

# 統計的モデリング基礎③

## ～重回帰・最尤推定～

鹿島久嗣  
(情報学科 計算機科学コース)

# 重回帰

# 重回帰：

## 複数の独立変数を用いて予測

- (単) 回帰では、ひとつの独立変数から予測を行う

$$g(x) = \beta x + \alpha$$

–例：年齢から年収を予測する

$$(\text{年収}) = \beta \times (\text{年齢}) + \alpha$$

- 重回帰では複数の ( $m$ 個の) 独立変数を用いる

$$g(x) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \alpha$$

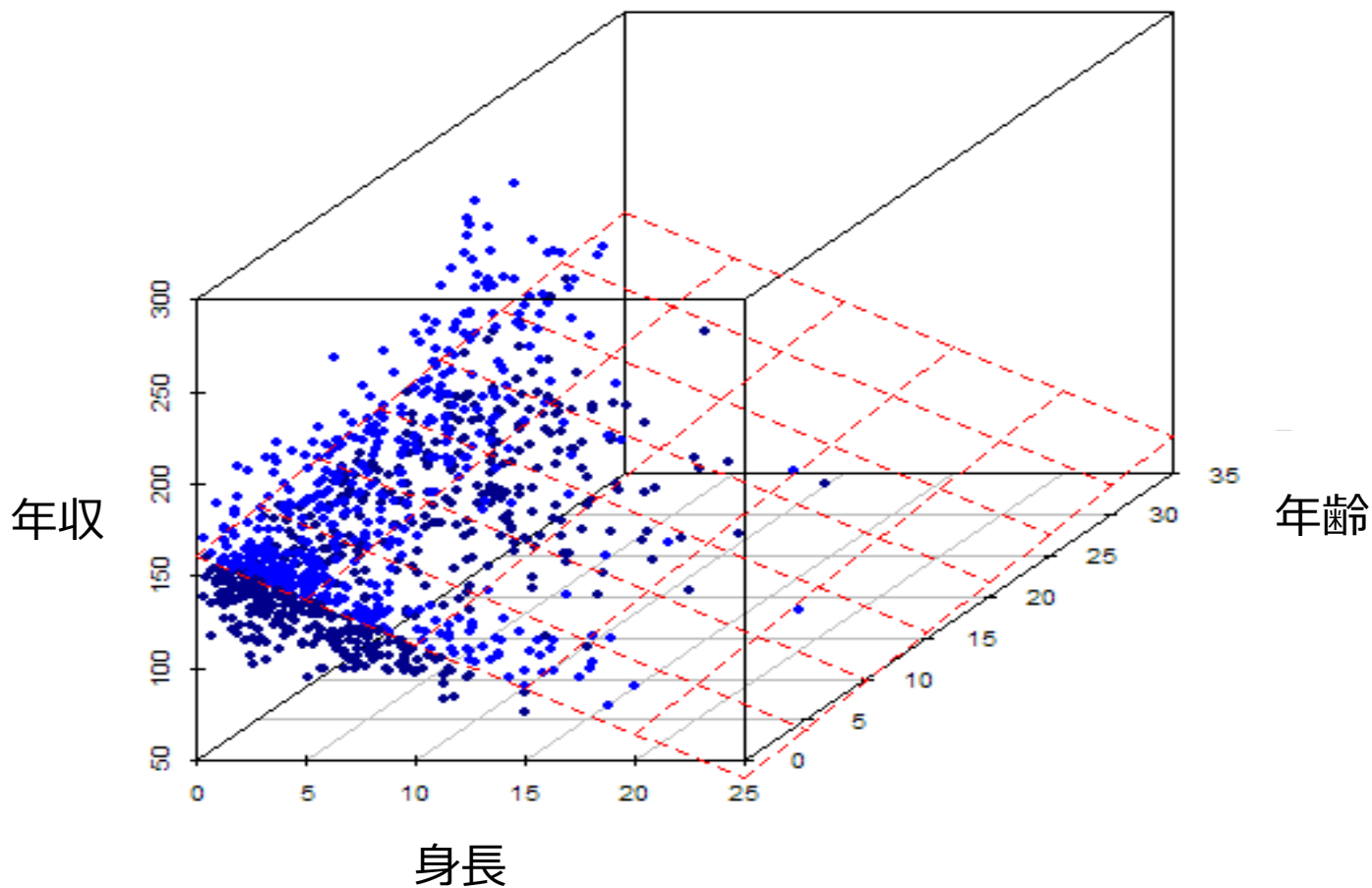
–例：年齢と身長から年収を予測する

$$(\text{年収}) = \beta_{(\text{年齢})} \times (\text{年齢}) + \beta_{(\text{身長})} \times (\text{身長}) + \alpha$$

# 重回帰のイメージ：

(超) 平面でデータに当てはめる

- 単回帰では直線で近似、重回帰では (超) 平面で近似



# 重回帰モデルの推定問題：

## 最小二乗法によってパラメータを推定する

- 単回帰と同じく、モデルの予測と実際のデータとの食い違いを二乗誤差で測る

$$\begin{aligned} \ell(\alpha, \{\beta_i\}_{i=1}^m) &= \sum_{i=1}^n \left( y^{(i)} - \hat{y}^{(i)} \right)^2 \\ &= \sum_{i=1}^n \left( y^{(i)} - \left( \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \dots + \beta_m x_m^{(i)} + \alpha \right) \right)^2 \end{aligned}$$

データ      モデル

- 最適化問題（最小化）を解いてパラメータ推定値を求める：

$$(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_m) = \operatorname{argmin}_{\alpha, \{\beta_i\}_{i=1}^m} \ell(\alpha, \{\beta_i\}_{i=1}^m)$$

- すべてのパラメータについて偏微分して0とおき連立方程式を得る

# 行列とベクトルを用いた表記：

行列とベクトルを用いて書き換えると便利

- モデル： $y = \boldsymbol{\beta}^\top \mathbf{x}$

- パラメータ： $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_m, \alpha)^\top$

- 独立変数： $\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \dots, x_m^{(i)}, 1)^\top$

} 最後の次元は切片部分に相当

- 目的関数： $\ell(\boldsymbol{\beta}) = \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2$   
 $= \sum_{i=1}^n (y^{(i)} - \boldsymbol{\beta}^\top \mathbf{x}^{(i)})^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$

- 計画行列： $\mathbf{X} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)})^\top$

- 従属変数： $\mathbf{y} = (y^{(1)}, y^{(2)}, \dots, y^{(N)})^\top$

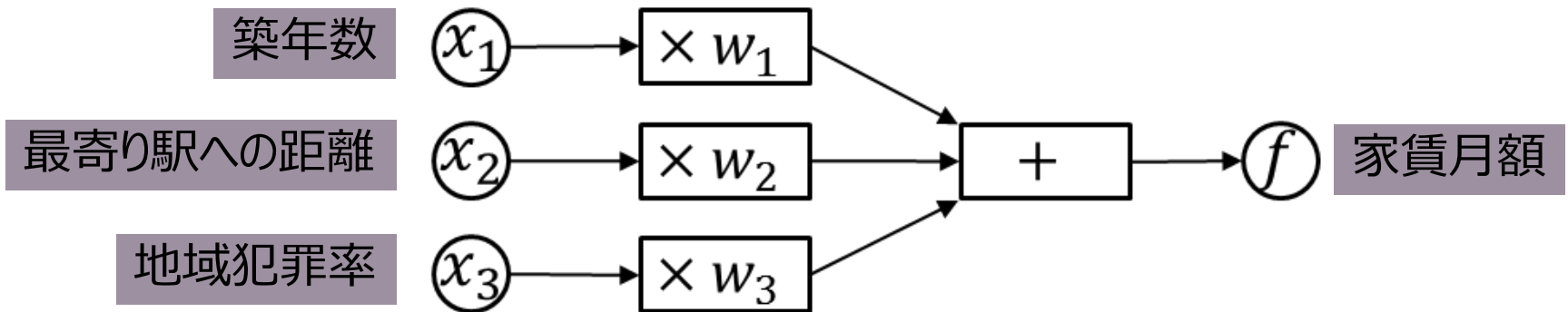
# 例： 家賃予測

- 計画行列：4件の賃貸住宅

$$\mathbf{X} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}]^T = \left[ \begin{pmatrix} 15 \\ 10 \\ 1.0 \end{pmatrix}, \begin{pmatrix} 3 \\ 1 \\ 0.1 \end{pmatrix}, \begin{pmatrix} 35 \\ 5 \\ 7.0 \end{pmatrix}, \begin{pmatrix} 40 \\ 70 \\ 1.0 \end{pmatrix} \right]^T$$

- 独立変数（ベクトル）：4件分の家賃

$$\mathbf{y} = (y^{(1)}, y^{(2)}, y^{(3)}, y^{(4)})^T = (140, 85, 220, 115)^T$$



# 重回帰モデルの解： 解析解が得られる

- 目的関数： $L(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$
- 解： $\boldsymbol{\beta}^* = \operatorname{argmin}_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ 

ただし、本当に（数値的に）解くときには連立方程式のほうを解く
- ただし、解が存在するためには $\mathbf{X}^\top \mathbf{X}$ が正則である必要
  - モデルの次元数 $m$ よりもデータ数 $n$ が大きい場合はおおむね成立
- 正則化：正則でない場合には $\mathbf{X}^\top \mathbf{X}$ の対角成分に正の定数 $\lambda > 0$ を加えて正則にする
  - 新たな解： $\boldsymbol{\beta}^* = \operatorname{argmin}_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$
  - 目的関数に戻すと： $L(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2$

パラメータのノルムに関する  
ペナルティ項

## 多重共線性：

### 独立変数間に強い相関がある場合には注意

- 重回帰モデルにおいて、独立変数間に強い相関がある場合には推定されたパラメータの分散が大きくなり、信頼性が下がる
  - どちらでも説明できるので、パラメータの重みを奪い合う
  - 例：年齢と勤続年数など
- 予測には影響しないが、得られたモデル（パラメータ）を解釈したい場合には注意を要する
  - 相関が強い場合には、片方ずつ用いた結果を調べるなどを行う

# 質的変数の取り扱い

# 質的変数の扱い： ダミー変数の利用

- 独立変数が質的変数（記号を値としてとる）の場合
  - 例： {右, 左}、{京都, 大阪, 東京}
- ダミー変数： {0,1}の2値をとる変数
  - {右, 左}を{0,1}として表現
  - 3値以上の場合には、選択肢数-1個のダミー変数を用いる：  
京都 = (1,0)、大阪 = (0,1)、東京 = (0,0)
- 例： 年齢と性別から年収を予測する
  - 性別が男性であるか {0(No), 1(Yes)} のダミー変数

東京をベースラインとして各地域の差分を示す

$$(\text{年収}) = \beta_1 \times (\text{年齢}) + \beta_2 \times (\text{性別}) + \alpha$$

– 性別が男性であるか {0(No), 1(Yes)} のダミー変数

## 従属変数が質的変数の場合：

ダミー変数を従属変数として回帰を適用（が、やや不適）

- 従属変数が質的変数の場合

- 例：年収と年齢から性別を当てる

- 従属変数をダミー変数として回帰を適用する

- 例：(性別) =  $\beta_1 \times (\text{年齢}) + \beta_2 \times (\text{年収}) + \alpha$

- 回帰モデルの適用は厳密にはちょっと変

- 回帰モデルは連続値を出力するが、本来、性別にあたるダミー変数は{0,1}のいずれかの値のみをとる

- 最小二乗法が仮定している均一分散性が成立しない

- 「効率性」が満たされないため推定値のバラつきが大きい

# 非線形回帰

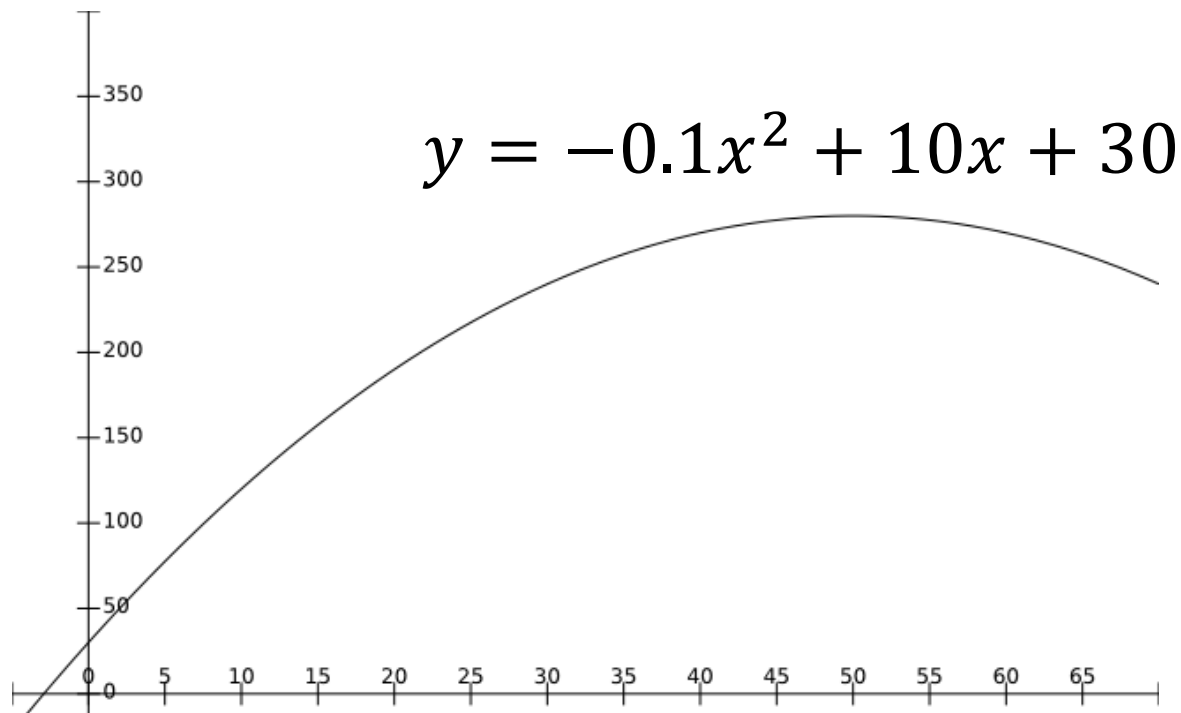
# 非線形回帰： 線形回帰に非線形性を導入する

- ここまでは線形モデルを仮定してきた： $y = \boldsymbol{\beta}^\top \mathbf{x}$ 
  - パラメータ： $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_m, \alpha)^\top$
  - 独立変数： $\mathbf{x} = (x_1, x_2, \dots, x_m, 1)^\top$– シンプルで安定して扱いやすい
- 線形モデルに非線形性を導入するにはどうしたらよいか？
  1. 変数変換（例： $x \rightarrow \log x$ ）
  2. 交差項（例： $x_1, x_2 \rightarrow x_1 x_2$ ）
  3. カーネル法

# 変数変換： 簡単に非線形性を導入する方法

- 独立変数に対して非線形の変換を適用する：

$$x \rightarrow \log x, e^x, x^2, \frac{1}{x}, \dots$$



# 変数の対数変換： 傾きパラメータ $\beta$ の意味が異なる

- $y = \beta x + \alpha$  の独立変数 ( $x$ ) と従属変数 ( $y$ ) は対数変換して用いられることがある
- 変換と係数の意味

		従属変数	
		$y$	$\log y$
独立変数	$x$	$y = \beta x + \alpha$ $x$ が1単位増加すると $y$ が $\beta$ 単位増加する	$\log y = \beta x + \alpha$ $x$ が1単位増加すると $y$ が $1 + \beta$ 倍になる
	$\log x$	$y = \beta \log x + \alpha$ $x$ を2倍すると $y$ が $\beta$ 単位増加する	$\log y = \beta \log x + \alpha$ $x$ を2倍すると $y$ が $1 + \beta$ 倍になる

# 交差項： 変数の組み合わせを導入

- もともとの独立変数  $x_1, x_2, \dots, x_m$  に加えて、2変数の交差項  $\{x_d x_{d'}\}_{d,d'}$  を用いる

– ダミー変数の交差項は2変数のANDに相当

- すべての交差項を採用すると行列パラメータ  $\mathbf{B}$  を導入して  $y = \mathbf{x}^\top \mathbf{B}^\top \mathbf{x}$  と書くことができる

$$y = \text{Trace} \left( \underbrace{\begin{bmatrix} \beta_{1,1} & \cdots & \beta_{1,m} \\ \vdots & \ddots & \vdots \\ \beta_{m,1} & \cdots & \beta_{m,m} \end{bmatrix}}_{\mathbf{B}}^\top \underbrace{\begin{bmatrix} x_1^2 & x_1 x_2 & \cdots & x_1 x_m \\ x_2 x_1 & x_2^2 & \cdots & x_2 x_m \\ \vdots & \vdots & \ddots & \vdots \\ x_m x_1 & x_m x_2 & \cdots & x_m^2 \end{bmatrix}}_{\mathbf{x} \mathbf{x}^\top} \right)$$

# カーネル回帰：

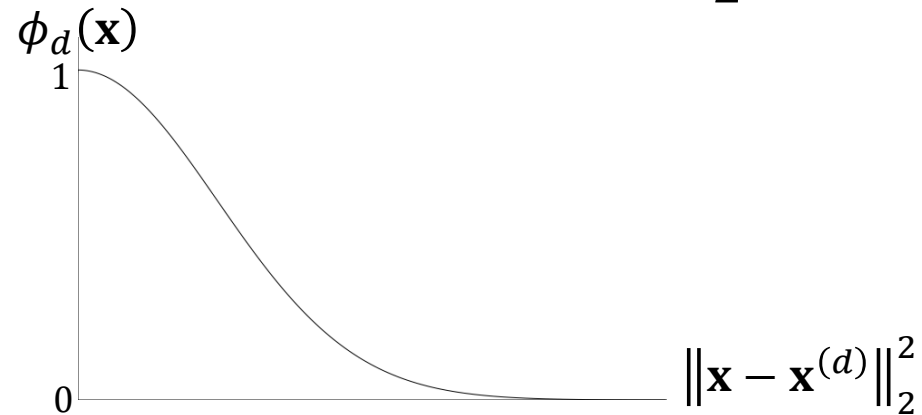
## カーネル関数を用いた非線形性の導入

- 前述の変数変換アプローチを一般化する
- 線形モデル  $y = \boldsymbol{\beta}^\top \mathbf{x}$  において、 $d$  番目の独立変数  $x_d$  を「カーネル関数」をもちいた基底  $\phi_d(\mathbf{x})$  で与える

- カーネル関数  $\phi_d(\mathbf{x})$ ：  
独立変数  $\mathbf{x}$  に何らかの非線形変換を適用したもの

- カーネルの例：ガウスクーネル  $\phi_d(\mathbf{x}) = \exp(-\gamma \|\mathbf{x} - \mathbf{x}^{(d)}\|_2^2)$

– 要するに、 $d$  番目のデータとの「類似度」のようなもの



# カーネル回帰：

## カーネル関数を用いた非線形性の導入

- カーネル回帰モデル：

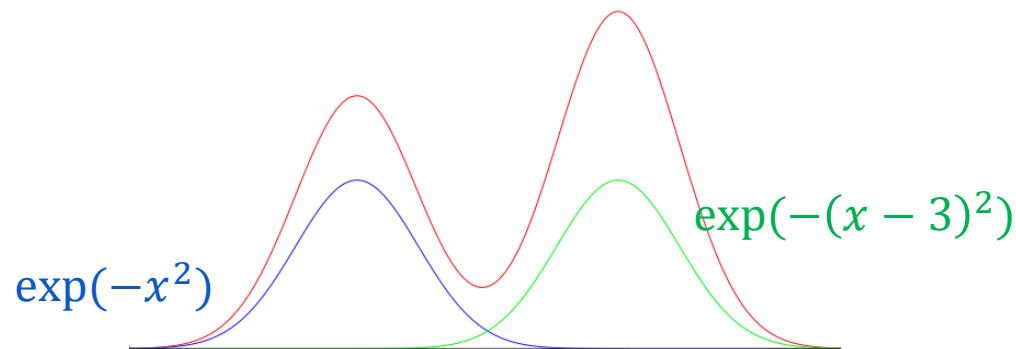
$$y = \boldsymbol{\beta}^\top \boldsymbol{\phi}(\mathbf{x}) = \beta_1 \phi_1(\mathbf{x}) + \beta_2 \phi_2(\mathbf{x}) + \cdots + \beta_n \phi_n(\mathbf{x}) + \alpha$$

–モデルの次元数 $n$ は、もとの $\mathbf{x}$ の次元数 $m$ とは異なることに注意

- 通常はモデルの次元数 $n = \text{データサイズ}$ にとる

– $\phi_d(\mathbf{x})$ は $\mathbf{x}$ と $\mathbf{x}^{(d)}$ の類似度を表すカーネル関数

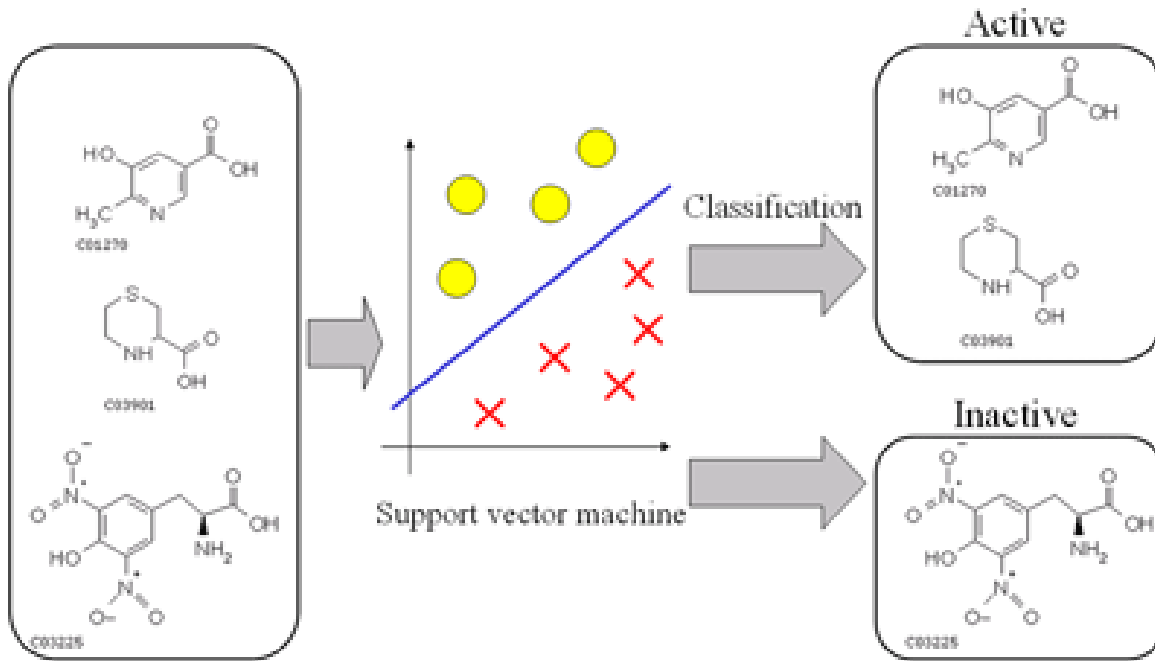
- $n = 2, m = 1$ の例： $y = 1.5 \exp(-x^2) + 2 \exp(-(x - 3)^2)$



# さまざまなカーネル関数：

## カーネル関数を変えれば様々なデータに対応可能

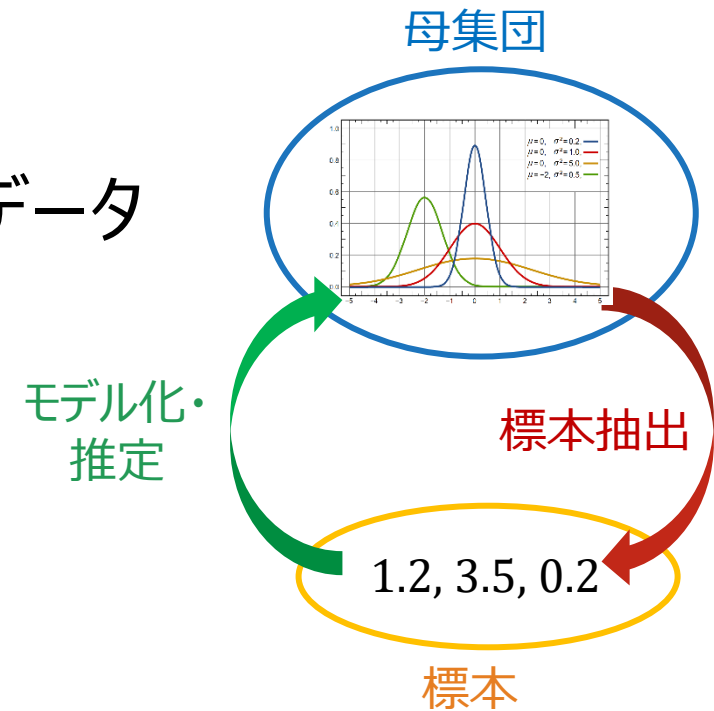
- カーネル回帰はカーネル関数の定義を変えることで、任意の対象を扱うことができる
  - 独立変数がベクトル  $\mathbf{x} = (x_1, x_2, \dots, x_m, 1)^T$  である必要すらない
- カーネル関数によって、系列、木、グラフなども扱うことができる



# 最尤推定

# (あらためて) 統計モデリングの考え方： 部分から全体について知る

- 母集団：確率分布で表される、我々が本当に興味のある集合
  - 分布のクラスやパラメータで指定されるとする
- 標本：実際に観測できる母集団の一部
  - 確率分布に従って抽出された具体的なデータ
- 目的：  
標本から母集団について推測する  
(標本抽出の逆)
  - パラメータを推定する (どうやって?)



# パラメータの推定問題：

サイコロの各目の出る確率を実際の出目から推定する

- 母集団は離散分布に従うとする

–  $P(X = k) = f(k)$  (ただし  $\sum_{k \in \mathcal{X}} f(k) = 1, f(k) \geq 0$ )

– たとえば (厳密な) サイコロであれば  $P(X = k) = \frac{1}{6} \approx 0.17$

- 標本抽出：

– サイコロを20回 (独立に) 振ったところ、

6 3 5 1 3 1 4 1 2 2 6 1 2 2 5 4 4 4 6 5 が出た

出目	1	2	3	4	5	6
回数	4	4	2	4	3	3

- 母集団のパラメータ (それぞれの目の出る確率) を推定したい

# サイコロのパラメータ推定問題へのひとつの解： 出た目の回数の割合で推定する

- ひとつのアイデア：

20回中で1が4回出たのだから  $P(X = 1) \approx \frac{4}{20} = 0.2$  と推定する

出目	1	2	3	4	5	6
回数	4	4	2	4	3	3
確率の推定値	0.2	0.2	0.1	0.2	0.15	0.15

- 正解が約0.17なので悪くない...
- この推定値はどのような原理に基づいているのか？

# 最尤推定： 確率分布の代表的な推定手法のひとつ

---

- 標本からの母集団確率分布の推定
- 代表的な推定手法
  - 最尤推定
  - モーメント推定
  - ベイズ推定
  - ...

# 最尤推定とは：

標本をもっともよく再現するパラメータを推定値とする

- $n$ 個のデータ： $x_1, x_2, \dots, x_n$  が生成される確率（尤度）：

$$L = P(X = x_1)P(X = x_2) \cdots P(X = x_n) = \prod_{i=1}^n P(X = x_i)$$

独立性を仮定しているので積になる

- サイコロの例：

– 目 $k$ が出る確率を $p_k$ , 目 $k$ が出た回数を $n_k$ とする

– 尤度 $L(p_1, p_2, \dots, p_n) = p_1^{n_1} p_2^{n_2} \cdots p_6^{n_6} = \prod_{k=1}^6 p_k^{n_k}$

– これを最大化する $p_1, p_2, \dots, p_n$ を求める（最大化問題を解く）と

$$\hat{p}_k = \frac{n_k}{n_1 + n_2 + \cdots + n_6}$$

# サイコロ（離散分布）の最尤推定： ラグランジュの未定乗数法によって推定値が求まる

- 尤度の代わりに対数尤度を最大化すると扱いやすい（解は変わらない）：

$$\log L(p_1, p_2, \dots, p_n) = \sum_{k=1}^6 n_k \log p_k$$

- 確率分布の制約:  $\sum_{k=1}^6 p_k = 1, p_k > 0$

$\{p_k\}_{k=1}^6, \lambda$  について最大化する

- ラグランジュ未定乗数法：

$$G(\{p_k\}_{k=1}^6, \lambda) = \sum_{k=1}^6 n_k \log p_k + \lambda \left( 1 - \sum_{k=1}^6 p_k \right)$$

# 練習：

## 正規分布のパラメータの最尤推定

- 正規分布： $f(x) = N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- パラメータ：平均 $\mu$ と分散 $\sigma^2$ の最尤推定量を求めてみよう
  1. データ： $x_1, x_2, \dots, x_n$  に対する対数尤度をつくる
  2. パラメータについての最大化問題を解く

$f(x)$

