ブートストラップ法

樋口さぶろお https://hig3.net

龍谷大学大学院 理工学研究科 数理情報学専攻

理論物理学特論　L13 (2022-12-21 Wed)

最終更新: Time-stamp: "2022-12-21 Wed 12:17 JST hig"

今日の目標

- パラメトリックな方法の限界を説明できる
- ブートストラップ法のアイデアを説明できる

L12-Q1

Quiz 解答: 正規分布のベイズ推定

1. $Y - x_0 = \epsilon \sim \mathrm{N}(0, \sigma^2)$ より,

$$f_{\mathrm{Y|X}}(y|x_0) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(y-x_0)^2}{2\sigma^2}}$$

2. ベイズの定理より
3. 事後分布は,

$$f_{\mathrm{X|Y}}(x|y_0) = \frac{f(x, y_0)}{f_{\mathrm{Y}}(y_0)} = \frac{f(y_0|x) f_{\mathrm{X}}(x)}{f_{\mathrm{Y}}(y_0)} = \frac{1}{Z(y_0)} \mathrm{e}^{-\frac{(y_0-x)^2}{2\sigma^2}} \times \mathrm{e}^{-\frac{(x-m)^2}{2C}}.$$

ここで, $X_0(y)$ は y に依存する規格化定数.
$(a^2)^{-1} = (\sigma^2)^{-1} + (C)^{-1}$ とおくと,

$$f_{\mathrm{X|Y}}(x|y_0) = Z_1(y_0)^{-1} \mathrm{e}^{-\frac{1}{2}\frac{1}{a^2}(x-(\frac{a^2}{\sigma^2}\cdot y_0 + \frac{a^2}{C}\cdot m))^2}.$$

よって,

$$X|y_0 \sim \mathrm{N}(\frac{a^2}{\sigma^2} \cdot y_0 + \frac{a^2}{C} \cdot m, a^2).$$

解釈
もともと $x = m$ と思っていたが, 測定結果 y_0 に影響されて修正された (ベイズ更新). 修正の程度は, 正確さ $\frac{a^2}{\sigma^2}$ と $\frac{a^2}{C}$ の勝負で.
総合的な正確さは C^{-1} から, $(a^2)^{-1} = (\sigma^2)^{-1} + (C)^{-1}$ に改善された.

ここまで来たよ

パラメトリックな方法の限界 I

これまでは推測統計できれいにできるところだけを見てきた.
確率統計では,「X が正規分布にしたがうとき」「母平均値, 母分散の…」のような設定で, 区間推定や検定ができることを知った. 分布の種類が既知（特に正規分布関係) で, そのパラメタについてだけ議論しようというのりをパラメトリック parametric という. 美しい世界だが制限が多い.

うまくいってるところ

- 母分布が正規でないときも, 標本期待値 $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$ は母期待値 $\theta = \mathrm{E}[g(X)]$ の不偏推定量になってる
  バイアス=差の母期待値 $b = \mathrm{E}[\hat{\theta}] - \theta$
  推定量が不偏 unbiased $\Leftrightarrow$ バイアス $b = 0$

パラメトリックな方法の限界 II

- 母分布が正規でないときも, 不偏標本分散
  $\hat{\theta} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ は母分散 $\theta = \mathrm{V}[X]$ の不偏推定量になってる
  $n-1$ はバイアスを 0 にするため.

うまくいってないところ

正規を仮定しなくてできるのは教科書のここまで.

- 母平均値や母分散の区間推定は, 母分布の分布の正規性を仮定してしまっている (区間推定って, 分布が非対称でも対称なの?)
  - 実用的には, 正規でない分布のこともあるし, もとの分布がわからないこともある.
- へんな統計量も考えたいこともある (例: 母期待値として書けない量, 中央値, q-分位値, 四分位範囲, 相関係数, ...) 不偏推定量は簡単にはみつからない.
  - 例: $\sqrt{S^2}$ は母標準偏差の不偏推定量ではない.
  - 例: 確率密度関数が偶のとき, $\theta = \mathrm{E}[X]^2 = 0$ の推定量として $\hat{\theta} = \overline{X}^2$ を考えると, $\mathrm{E}[\hat{\theta}] = \mathrm{E}[X^2]/n > 0$ となり, 不偏でない.

L13-Q1

Quiz(不偏推定量)

母分布から標本 $X_1, \cdots, X_n$ を抽出した.

1. 標本平均値 $\overline{X} = \frac{1}{n}(X_1 + X_2 + X_3 + X_4 + \cdots + X_n)$ は母平均値 $\mu = \mathrm{E}[X_i]$ の不偏推定量か. バイアスを求めよう.
2. ひいきありの標本平均値 $\overline{X}' = \frac{1}{n}(2X_1 + 0 \cdot X_2 + X_3 + X_4 + \cdots + X_n)$ は母平均値の不偏推定量か. バイアスを求めよう.
3. 標本分散 $S^2 = \frac{1}{n}[(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})]$ は母分散 $\sigma^2 = \mathrm{V}[X_i]$ の不偏推定量か. バイアスを求めよう.
4. ひいきありの不偏標本分散 $S'^2 = \frac{1}{n}[(X_1 - \overline{X}')^2 + \cdots + (X_n - \overline{X}')]$ は母分散の不偏推定量か. バイアスを求めよう.

ここまで来たよ

経験累積分布関数 (empirical cumulative distribution function)

統計量 θ が, 確率密度関数 $f(x)$ を使って書かれてるなら, とりあえず標本から $\hat{f}$ を作れば, θ の標本での対応物が計算できる.
確率密度関数より相手にしやすい累積分布関数で考える.

$$F(x) = \int_{-\infty}^{x} f(x')\,\mathrm{d}x' = P(X \leq x)$$

定義 (経験分布)

サイズ n の標本 $\{x_1, x_2, \ldots, x_n\}$ に対して, 次を経験累積分布関数という.
これの定める分布を経験分布という.

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{[x \geq x_i]}(x)$$

$\hat{F} \to F$ as $n \to +\infty$.
$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)$ といいたくなるが, これは $n \to +\infty$ で $f(x)$ に収束しない

経験分布の性質

定義 (差込推定量)

「経験分布の母ナントカ」を母分布の差込推定量, プラグイン推定量という.

$$\mathrm{E}_{\hat{F}}[g(X)] = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

F の母平均値の差込推定量は, $\{x_1, \ldots, x_n\}$ の標本平均値
母中央値 m を $m = \frac{1}{2}(\inf\{x|F(x) \geq \frac{1}{2}\} + \sup\{x|F(x) \leq \frac{1}{2}\})$ で定義すると, F の母中央値の差込推定量は $\{x_1, \ldots, x_n\}$ の標本中央値
とりあえず推定量としてあいまいさなく定まる.

Quiz(経験分布)

未知の母分布から, X のサイズ $n = 4$ の標本 $9, 9, 10, 12$ を得た.

1. 経験累積分布関数のグラフを描こう.
2. '経験確率密度関数' のグラフを雰囲気で描こう.
3. 経験分布の母平均値を求めよう.
4. 経験分布の母中央値を求めよう.

例: へんな分布とへんな量 I

例として, 次の確率密度関数（偶関数でない!) を持つ連続型確率変数 X を考える (この分布の特殊性は使わない).

$$f(x) = \begin{cases} \frac{1}{4} & (-2 \leq x < 0) \\ \frac{1}{2} & (0 \leq x < 1) \\ 0 & (\text{他}) \end{cases} .$$

推定したい, 母分布の量 θ として, 母平均値, 母分散, 母中央値を考える. これらに対応する推定量を $\hat{\theta}$ と書く.

$$\begin{aligned} \mathrm{E}[X] &= -\tfrac{1}{4}, \\ \mathrm{V}[X] &= \tfrac{5}{6} - (-\tfrac{1}{4})^2 = \tfrac{37}{48}, \\ \mathrm{Median}[X] &= 0. \end{aligned}$$

うまくいってるところ I

$\hat{\theta}$ が θ の不偏推定量であるとは, $\mathrm{E}[\hat{\theta}] = \theta$ であること.

- X の標本平均値 $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ は, 母平均値 $\mathrm{E}[X]$ の不偏推定量. すなわち, $\mathrm{E}[\overline{X}] = \mathrm{E}[X]$.
- X の不偏標本分散 $S^2 = \frac{1}{n-1}[\cdots]$ は, 母分散 $\mathrm{V}[X]$ の不偏推定量. すなわち, $\mathrm{E}[S^2] = \mathrm{V}[X]$.

うまくいってないところ

- $\overline{X}$ ってどのくらいの精度? → 区間推定があればいい → 区間推定は母分布が正規分布のときしかできない! いま左右対称になる?
- S^2 ってどのくらいの精度? → 区間推定があればいい → 区間推定は母分布が正規分布のときしかできない!
- なにが 母 Median$[X]$ の不偏推定量? 標本 Median でいいの? 左にずれそうな気がする
  - バイアス bias $b = \mathrm{E}[$標本 Median$] - $ Median$[X]$.
  - 不偏 $\Leftrightarrow b = 0$.
- 標準誤差
  - $\mathrm{SE}^2 = \mathrm{E}[($標本 Median $- \mathrm{E}[$標本 Median$])^2]$.
  - 標本平均値, 標本期待値の標準誤差なら, 不偏標本分散で書けるけど, それ以外はわからない.
- 信頼区間 正規分布じゃないから理論がない. 母平均値以外には理論がない.
  - 母 Median の信頼区間はたぶん左右非対称になる

経験分布の母期待値は Monte Carlo 近似できる

X_i が独立に経験累積分布 $\hat{F}$ にしたがうとき,

$$\mathrm{E}_{\hat{F}}[g(X_1,\ldots,X_k)] = \frac{1}{n^k}\sum_{i_i=1}^{n}\cdots\sum_{i_k=1}^{n} g(x_{i_1},\ldots,x_{i_k})$$

これを, もとの分布 F に対する $\mathrm{E}_F[g(X_1,\ldots,X_k)]$ の差込推定量, プラグイン推定量という.

命題 (Monte Carlo 近似)

母期待値は, 標本のサイズが大きいとき, 大数の法則により次で近似できる. $N \to +\infty$ で, 真の値に収束する.

$$\mathrm{E}_{\hat{F}}[g(X_1,\ldots,X_k)] \approx \frac{1}{N}\sum_{j=1}^{N} g(x_1(j),\ldots,x_k(j))$$

$(x_1(j),\ldots,x_k(j))$ は経験分布から抽出した (n^k 項から等確率で抽出した) j 個目の組.

ここまで来たよ

ブートストラップ法の主要アイデア I

母分布の情報を使いたい. 母分布は経験分布と似ているだろう, 推定量の分布は, 標本からさらに繰り返し標本抽出して Monte Carlo 近似で求めちゃえ, というのがブートストラップ法のアイデア.

ブートストラップ標本 I

ブートストラップ標本とは, 経験分布から抽出したサイズ $k = n$ の標本 (n は経験分布の元になった標本のサイズ).
別の言い方: 標本を母集団と思い直して復元抽出して孫標本.
ふつう, 多数個 (B 個) のブートストラップ標本を考える.

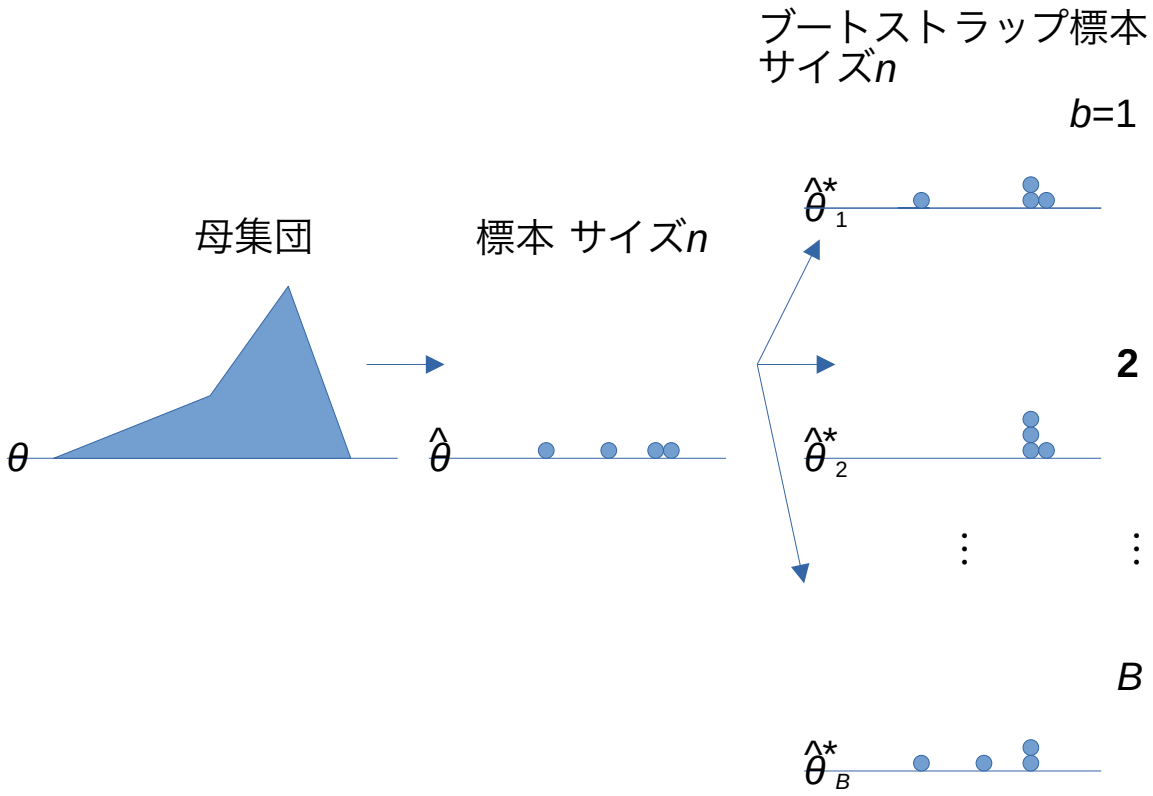
(母集団) : (標本) ≈(標本) : (ブートストラップ標本)

親 : 子 ≈子 : 孫

親 $\overset{\text{標本抽出}}{\rightarrow}$ 子

真のパラメタθ – 標本からの推定値$\hat{\theta}$

≈ 標本からの推定値$\hat{\theta}$ – ブートストラップ推定値$\hat{\theta}^*$


ブートストラップ標本
サイズn
b=1
母集団
標本 サイズn
θ
$\hat{\theta}$
$\hat{\theta}^*_1$
2
$\hat{\theta}^*_2$
B
$\hat{\theta}^*_B$

ブートストラップ推定値

ブートストラップ標本を抽出する試行を考えたとき, その母期待値 $\mathrm{E}[\hat{\theta}^*]$ をブートストラップ推定値という.

Monte Carlo 近似

ブートストラップ推定値は, Monte Carlo 近似できる. B 個のブートストラップ標本 b の統計量 $\hat{\theta}_b^*$ を求めて, 標本期待値をとる.

$$\hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$$

.

B: ブートストラップ反復回数

ここまで来たよ

標本誤差

どんな統計量かを気にせず, ブートストラップ標本から, ブートストラップ推定値の Monte Carlo 近似として一律に計算できちゃう.

$$(\text{標準誤差})^2 = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}^*)^2.$$

バイアス, バイアス補正値

バイアス $b = \mathrm{E}[$標本推定値$\hat{\theta}] -$ 真の値θ.

$$
\begin{aligned}
\text{真の値}\theta =& \mathrm{E}[\text{標本推定値}\hat{\theta}] - \text{バイアス}b \\
\approx& \text{標本推定値}\hat{\theta} - (\text{ブートストラップ推定値}\hat{\theta}^* - \text{標本推定値}\hat{\theta}) \\
=& 2 \times \hat{\theta} - \hat{\theta}^* \\
=& \text{バイアス補正 (された) 値}
\end{aligned}
$$

ブートストラップ推定値は, ブートストラップ反復回数 $B = 50 - -300$ の Monte Carlo 近似で求める.
ブートストラップの近似値が標本推定値より左に出たら, 補正値は標本推定値より右に置け. 標本推定値がもともと左にでちゃってただろうから, その分戻す, という考え方.

L13-Q2

Quiz(ブートストラップ標本抽出)

ある未知の母集団から, サイズ $n = 6$ の標本抽出したところ, 次のようになった. 中央値 θ を考える.

$$8, 8, 8, 10, 13, 13.$$

1. 標本中央値 $\hat{\theta}$ を求めよう.
2. サイコロを使って, $B = 3$ のブートストラップ標本を生成しよう.
3. ブートストラップ標本の標本中央値 $\hat{\theta}^*_1, \hat{\theta}^*_2, \hat{\theta}^*_3$ を求めよう.
4. ブートストラップ標本の標本中央値 $\hat{\theta}^*_b$ のブートストラップ標本平均値 $\hat{\theta}^*$ を求めよう.
5. バイアス補正を行った母中央値の推定値を求めよう.
6. 母中央値の推定値の標準誤差を推定しよう

L13-Q3

Quiz(ブートストラップ法)

(理解チェックのための不自然なブートストラップ標本です)
標本で, 推定量 $\hat{\theta} = 10$ となった. $B = 20$ ブートストラップ標本で, 推定量 $\hat{\theta}^*$ は, 大きさの順に,
$8, 8, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 12, 12, 12, 12, 12, 12$
となった.

1. バイアスを推定しよう.
2. バイアス修正を行った推定値を求めよう.
3. 標準誤差を推定しよう.
4. $\alpha = 0.1$ のブートストラップ基本区間推定を行おう.

ブートストラップ信頼区間 I

いい加減に考えると, B 個のブートストラップ推定量 $\hat{\theta}_b$ のヒストグラムの両端 alpha/2 を削りたくなる (パーセンタイル信頼区間という名前もついている) が, 分布が左右非対称なときは根拠がない.
区間内が $1-\alpha$ になるように調節された, 両端の位置 $t(\alpha/2), t(1-\alpha/2)$ (α が小さければふつうは負, 正).

$$1-\alpha = P(t(\alpha/2) \leq \hat{\theta} - \theta \leq t(1-\alpha/2)).$$

不等式を母分布の量 θ について解いて,

$$1-\alpha = P(\hat{\theta} - t(1-\alpha/2) \leq \theta \leq \hat{\theta} - t(\alpha/2)).$$

$\hat{\theta}$ をブートストラップ推定値, θ を標本の推定値で置き換える. $t(\alpha/2)$ を, ブートストラップ推定値のうち下位 $\alpha/2$ がはいる位置 $t(\alpha/2)^*$ で置き換える.

ブートストラップ信頼区間 II

$$1-\alpha = P(\hat{\theta}^* - t(1-\alpha/2)^* \leq \theta \leq \hat{\theta}^* - t(\alpha/2)^*).$$

これをブートストラップ基本信頼区間という.
ブートストラップ標本で左にでがちだったら, もともとの標本推定値が左にでてるだろうから, そのずれくらい, 右に信頼区間を広げておかなければいけない.

ブートストラップ

Python

- https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/modules/generated/sklearn.cross_validation.Bootstrap.html
- ブートストラップ信頼区間の計算https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html

R

- boot
- bootstrap

ブートストラップと関係する手法

- jack knife 法
- cross validation
- bagging, b for bootstrap
- random forest