KYOTO UNIVERSITY

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）

DEPARTMENT OF INTELLIGENCE SCIENCE
AND TECHNOLOGY

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

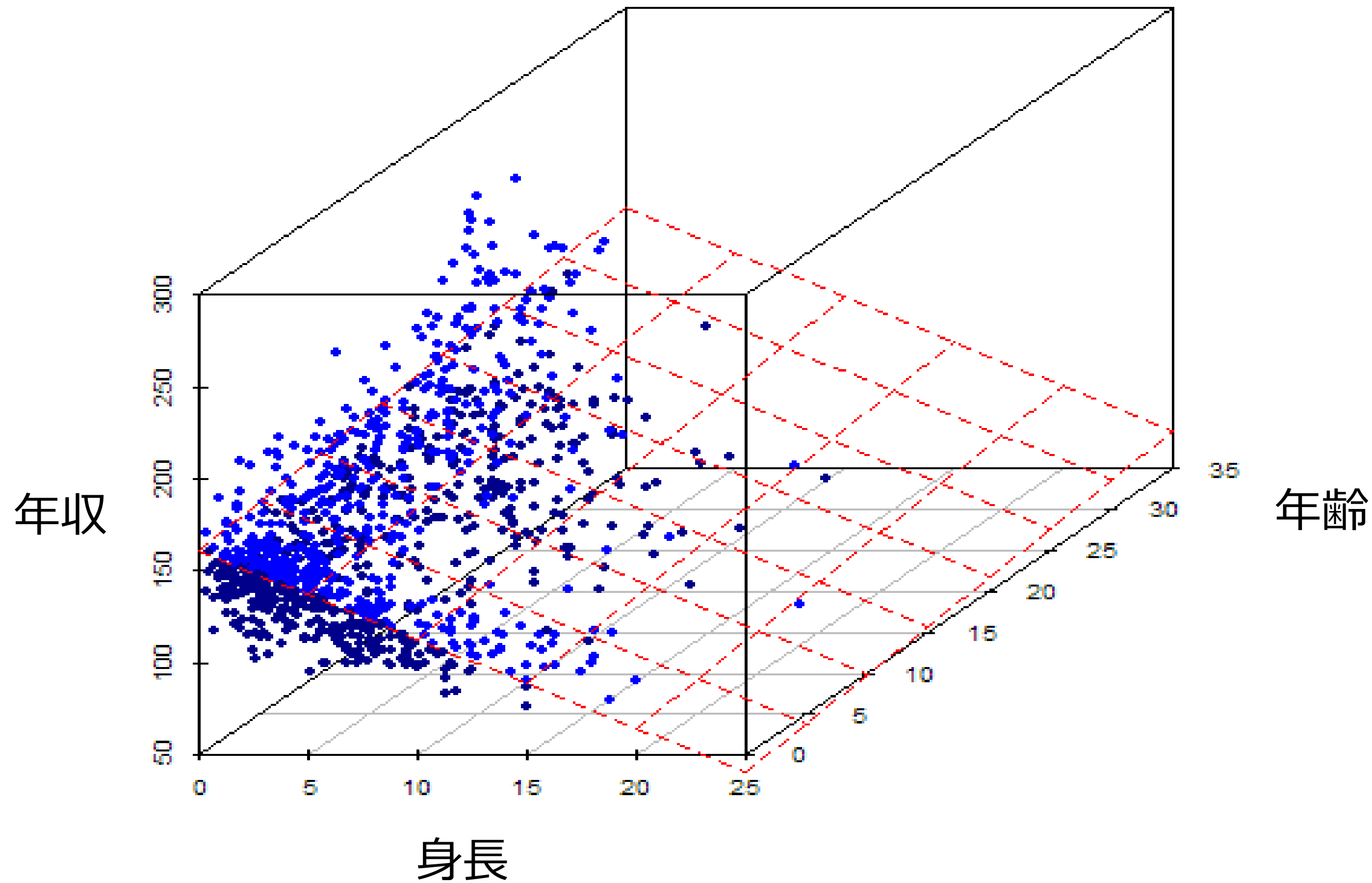
$$g(x) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \alpha$$

  – 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

$$\ell(\alpha, \{\beta_i\}_{i=1}^m) = \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

$$= \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \cdots + \beta_m x_m^{(i)} + \alpha\right)\right)^2$$

- 最適化問題（最小化）を解いてパラメータ推定値を求める：

$$\left(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_m\right) = \mathrm{argmin}_{\alpha, \{\beta_i\}_{i=1}^m}\, \ell(\alpha, \{\beta_i\}_{i=1}^m)$$

  – 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：$\mathbf{x}^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \dots, x_m^{(i)}, 1\right)^\top$

最後の次元は切片部分に相当

- 目的関数：$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2$

$$= \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\beta}^\top \mathbf{x}^{(i)}\right)^2 = \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

  - 計画行列： $\boldsymbol{X} = \left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}\right)^\top$
  - 従属変数： $\mathbf{y} = \left(y^{(1)}, y^{(2)}, \dots, y^{(N)}\right)^\top$

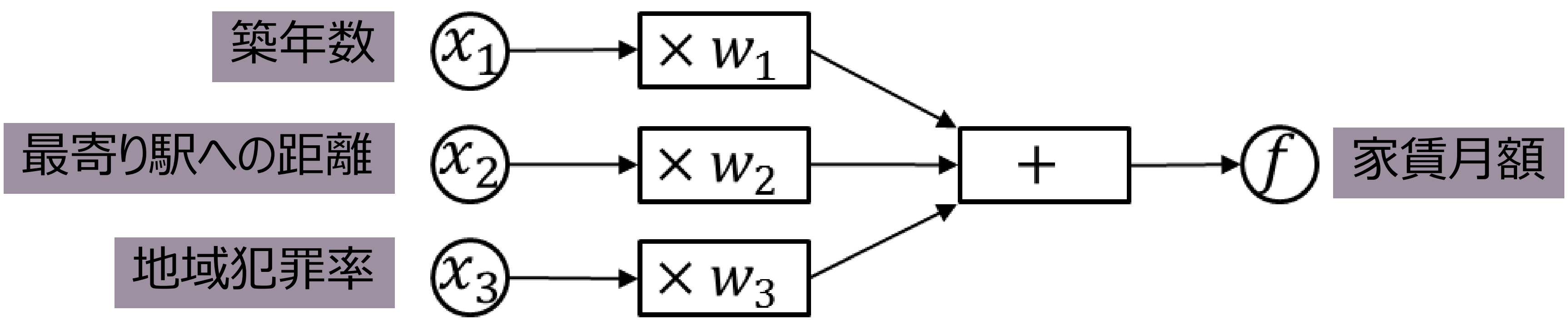
例：家賃予測

- 計画行列：4件の賃貸住宅

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

- 独立変数（ベクトル）：4件分の家賃

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



重回帰モデルの解：解析解が得られる

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

  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$\boldsymbol{X}^\top\boldsymbol{X}$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$\boldsymbol{X}^\top\boldsymbol{X}$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: $\boldsymbol{\beta}^* = \mathrm{argmin}_{\boldsymbol{\beta}}\, L(\boldsymbol{\beta}) = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top\mathbf{y}$
  - 目的関数に戻すと：$L(\boldsymbol{\beta}) = \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$

    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)

東京をベースラインとして各地域の差分を示す

- 例：年齢と性別から年収を予測する

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

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

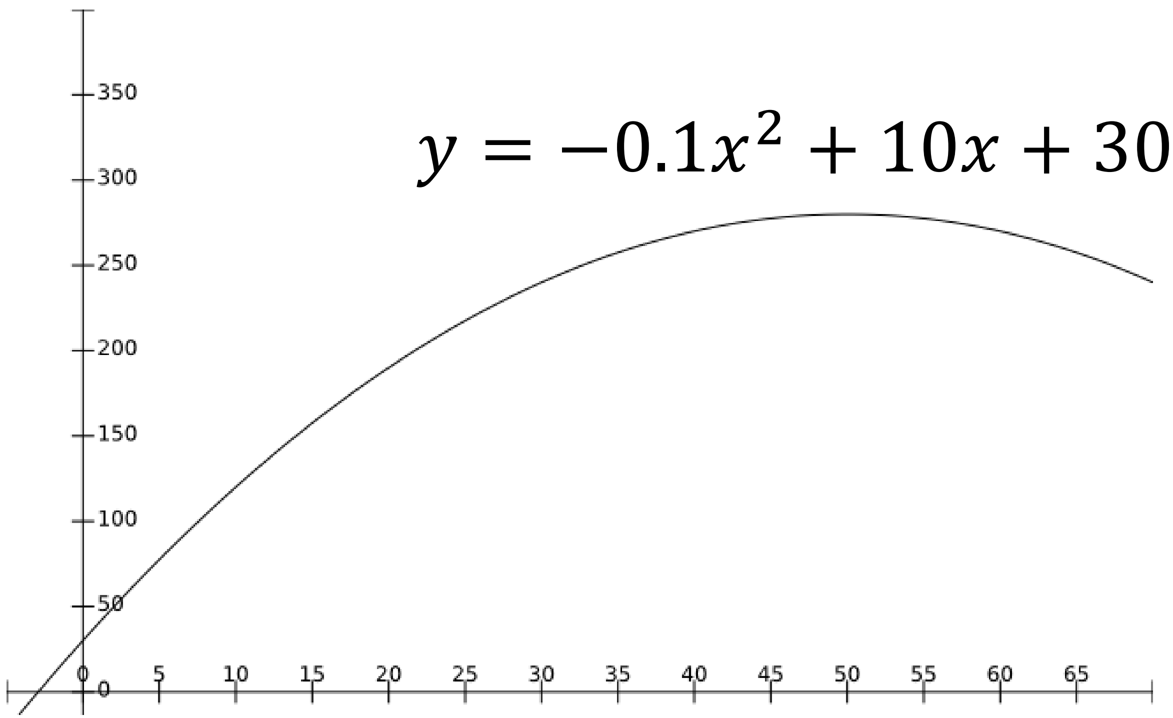
1. 変数変換（例： $x \to \log x$）
2. 交差項（例：$x_1, x_2 \to x_1 x_2$）
3. カーネル法

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：

$$x \to \log x\,, e^{x}, x^{2}, \frac{1}{x}, \ldots$$

$$y = -0.1x^2 + 10x + 30$$



変数の対数変換：傾きパラメータβ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β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β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$x_1, x_2, \ldots, x_m$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**B**を導入して $y = \mathbf{x}^\top \mathbf{B}^\top \mathbf{x}$ と書くことができる

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

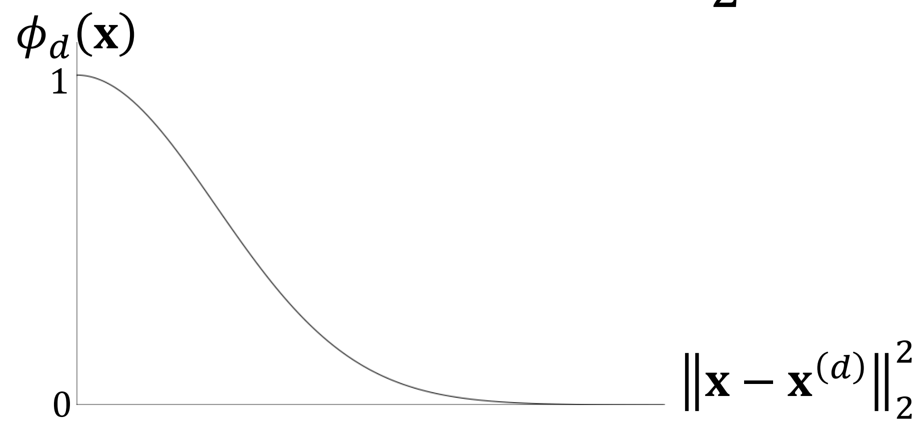
 ※ $\mathbf{B} = \mathbf{U}\mathbf{U}^\top$ として**B**の低ランク分解を仮定するとFactorization Machine

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：ガウスカーネル $\phi_d(\mathbf{x}) = \exp(-\gamma \left\| \mathbf{x} - \mathbf{x}^{(d)} \right\|_2^2)$
  - 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

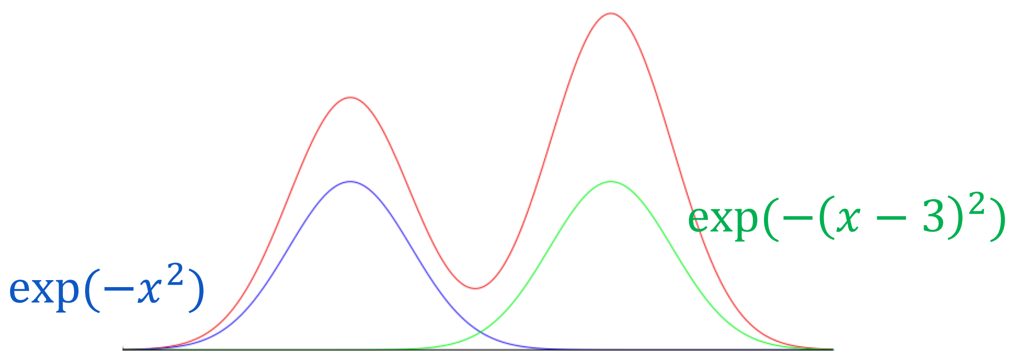
- カーネル回帰モデル：

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

  – 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 ＝データサイズにとる

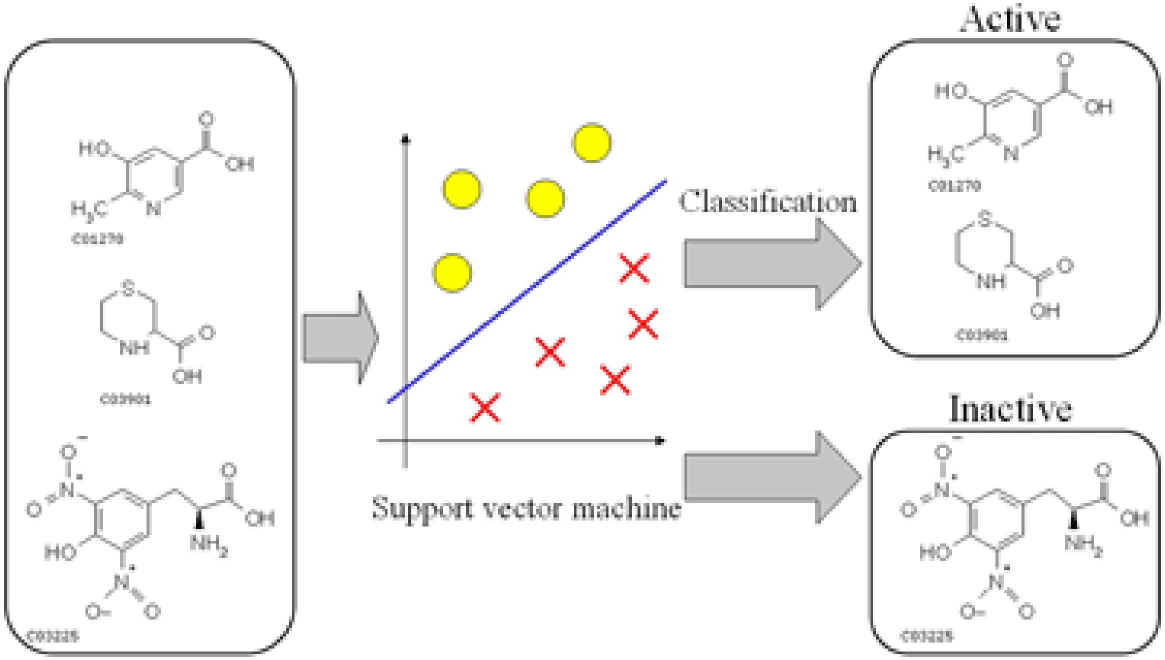
  – $\phi_d(\mathbf{x})$は$\mathbf{x}$と$\mathbf{x}^{(d)}$の類似度を表すカーネル関数

- $n = 2, m = 1$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 $\mathbf{x} = (x_1, x_2, \dots, x_m, 1)^\top$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

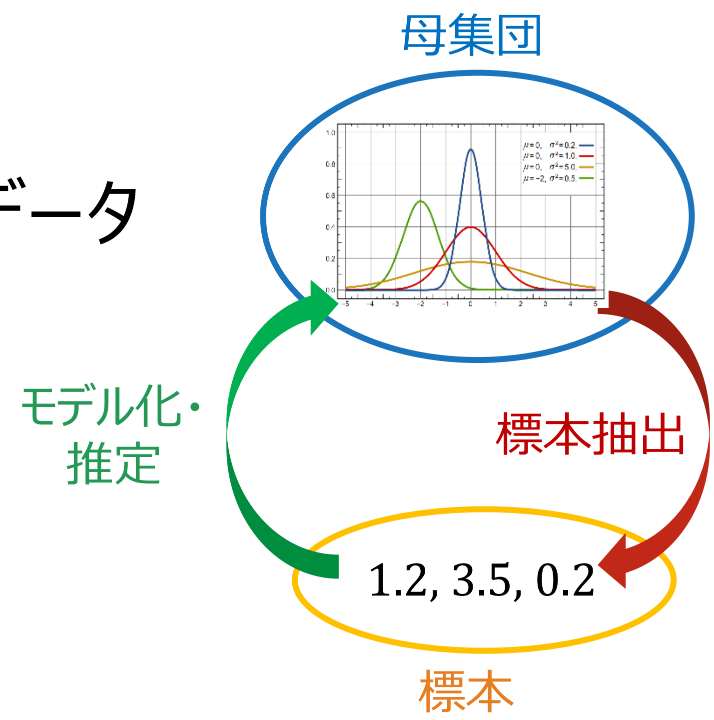


http://www.bic.kyoto-u.ac.jp/coe/img/akutsu_fig_e_02.gif

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

- ひとつのアイディア：
  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： $x_1, x_2, \ldots, x_n$ 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$L(p_1, p_2, \ldots, p_n) = p_1^{n_1} p_2^{n_2} \cdots p_6^{n_6} = \prod_{k=1}^{6} p_k^{n_k}$
  - これを最大化する$p_1, p_2, \ldots, p_n$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: $\sum_{k=1}^{6} p_k = 1, p_k > 0$
- ラグランジュ未定乗数法：

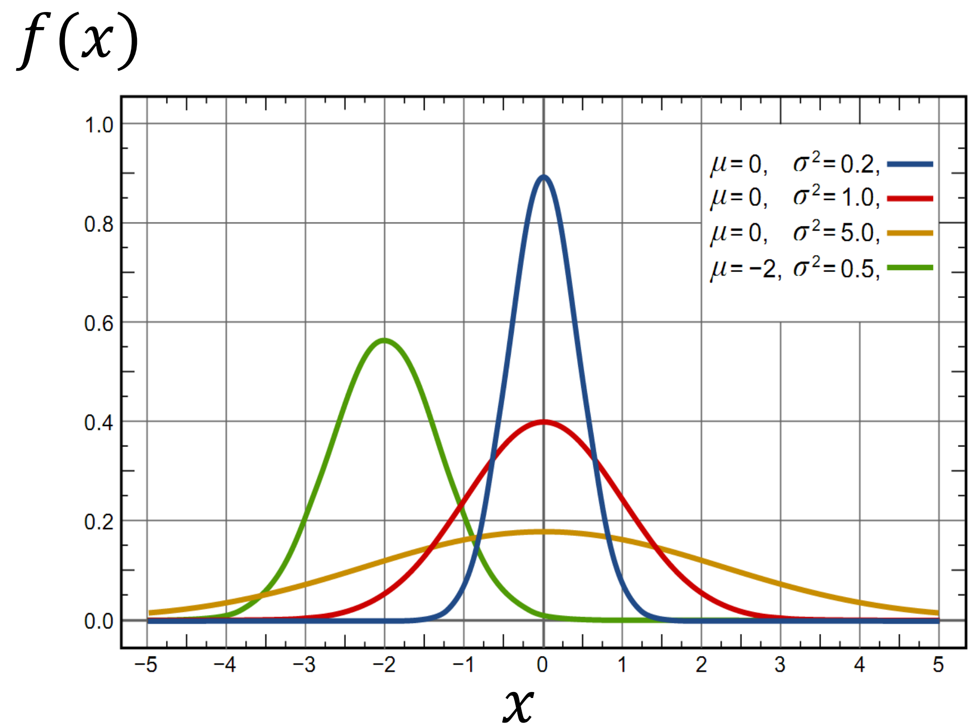
$\{p_k\}_{k=1}^{6}, \lambda$ について最大化する

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

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μと分散σ^2の最尤推定量を求めてみよう

1. データ：$x_1, x_2, \ldots, x_n$ 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



ベイズ決定

応用問題：
どちらのサイコロが使われた？

- ２つの（いびつな）サイコロA, Bがある
  - サイコロAを20回振ったところ：

出目	1	2	3	4	5	6
回数	5	1	4	2	4	4

  - サイコロBを16回振ったところ：

出目	1	2	3	4	5	6
回数	2	8	2	2	1	1

応用問題：
どちらのサイコロが使われた？

- （いびつな）サイコロA, Bのパラメータの最尤推定値は：
  - サイコロA：

出目	1	2	3	4	5	6
確率	5/20	1/20	4/20	2/20	4/20	4/20

  - サイコロB：

出目	1	2	3	4	5	6
確率	2/16	8/16	2/16	2/16	1/16	1/16

応用問題：
どちらのサイコロが使われた？

- （いびつな）サイコロA, Bのパラメータの最尤推定値は：
  - サイコロA：

出目	1	2	3	4	5	6
確率	5/20	1/20	4/20	2/20	4/20	4/20

  - サイコロB：

出目	1	2	3	4	5	6
確率	2/16	8/16	2/16	2/16	1/16	1/16

- 今、2つのサイコロのいずれかを選んで（Cとする）5回振ったところ：

出目	1	2	3	4	5	6
回数	1	1	0	2	0	1

- 使われたサイコロはA, Bのいずれだろうか？（C=A or C=B?）

ベイズ決定：
事後確率によって決定する

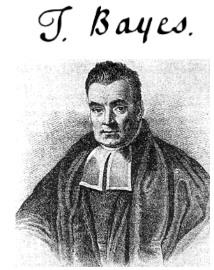
- A, B どちらのサイコロを選んだかを確率変数Xで表す
  - 事前確率：でたらめに選ぶと $P(X = \mathrm{A}) = P(X = \mathrm{B}) = 1/2$
  - 何も情報がなければこれ以上はわからない
- 事後分布：C（A, Bのいずれか）を振って出たデータ$\mathcal{D}$を見たあとの、Xの確率分布 $P(X|\mathcal{D})$

出目	1	2	3	4	5	6
回数	1	1	0	2	0	1

- ベイズ決定：事後確率が$P(X = \mathrm{A}|\mathcal{D}) > P(X = \mathrm{B}|\mathcal{D})$であれば、Aが使われた可能性が高いと判断できる
- 事後確率の計算： $P(X|\mathcal{D}) = \dfrac{P(\mathcal{D}|X)P(X)}{P(\mathcal{D})}$ （ベイズの定理）

事後確率の計算：ベイズの定理と最尤推定で事後確率を計算

- 事後確率の計算には「ベイズの定理」をつかう：



$$P(X|\mathcal{D}) = \frac{P(\mathcal{D}|X)P(X)}{P(\mathcal{D})}$$

$P(X)$は事前確率（今回は1/2）

  - 判断基準：$P(X = \mathrm{A}|\mathcal{D}) \gtrless P(X = \mathrm{B}|\mathcal{D})$
    $\leftrightarrow\ P(\mathcal{D}|X = \mathrm{A})P(X = \mathrm{A}) \gtrless P(\mathcal{D}|X = \mathrm{B})P(X = \mathrm{B})$
    - 注意：分母$P(\mathcal{D}) = \sum_X P(\mathcal{D}|X)P(X)$ は今回は計算する必要はない

- サイコロのパラメータ$\{p_k^{\mathrm{A}}\}_{k=1}^6$, $\{p_k^{\mathrm{B}}\}_{k=1}^6$は最尤推定によって推定

サイコロCの出目回数

- $P(\mathcal{D}|X = \mathrm{A}) = \prod_{k=1}^6 {p_k^{\mathrm{A}}}^{n_k^{\mathrm{C}}} \gtrless P(\mathcal{D}|X = \mathrm{B}) = \prod_{k=1}^6 {p_k^{\mathrm{B}}}^{n_k^{\mathrm{C}}}$で判断

線形回帰モデルの確率的解釈

最尤推定：
データ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 $x_1, x_2, \ldots, x_n$ から確率モデル$f(x \mid \theta)$のパラメータθを推定したい

- n個のデータが（互いに独立に）生成される確率（尤度）：

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

- 尤度最大になるパラメータを推定値$\hat{\theta}$とする

実際には対数尤度で扱うことが多い

$$\hat{\theta} = \mathrm{argmax}_\theta \prod_{i=1}^{n} f(x_i \mid \theta) = \mathrm{argmax}_\theta \sum_{i=1}^{n} \log f(x_i \mid \theta)$$

– もっともデータを生成する確率が高い（「最も尤もらしい」）

線形回帰モデルの最尤推定：線形回帰の確率モデルを考える

- データ：$\mathbf{x} = (x^{(1)}, x^{(2)}, \dots, x^{(n)})$ と $\mathbf{y} = \left(y^{(1)}, y^{(2)}, \dots, y^{(n)}\right)$ に線形モデル： $g(x) = \beta x + \alpha$ を当てはめる

- 最小二乗法：$\ell(\alpha, \beta) = \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta x^{(i)} + \alpha\right)\right)^2$ を最小化

- 一方、線形回帰モデルに対応する確率モデルを仮定する：
  - 正規分布：$y^{(i)}$は平均$\beta x^{(i)} + \alpha$, 分散σ^2の正規分布に従う
  - 確率密度：$f\left(y^{(i)} \mid x^{(i)}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - (\beta x^{(i)} + \alpha)\right)^2}{2\sigma^2}\right)$
  - 「平均的に」回帰直線 $y = \beta x + \alpha$ に乗るデータを生成するモデル

線形回帰モデルの最尤推定：
線形回帰の確率モデルの最尤推定 ＝ 最小二乗法

- 線形回帰モデルに対応する確率モデルを考える：

- 確率密度関数：$f\left(y^{(i)} \mid x^{(i)}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)}-(\beta x^{(i)}+\alpha)\right)^2}{2\sigma^2}\right)$

- 対数尤度：$L(\alpha, \beta) \ = \sum_{i=1}^{n} \log f\left(y^{(i)} \mid x^{(i)}\right)$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - (\beta x^{(i)} + \alpha)\right)^2 + \text{const.}$$

- 対数尤度をα, βについて最大化すること（最尤推定）
 ＝二乗誤差をα, βについて最小化すること（最小二乗法）

線形回帰モデルの最尤推定：
分散の最尤推定量

- 確率密度関数：$f\left(y^{(i)} \mid x^{(i)}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)}-(\beta x^{(i)}+\alpha)\right)^2}{2\sigma^2}\right)$

- 分散については、対数尤度：

$$L(\sigma^2) \; = n \log \frac{1}{\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - (\beta x^{(i)} + \alpha)\right)^2 + \text{const.}$$

- $L(\sigma^2)$を最大化する最尤推定量は：

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - (\beta x^{(i)} + \alpha)\right)^2$$

※ 以上の議論は重回帰モデルの場合も同様

最尤推定の性質

最尤推定量の性質：一致性

- パラメータθの推定量として$\hat{\theta}$を得たとする（例えば最尤推定で）
- 推定量の良さはどのように評価するか？
  - 不偏性 $E[\hat{\theta}] = \theta$：推定量の期待値が真の値に一致する
    - Eは様々な標本の採り方についての期待値を表す
    - たとえば、平均の最尤推定量は不偏性をもつが、分散の最尤推定量はもたない
  - 一致性：標本サイズを大きくしていくと真の値に一致する：

$$\hat{\theta} \xrightarrow[n\to\infty]{} \theta$$

- 最尤推定は、適当な条件のもと一致性をもつ

漸近正規性：
最尤推定は漸近正規性をもつ

- 最尤推定量の分布は$n \to \infty$で、真のパラメータθを平均とする正規分布に従う

- もう少し厳密にいうと：
$\sqrt{n}\left(\hat{\theta} - \theta\right)$の分布が平均0、分散$I(\theta)^{-1}$の正規分布に近づく
  - $I(\theta)$はフィッシャー情報量：
$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right]$$
$$= -\int\left(\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right)f(x|\theta)\mathrm{d}x$$

$I(\theta)$が大きいほど、対数尤度関数が、真値の周りで「尖っている」イメージ

  - $n \to \infty$で$\hat{\theta} \to \theta$

ポアソン回帰

最尤推定：
データ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 $x_1, x_2, \ldots, x_n$ から確率モデル$f(x \mid \theta)$のパラメータθを推定したい

- n個のデータが（互いに独立に）生成される確率（尤度）：

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

実際には対数尤度で扱うことが多い

- 尤度最大になるパラメータを推定値$\hat{\theta}$とする

$$\hat{\theta} = \mathrm{argmax}_\theta \prod_{i=1}^{n} f(x_i \mid \theta) = \mathrm{argmax}_\theta \sum_{i=1}^{n} \log f(x_i \mid \theta)$$

– もっともデータを生成する確率が高い（「最も尤もらしい」）

ポアソン分布の最尤推定：
標本平均がパラメータの最尤推定量になる

- ポアソン分布：$P(Y = y \mid \lambda) = \dfrac{\lambda^y}{y!}\exp(-\lambda)$

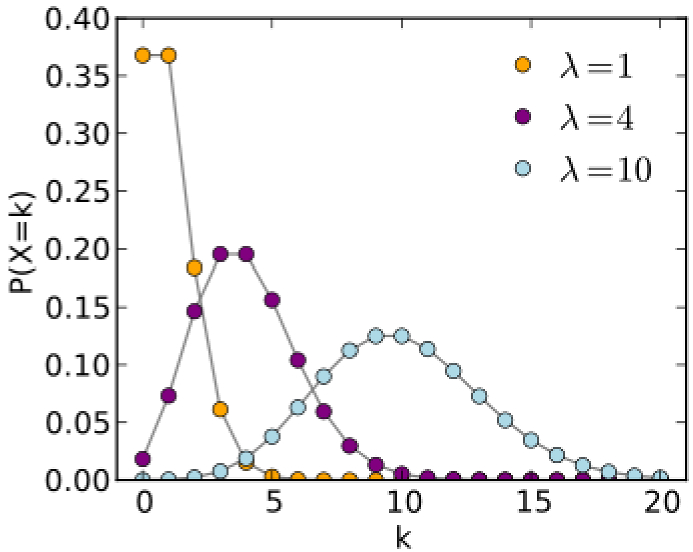
  $\lambda > 0$は平均に相当するパラメータ

- データ：$y_1, y_2, \ldots, y_n$ に対する対数尤度：

$$L(\lambda) = \sum_{i=1}^{n} \log P(Y = y_i \mid \lambda) = \log\lambda \sum_{i=1}^{n} y_i - n\lambda + \mathrm{const.}$$

- パラメータの最尤推定量：

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} y_i}{n}$$



https://en.wikipedia.org/wiki/Poisson_distribution#/media/File:Poisson_pmf.svg

ポアソン回帰：
非負整数の回帰モデル

- 例えば、ある機械の各日の故障件数をモデル化したいとする
  - 曜日や気温などに依存して平均的な故障件数が変わるとする
- 独立変数（曜日など）に依存する回数のモデル：ポアソン回帰

$$P(Y = y \mid \mathbf{x}, \boldsymbol{\beta}) = \frac{\left(\exp(\boldsymbol{\beta}^\top \mathbf{x})\right)^y}{y!} \exp(-\exp(\boldsymbol{\beta}^\top \mathbf{x}))$$

  - ポアソン分布の平均が線形モデルで表されるとする：
    - ポアソン分布：$P(Y = y \mid \lambda) = \frac{\lambda^y}{y!} \exp(-\lambda)$
    - 重回帰モデル：$\lambda = \exp(\boldsymbol{\beta}^\top \mathbf{x})$

組み合わせる

ポアソン回帰の最尤推定：解析解は得られなさそう…

- 独立変数：$\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}\right)$ # n日分の測定
- 従属変数：$\left(y^{(1)}, y^{(2)}, \dots, y^{(n)}\right)$ # n日分の故障数
- 対数尤度（最大化問題）：

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log \frac{\left(\exp\left(\boldsymbol{\beta}^\top \mathbf{x}^{(i)}\right)\right)^{y^{(i)}}}{y^{(i)}!} \exp\left(-\exp\left(\boldsymbol{\beta}^\top \mathbf{x}^{(i)}\right)\right)$$

$$= \sum_{i=1}^{n} y^{(i)} \boldsymbol{\beta}^\top \mathbf{x}^{(i)} - \sum_{i=1}^{n} \exp\left(\boldsymbol{\beta}^\top \mathbf{x}^{(i)}\right) + \mathrm{const.}$$

- これを最大化する$\boldsymbol{\beta}$を求めたいが、解析解は得られない

最尤推定の利点：
モデリングの自動化

- 最尤推定の利点：確率モデルの形（データの生成プロセスの仮定）を決めればモデルパラメータが自動的に決まる
  - ただし、最後に最大化問題を解いて、パラメータ推定量を求める必要がある
  - 離散分布、ポアソン分布、正規分布などは解析的に解が求まる
  - 線形回帰（正規分布でノイズが載る）は連立方程式（いちおう解析的な解）
  - ただし、他の多くのモデルでは、最適化問題を数値的に解く必要がある