分子物性学





機械学習(machine learning)...?

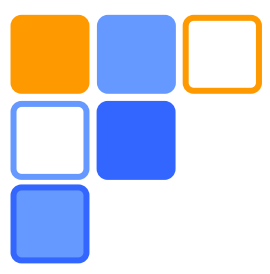
- 機械(パソコン)を使って学習する...NO!
- 機械が学習する...どうやって?
- 訓練が必要(膨大な量のデータが必要)
- 学習の検証も必要
- どんな学習...? (教師あり/なし学習)
- 今の時代、ネットでなんでも手に入るけど?
- 多くが Black Box (闇の中)「気持ち悪い」
- 先ずは簡単な回帰分析から。



確率論の基本

- 同時確率 $p(X, Y)$: X かつ Y の確率
- 条件つき確率 $p(Y|X)$: X が与えられた下での Y の確率
- 加法定理: $p(X) = \sum_Y p(X, Y)$ (周辺確率)
- 乗法定理: $p(X, Y) = p(Y|X)p(X)$
- また、 $p(X, Y) = p(Y, X)$ であるから
$$p(Y, X) = p(X|Y)p(Y)$$
- したがって以下のベイズの定理が成り立つ

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$



事前確率、事後確率、尤度関数

- D : 観測データ (x_i, y_i) , w : パラメータ
- ベイズの定理:
$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$
- 事後確率: $p(w|D)$ D を観測した後に与えられる (eg. モンティ・ホール問題)
- 事前確率: $p(w)$ (先験確率)
- ベイズの定理の右辺 $p(D|w)$ は尤度 (ゆうど)
- 尤度最大化と誤差最小化は等価

2次元データの分布・平均値

□ 2つの量 (x_i, y_j) の度数分布 f_{ij}

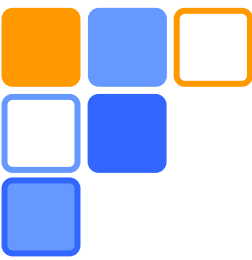
平均値: $\bar{x} = \sum_i \sum_j x_i f_{ij} / N$, $\bar{y} = \sum_i \sum_j y_j f_{ij} / N$
 $(N = \sum_i \sum_j f_{ij})$

	140	150	160	170	180	190	$\sum_i f_{ij}$
40	1						1
50		1	3	1			5
60		1	1	2			4
70			1	2	3	1	7
80					2		2
90						1	1
$\sum_j f_{ij}$	1	2	5	5	5	2	

$$\bar{x} = \frac{140 \times 1 + 150 \times 2 + \dots + 190 \times 2}{20} = 168.5 \approx 170$$

$$\bar{y} = \frac{40 \times 1 + 50 \times 5 + \dots + 90 \times 1}{20} = 63.5 \approx 60$$

身長 (cm)	体重 (kg)
156	45
143	42
175	71
181	75
179	65
168	63
158	48
176	73
173	69
155	49
163	56
169	52
171	71
188	73
191	88
177	82
165	63
154	48
149	56
164	66
167.75	62.75



2次元データの相関

□ 分散 (variance)

$$\sigma_x^2 = \frac{1}{N} \sum_i \sum_j (x_i - \bar{x})^2 f_{ij} = \frac{1}{N} \sum_i \sum_j x_i^2 f_{ij} - \bar{x}^2$$

$$\sigma_y^2 = \frac{1}{N} \sum_i \sum_j (y_j - \bar{y})^2 f_{ij} = \frac{1}{N} \sum_i \sum_j y_j^2 f_{ij} - \bar{y}^2$$

□ 共分散 (covariance)

$$C_{xy} = \frac{1}{N} \sum_i \sum_j (x_i - \bar{x})(y_j - \bar{y}) f_{ij} = \frac{1}{N} \sum_i \sum_j x_i y_j f_{ij} - \bar{x} \bar{y}$$

□ 相関係数 (correlation coefficient)

$$r = \frac{C_{xy}}{\sigma_x \sigma_y} \quad (|r| \leq 1)$$

r は x と y がどれくらい関連しているかを表す:

$r = 0$: 無相関

$r = 1$: 正の完全相関

$r = -1$: 負の完全相関

回帰直線

□ 最小2乗法による直線近似

2つの量 (x, y) の間に直線的な関係

$$y = ax + b \quad (a, b \text{ は定数})$$

があると仮定し、**直線からのずれ**の2乗和

$$Q = \sum_i \sum_j [y_j - (ax_i + b)]^2 f_{ij}$$

を**最小**にする a, b を求める:

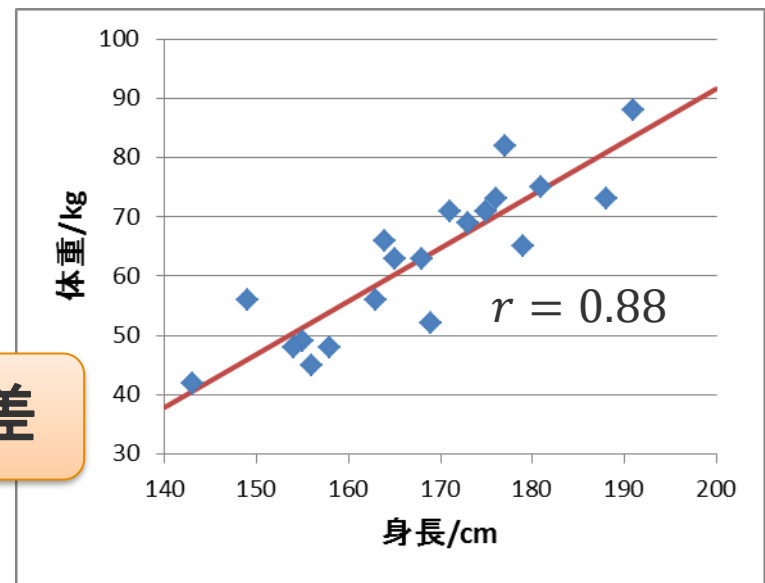
$$\frac{\partial Q}{\partial a} = 0, \quad \frac{\partial Q}{\partial b} = 0$$

回帰直線(regression line)

$$\frac{y - \bar{y}}{\sigma_y} = r \frac{x - \bar{x}}{\sigma_x}$$

が得られる。

誤差



データ $\{x_i, y_j\}$

平均値 $\bar{x}, \bar{y}$ を計算

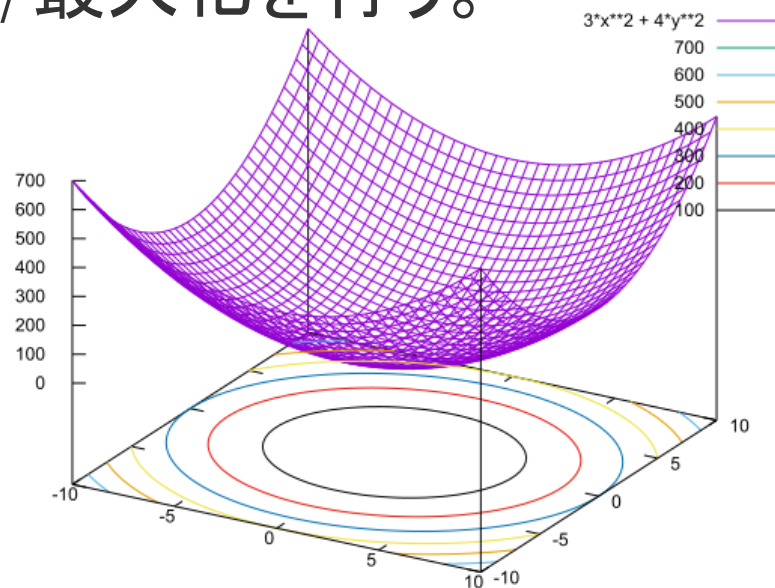
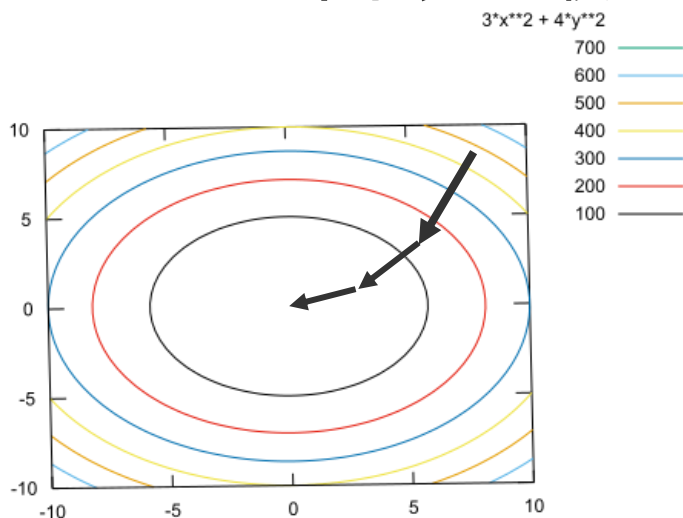
分散 σ_x^2, σ_y^2 , 共分散 C_{xy} を計算

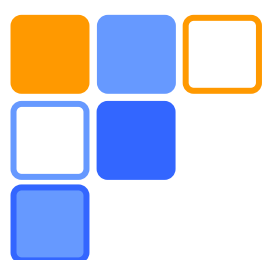
相関係数 r を計算

回帰直線 $y = \bar{y} + r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$ を得る

多項式による回帰分析

- 準備: エクセルの「ソルバー」を使用可能にする。
 - 「データ」タブをクリック、「分析」リボンに「ソルバー」が表示されていればOK
 - 表示されていない場合: 「ファイル」タブ→左下に「その他」→「オプション」→管理(A):Excelアドイン「設定」→☒ソルバーアドインにチェック
- ソルバー: 関数の最小/最大化を行う。





実際にエクセルで計算

□ データの作成

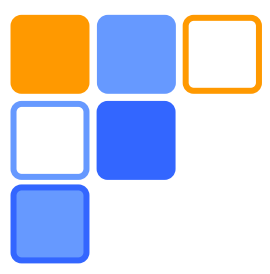
- 関数 $y = f(x)$, $x = x_1, x_2, \dots, x_N$ に対し
- $y_i = f(x_i) + r_i$, ($i = 1, 2, \dots, N$) を作成
- ただし、 r_i はランダムな数(平均0)

□ (x_i, y_i) から $y = f(x)$ を推定する

- $y_{\mathbf{w}}(x) = w_0 + w_1x + w_2x^2 + \dots + w_Mx^M$
- 誤差関数 $E(\mathbf{w})$ が最小となるように $\mathbf{w}$ を決める

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N [y_{\mathbf{w}}(x_i) - y_i]^2 \quad \mathbf{w} = (w_0, w_1, \dots, w_M)$$

- $f(x)$ は与えられた関数: $\sin(x)$, $\cos(x)$, $\tanh(x)$ など



計算例(1)

□ データ作成

y=tanh(a*x)			
dx =	1.00E-01		
a =	2		
x	y	r(0,0.3)	yr
-2	-0.99933	-0.10456	-1.10389
-1.9	-0.999	0.090118	-0.90888
-1.8	-0.99851	-0.47348	-1.47199
-1.7	-0.99777	-0.17749	-1.17527

平均値0, 標準偏差0.3の正規乱数
エクセルのデータ分析ツールで生成


$$y_i = f(x_i) + r_i$$

計算例(2)

- ソルバーを使って w を計算する。 $M = 3$



$$= \$I\$6 + \$I\$7 * A6 + \$I\$8 * A6^2 + \$I\$9 * A6^3$$

I	J	k
W0:w3	yw3	(yr-yw3)^2
0.017902	-0.96714	0.0187
1.182232	-1.05053	0.020065
-0.027	-1.11327	0.12868
-0.18593	-1.15646	0.000354
	-1.18123	0.063067

← 誤差関数

↑ ソルバーからの出力

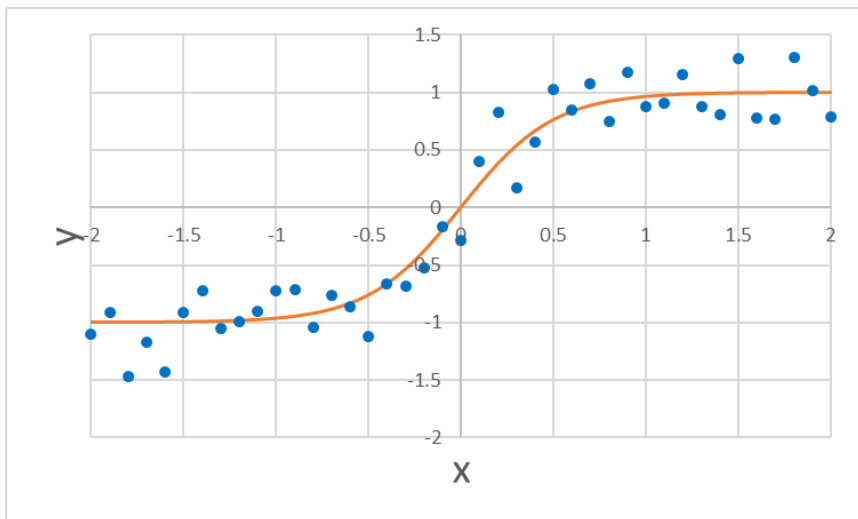
	0.891365	0.015759
	0.78691	8.78E-06
		2.817538

=SUM(K6:K46)

↑ ソルバーへの入力

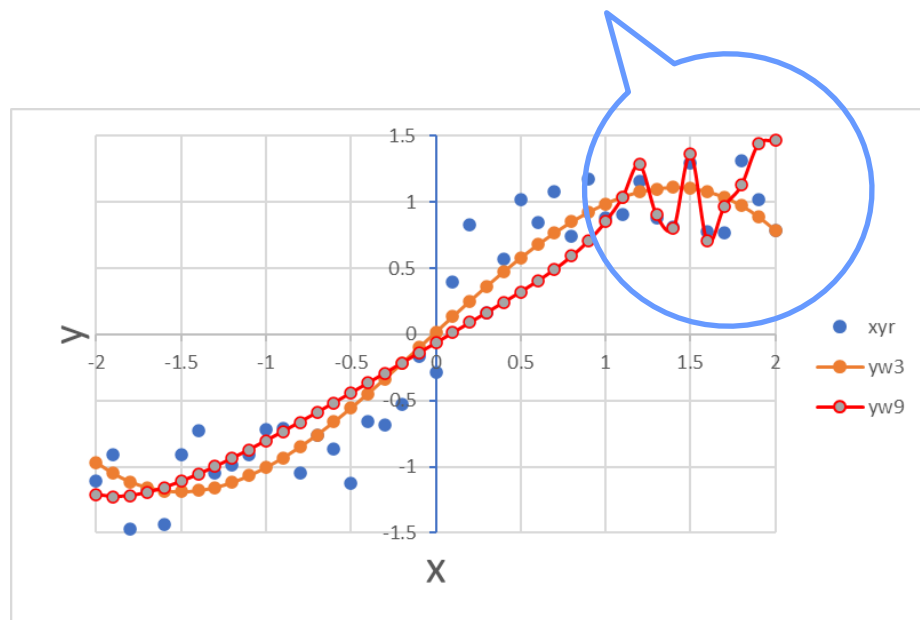
計算例(3)

結果のグラフ



実線: $y = \tanh(2x)$
ドット: $y = \tanh(2x_i) + r_i$

過学習
(over-fitting)



オレンジ: $M = 3$
赤: $M = 9$

計算例(4)

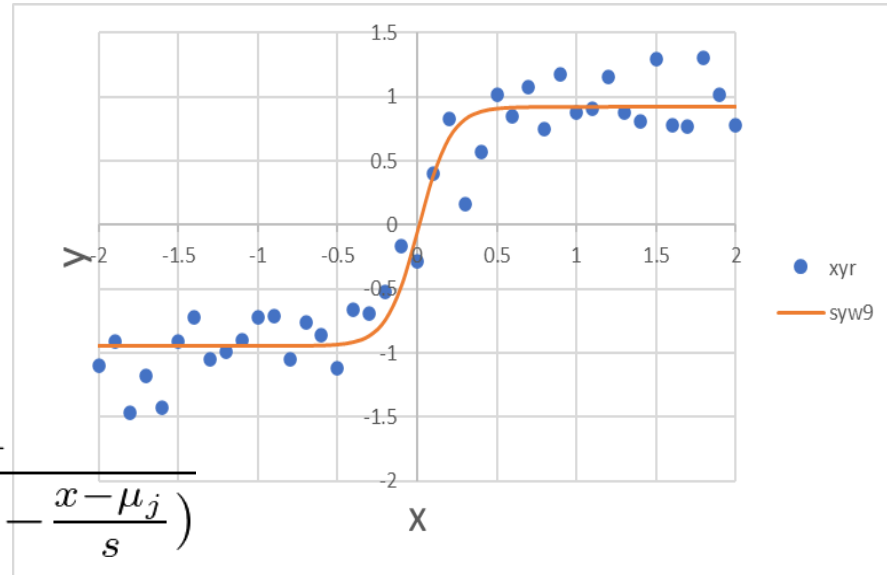
基底関数 $\phi_j(x)$ を使う。

$$\phi_j(x) = \sigma\left(\frac{x - \mu_j}{s}\right) = \frac{1}{1 + \exp\left(-\frac{x - \mu_j}{s}\right)}$$

$$y_w(x) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(x)$$

P Q R S

Note: $\tanh(x) = 2\sigma(2x) - 1$



	dm =	1.00E-02	
	s =	1.00E-01	
W0:w9	phi_j(x)	yw9	(yr-yw9)^2
-0.94891	1.87E-09	-9.49E-01	0.024018
0.207458	5.07E-09	-9.49E-01	0.001602
0.207458	1.38E-08	-9.49E-01	0.273609

=\$P\$6+\$P\$7*Q6+\$P\$8*Q6+\$P\$9*Q6+\$P\$10*Q6+\$P\$11*Q6+\$P\$12*Q6+\$P\$13*Q6+\$P\$14*Q6+\$P\$15*Q6

$\uparrow = 1 / (1 + \text{EXP}(-(A6 - \$R\$1) / \$R\$2))$



参考文献

- C.M.ビショップ 「パターン認識と機械学習 上
ベイズ理論による統計的予測」(丸善出版 2012)
- 涌井良幸、涌井貞美 「Excelでわかる機械学習
超入門—AIのモデルとアルゴリズムがわかる」(技術評論社
2019)
- 富谷昭夫 「これならわかる機械学習入門」(講
談社 2021)
- 江崎貴裕 「データ分析のための数理モデル入
門 本質をとらえた分析のために」(ソシム株式会社
2020)