ブートストラップ法

TANIGUCHI Taichi

October 5, 2018

Contents

Part I
ブートストラップ法 (bootstrap method)

ブーストラップ法とはデータ解析の確からしさを評価する手法です。

例えばデータが X： 1000×100 行列で与えられているとします。そして、データを入力するとある実数の値を出力する装置を考えます。この実数の値は何らかの事象についての予測値とします。出力値を $\theta^\star$ とします。

有限のデータから予測された値は期待値からずれていることがほとんどです。ではこのデータから予測値のばらつきを調べたいとします。次のアルゴリズムを適応することで予測値のばらつきを確かめます。

1　ノンパラメトリック・ブートストラップ

アルゴリズムを以下で与えます。

1. 正の整数 n を決める。また、十分に大きな正の整数 B を決める。
2. 1,2,3,„,n の中から等しい確率で数字を選ぶこれを n 回繰り返して得られた整数列を $i_1, i_2, i_3, , , , i_n$ をする。得られた数字を index に持つ要素を X の中から取り出して $x_{i_1}, x_{i_2}, , , , x_{i_n}$ とする。このデータを新たに $X^\star$ とする。
3. 　　2 と 3 を B 回繰り返したのち、$X_1, X_2, , , , X_B$ のデータ集合を得る。
4. 　各 X_b　$\forall b \in B$　で $\theta^{\star^b}$ を計算する。

このアルゴリズムを適応することで B 個の $\theta^\star$ を得られます。この B 個の $\theta^\star$ から分散 $\sigma^\star$ を計算します。ただし、普遍推定量で求めることに注意します。

$$(\sigma^\star)^2 = \frac{1}{B-1}\sum_{b=1}^{B}(\theta^{\star^b} - \frac{1}{B}\sum_{b^\star}\theta^{b^\star})$$

この値と元のデータから予測値 $\theta\star$ がどれだけ信用できるかがわかります。B の値が大きければ大きいほど正確性は増しますが、一般に B は 10^4 で十分と言われています。

ノンパラメトリックというのは、データの生成のメカニズムを仮定していないことを意味します。

2　信頼区間

2.1　標準信頼区間

パラメータ θ の信頼区間が知りたいときがあります。つまり、どのような θ なら十分信頼できるのかということを知りたいわけです。信頼しない確率を α とすると、信頼区間を求めるアルゴリズムは以下のようになります。

- ノンパラメトリックブートストラップ法によって、予測値 θ の, サンプルによるばらつきを示す、標準偏差 σ を求めます。
- 標準正規分布の両側に対して、$\frac{\alpha}{2}$ 以上（以下）、となる確率を求めます。つまり、$P(X \leq z^{(1-\frac{\alpha}{2})}) = 1 - \frac{\alpha}{2}$ となる $z^{(1-\frac{\alpha}{2})}$ を求めておきます。
- θ の両側信頼区間を
$$[\theta^\star - z^{(1-\frac{\alpha}{2})}\sigma, \theta^\star + z^{(1-\frac{\alpha}{2})}\sigma$$
とします。

2.2 パーセンタイル信頼区間

標準信頼区間では、$\theta^\star$ が正規分布に従うと仮定している。つまり、ブートストップによって求めた $\theta^\star$ の標準偏差を σ とすると、ブートストラップによって求めた M 個の予測値 $\theta^\star_m$ は θ の真の値、θ^{true} を平均として、分散を σ^2 にもつ正規分布に従うと仮定している。

しかし、実際ブートストラップ法を使うと、分布が左右非対称になったり、平均が大きすぎたりしてしまう。よってブートストラップ法によって直接信頼区間を得る方法がある。それがパーセンタイル信頼区間である。パーセンタイル信頼区間は以下のアルゴリズムで求められる。

- ブートストラップ方によって求めた B 個の予測値、$\theta^\star_1, \theta^\star_2, ..., \theta^\star_B$ を小さい順に並び変える。$0 < p < 1$ に対して、pB に小さい数値を $\theta^\star_p$ と表す。pB が整数値でなければ、隣り合う値を線形補間する。
- 両側信頼区間を次のように与える。

$$[\theta^\star_{\frac{\alpha}{2}}, \theta^\star_{1-\frac{\alpha}{2}}]$$

これはつまり、予測値を小さい順に並べて、両側 $\frac{\alpha}{2}$、$1-\frac{\alpha}{2}$ の中側を信頼区間とするということである。

このように計算した信頼区間に θ が入る確率を被覆確率という。多くの場合、P パーセント信頼区間の被覆確率は P パーセントと考える。この被覆確率と信頼水準（1-有意水準）との誤差をバイアスといい、これが小さいほど精度がよいということになる。

一般にブートストラップ法で得られた θ の分布は非対称なので、正規分布を仮定した標準信頼区間のほうがバイアスは小さくなる。非対称性が誤差につながるのはのちに解説する。

3 プラグイン推定量

ここではブートストラップ法によって推定された値が、その真値の近似になっているのかを検証する。データの要素 $x_1, x_2, ..x_n$ は確率密度関数 $f(x)$ に対して、

$$x_1, x_2, ...x_n \quad \sim \quad f(x)$$

となることを仮定する。

次に確率密度関数 $f^\star(x)$ を次のように定義する。

$$f^\star(x) = \frac{1}{n}\sum_{i=1}^{n} \delta\ (x - x_i)$$

ただし、$\delta\ (x)$ はディラックのデルタ関数といいます。

ブートストラップでは確率 $\frac{1}{n}$ で x_i を選択してくるので、ブートストラップによって得られるサンプルデータは $f^{\star}(x)$ に従います。

つまり、ブートストラップによってえらばれた回数が多いサンプルの確率が高いという理にかなった解釈ができます。この $f^{\star}$ を経験分布 (empirical distribution) といいます。

よってブートストラップ標本 $X^{*} = x_1^{*}, ..., x_{n'}^{*}$ は

$$x_1^{*}, ..., x_{n'}^{*} \sim f^{\star}(x)$$

と表すことができる。

$n \to \infty$ の時、$f^{\star}$ が f に収束すれば、ブートストラップによって生成されたサンプルデータから予測された推定値は意味を持つことがわかる。

4 マルチスケール・ブートストラップ法 (multiscale bootstrap)

References

[1] $http://ebsa.ism.ac.jp/ebooks/sites/default/files/ebook/1881/pdf/vol3_ch8.pdf$

[2] $https://www1.doshisha.ac.jp/\ mjin/R/Chap_44/44.html$

[3] https://ja.wikipedia.org/wiki/