情報数学II
第10回 確率・統計3

小野順貴
(ONO, Nobutaka)
東京都立大学　システムデザイン学部
情報科学科　教授
onono@tmu.ac.jp

推定問題

- 与えられたデータから、何らかの変数（パラメータ）を推定する問題を考える。
  - 例：重心が偏ったいびつなコインがある。通常のコインのように、表、裏が出る確率を同様に確からしいとは考えられない。このコインを10回投げたところ、7回表がでた。表が出る確率を b とするとき、bをどのように推定すればよいだろうか？
  - この問題であれば、b=7/10=0.7と推定するのが妥当なように思える。しかし、より複雑な問題にも適用できるような、一般的なアプローチはどうすればよいだろうか？

最尤法

- 観測されたデータから、確率モデルのパラメータを最も尤も(もっとも)らしい、という基準で推定する手法
- 概要
  - 確率変数 : x (一般には多次元ベクトル)
  - 確率密度関数 : $p(x;\theta)$
  - パラメータ : θ (一般には多次元ベクトル)

  とするとき、$L(\theta;x)=p(x;\theta)$ を尤度関数という。
    - 確率密度関数は、x の関数。x で積分すると1になる。
    - 尤度関数は θ の関数。θ で積分しても1にはならない。
  - 観測 x に対し、以下の推定を最尤法という。

$$\hat{\theta}=\mathrm{argmax}_{\theta}L(\theta;x)$$

$\max_{\theta}f(\theta)$ は $f(\theta)$ の最大値

$\mathrm{argmax}_{\theta}f(\theta)$ は $f(\theta)$ の最大値を与える θ

二項分布の例

- 確率 b で表が出るコインを n 回投げ、x 回表が出る確率を考える
  - n回の中でx回表になる組み合わせ(表表裏、など): ${}_nC_x$ 通り
  - それぞれの組み合わせの確率: $b^n(1-b)^{(n-x)}$
  - 求める確率: ${}_nC_x b^n(1-b)^{(n-x)}$
- このとき b が、前ページのパラメータ θ にあたる。
  (簡単のため、n は定数とする)

$$L(b;x) = p(x;b) = \underline{{}_nC_x b^x(1-b)^{(n-x)}}$$

二項分布と呼ばれる

- 最尤法で b を求めるには、与えられた x に対して、
  $L(b;x)$ を最大にする b を求めればよい

対数尤度関数の最大化

- 二項分布に限らず多くの場合、尤度関数そのものの最大化より、その対数である対数尤度関数の最大化を考える方が容易
  - 対数関数は単調なので、$\mathrm{argmax}_\theta f(\theta) = \mathrm{argmax}_\theta \log f(\theta)$
  - 独立な事象の確率は積で表され、対数をとることによって和になるため
- 二項分布の例：

$$\log L(b;x) = \log {}_nC_x + x\log b + (n-x)\log(1-b)$$

  - 最大値を求めるために b で微分すると

$$\frac{d\log L(b;x)}{db} = \frac{x}{b} - \frac{n-x}{1-b}$$

  - =0として b について解くと

$$x(1-b) - (n-x)b = 0 \quad \Rightarrow \quad b = \frac{x}{n}$$

直感にも
合致した値が
求まった

MATLAB演習10

- $n = 10,\ x = 7$ として実際に尤度関数

$$L(b;x) = {}_nC_x b^x (1-b)^{(n-x)}$$

の概形をMATLABで求め、数値的に最大値を求めよ。

  - ${}_nC_x$ は、MATLABでは、nchoosek(n,x)で計算できる。
  - b は0から1までの値をとるので、
    適当な刻み幅（例えば0.01）をとって求めよ。

MATLAB演習10解答例

```
n=10;
x=7;
b=0:0.01:1;
logL=nchoosek(n,x)*b.^x.*(1-b).^(n-x);
plot(b,logL);
xlabel(‘b’);
ylabel(‘log likelihood’);
```

正規分布の例

- データ $x_1, \cdots, x_N$ がそれぞれ独立に同一の正規分布に従うとき、正規分布のパラメータ μ, σ^2 を求めたい
  - 複数の確率変数が独立に同一の分布に従うことを独立同分布 (independent and identically distributed; i.i.d.) という
  - 正規分布 : $p(x; \boldsymbol{\theta}) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
  - パラメータ: $\boldsymbol{\theta} = (\mu, \underline{\sigma^2})$

ひとつの文字として扱う

正規分布の対数尤度関数

- 複数の確率変数の確率密度関数を同時確率密度関数という。
- $\boldsymbol{x} = (x_1, \cdots, x_N)$ の同時確率密度関数

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

  - $x_1, \cdots, x_N$ は独立なので、同時確率密度関数はそれらの積
  - $\prod$ は要素すべてを乗じることを表す記号で、総乗ともよばれる。
    - 要素すべての和（総和）を表す記号は $\sum$
- 対数尤度関数：

$$\log L(\boldsymbol{\theta}; \boldsymbol{x}) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2$$

正規分布の平均、分散の最尤推定量

- パラメータが2つ(平均、分散)あるので、対数尤度関数を平均、分散で微分して0とおき連立方程式を解く。
- 対数尤度関数の微分

$$\frac{\partial \log L(\boldsymbol{\theta};\boldsymbol{x})}{\partial \mu} = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu) = 0$$

$$\frac{\partial \log L(\boldsymbol{\theta};\boldsymbol{x})}{\partial (\sigma^2)} = -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(x_n-\mu)^2 = 0$$

- 最尤推定量

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n$$

標本平均に等しい

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2$$

標本分散に等しい

MATLAB演習11

- $x_1, \cdots, x_N$ として、平均0、分散1の正規分布に従う標本を100個生成し（つまり $N = 100$ ）、その標本に対して対数尤度関数

$$\log L(\boldsymbol{\theta}; \boldsymbol{x}) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2$$

を求め、$\sigma^2 = 1$ 付近で最大値をとることを確認せよ。
ただし、対数尤度関数は $\mu = 0$ とし、σ^2 の関数として表せ。
また、σ^2 の範囲は0.5～3とせよ。

MATLAB演習11解答例

```
N=100;
x=randn(1,N);
sigma2=0.5:0.01:3;
logL=-N/2*log(2*pi)-N/2*log(sigma2)-sum(x.^2)./(2*sigma2);
plot(sigma2,logL);
xlabel(‘sigma^2’);
ylabel(‘log likelihood’);
```

不偏推定量

- 推定量も「確率変数」である
  - $\boldsymbol{x} = (x_1, \cdots, x_N)$ が確率変数ならば、

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$$ も確率変数

真の値と区別するため、ここではハットをつける

  - 推定量の期待値が真の値に一致するとき、これを不変推定量と呼ぶ。
  - 例えば標本平均は、以下より不偏推定量

$$E[\hat{\mu}] = \frac{1}{N}\sum_{n=1}^{N} E[x_n] = \frac{1}{N}\sum_{n=1}^{N} \mu = \mu$$

不偏分散

- 最尤推定量は必ず不偏推定量であるとは限らない
  - 標本数が増えるにつれ、漸近的に不偏に近づく(一致性)
- 正規分布の分散の最尤推定量は不偏推定量ではない

$$E[\hat{\sigma}^2] = \frac{1}{N}\sum_{n=1}^{N} E[(x_n - \hat{\mu})^2] = \frac{N-1}{N}\sigma^2$$

標本分散は真の分散より小さく見積もってしまう。
導出はやや煩雑なので省略するが、真の平均
ではなく、標本平均を用いていることが要因。

- 分散の不偏推定量

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$$

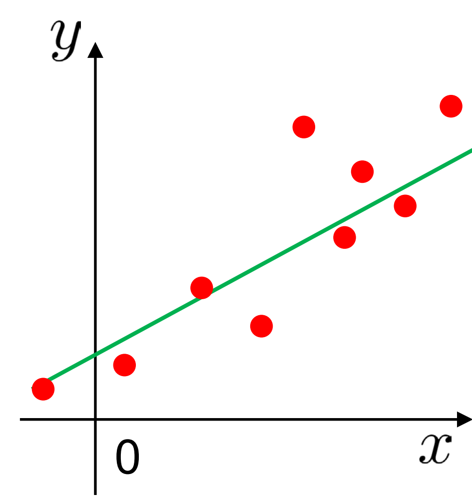
復習

線形回帰分析

- 1変数の線形回帰問題とは、与えられたデータ (x_k, y_k) $(k = 1, \cdots, K)$に対し、線形回帰式 $y = ax + b$ が最もよくフィットするようにパラメータ (a, b) を求める問題。
- フィッティングの良さとしては二乗誤差がよく用いられる。

$$f(a,b) = \sum_{k=1}^{K} (\underbrace{ax_k + b}_{\text{線形回帰モデル}} - \underbrace{y_k}_{\text{データ}})^2$$

モデルとデータの二乗誤差



- どうやって二乗誤差を最小にする (a, b) を求めるか？

最小二乗法と最尤推定

- 2つのデータ $\boldsymbol{x} = (x_1, \cdots, x_N)$ と $\boldsymbol{y} = (y_1, \cdots, y_N)$ の間に、

  $$y_n = ax_n + b + \varepsilon_n$$

  という線形関係が成り立つことを仮定する。ただし、ε_n はモデル化誤差を表し、独立に平均0、分散 σ^2 の正規分布に従うものとする。このとき、a, b を最尤法で求めるとどうなるだろうか？
- ε_n が正規分布に従うのであるから、a, b, x_n が与えられた場合の y_n の確率密度関数は以下で与えられる。

$$p(y_n; a, b, x_n) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(ax_n + b - y_n)^2}{2\sigma^2}}$$

最小二乗法と最尤推定

- 尤度関数

$$L(a,b;\boldsymbol{x},\boldsymbol{y}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \prod_{n=1}^{N} e^{-\frac{(ax_n+b-y_n)^2}{2\sigma^2}}$$

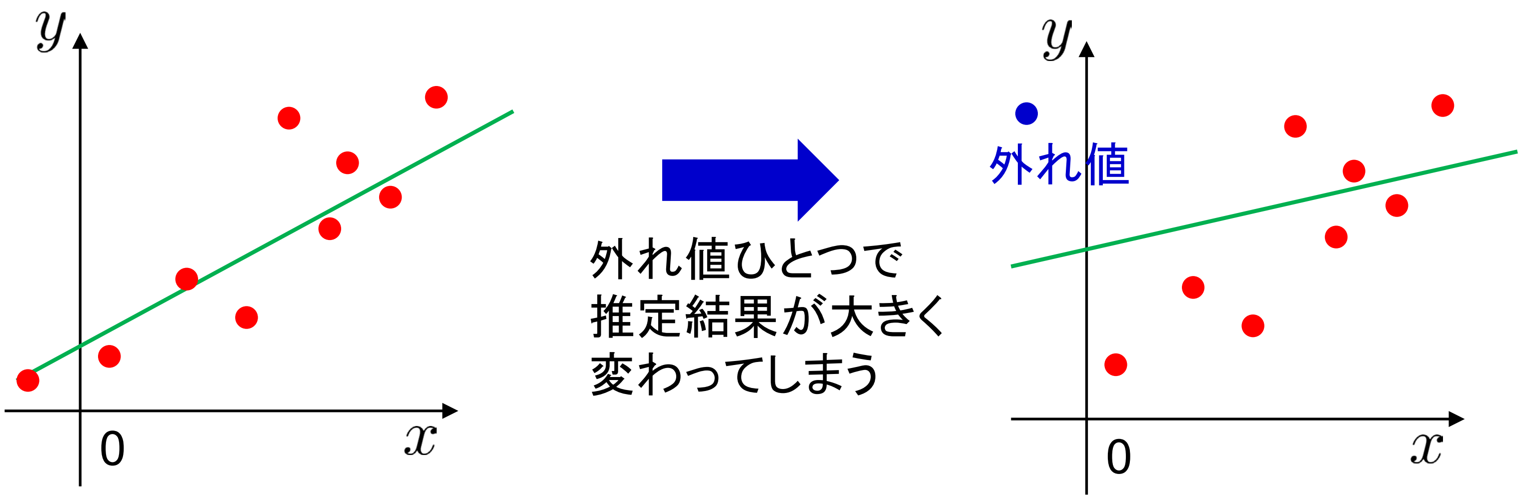
- 対数尤度関数

$$\log L(a,b;\boldsymbol{x},\boldsymbol{y}) = \underbrace{-\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2}_{a,b\text{に依らない定数項}} - \underbrace{\frac{1}{2\sigma^2}}_{a,b\text{に依らない係数}}\sum_{n=1}^{N}\underbrace{(ax_n+b-y_n)^2}_{\text{二乗誤差}}$$

二乗誤差の最小化は、モデル化誤差に正規分布を仮定した最尤推定と等価

外れ値に対する影響

- 最小二乗法は外れ値に弱いことが知られている。



- なぜか？前ページから、最小二乗法は、モデル化誤差に正規分布を仮定して最尤推定していることと等価。しかし、正規分布からは外れ値は極めて稀にしか発生しない。このミスマッチが原因。

外れ値に対する影響の緩和

- もし、データに外れ値が含まれており、正規分布の仮定が適切でない場合には、モデル化誤差に異なる分布を考えることにより、外れ値に頑健な推定が行える。
- 例）モデル化誤差がラプラス分布に従うと仮定すると

$$p(y_n; a, b, x_n) = \frac{1}{\sqrt{2}\sigma} e^{-\frac{|ax_n+b-y_n|}{\sigma/\sqrt{2}}}$$

$$\Rightarrow \quad f(a,b) = \sum_{k=1}^{K} |ax_k + b - y_k|$$

最尤推定は絶対値誤差最小化に等価

まとめ

- 最尤法
  - 尤度関数を最大化するという基準でパラメータを推定する枠組み
  - 最小二乗法は、誤差に正規分布を仮定した最尤推定と解釈できる