Phystech@DataScience

Бутстреп

24 марта 2024 г.

Постановка задачи

$X = (X_1, ..., X_n)$ — выборка

$T(X_1, ..., X_n)$ — статистика

Задача: оценить распределение $T(X)$ или функционал $V(T(X))$.

Пример: оценка дисперсии статистики

$$V(T(X_1, ..., X_n)) = \mathsf{D}T(X_1, ..., X_n) = \mathsf{E}T^2(X) - (\mathsf{E}T(X))^2$$

Какую сделать оценку?

Бутстреп

Метод бутстрепа

Этап 1.

Генерация индексов из равномерного распределения:

$$i_1, \dots, i_n \sim U\{1, \dots, n\}$$

Генерация выборки $X^* = (X_1^*, \dots, X_n^*) = (X_{i_1}, \dots, X_{i_n})$: упорядоченный выбор **с возвращением** n элементов из мн-ва $\{X_1, \dots, X_n\}$.

Например:

1. $X = (100, 11, -5, 91, 32)$ — реализация выборки
2. $(4, 5, 5, 1, 2) = (i_1, \dots, i_5) \sim U\{1, \dots, 5\}$.
3. $X^* = (X_{i_1}, \dots, X_{i_n}) = (91, 32, 32, 100, 11)$ — бутстрепная выборка.

Важно: размер выборки равен исходному

Метод бутстрепа

Этап 2.

Процедуру генерации выборок повторить B раз:

$$X_b^* = (X_{b1}^*, ..., X_{bn}^*), \text{ где } 1 \leqslant b \leqslant B.$$

Далее по каждой выборке посчитаем значение статистики T, получив выборку значений: $T_1^* = T(X_1^*), ..., T_B^* = T(X_B^*)$.

Этап 3.

Полученную выборку использовать для аппроксимации значения оценки, которая называется *бутстрепной оценкой*.

Например, бутстрепная оценка дисперсии T имеет вид

$$\widehat{v}_{boot} = \frac{1}{B}\sum_{b=1}^{B} T_b^{*2} - \left(\frac{1}{B}\sum_{b=1}^{B} T_b^*\right)^2$$

Схема метода бутстрепа

$X = (X_1, \dots, X_n)$ — выборка

$T(X_1, \dots, X_n)$ — статистика

Задача: оценить распределение $T(X)$ или функционал $V(T(X))$.

$$\left.\begin{array}{lll}
X^*_{11},\ \dots,\ X^*_{1n} & \longrightarrow & T(X^*_1)\ \dots \\
X^*_{b1},\ \dots,\ X^*_{bn} & \longrightarrow & T(X^*_b) \\
 & & \dots \\
X^*_{B1},\ \dots,\ X^*_{Bn} & \longrightarrow & T(X^*_B)
\end{array}\right\}$$

v_{boot} — бутстрепная оценка $v = V(T(X))$

Зоопарк: оценить дисперсию выборочного среднего

Выборка:

Задача: Для каждого пикселя и каждого цветового канала оценить дисперсию выборочного среднего.

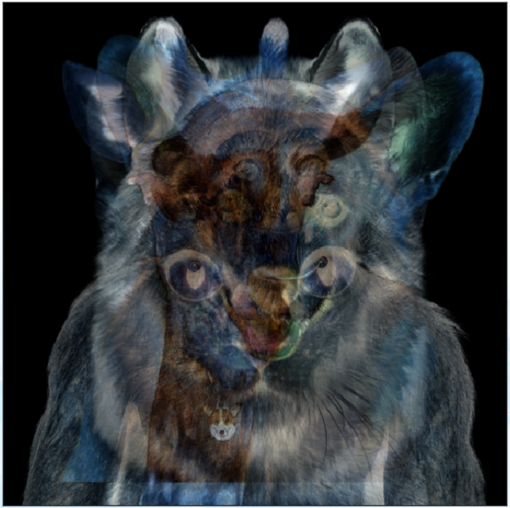
Зоопарк: оценить дисперсию выборочного среднего

Зоопарк: оценить дисперсию выборочного среднего

Дисперсия по бутстрепной выборке средних:

Зоопарк: оценить дисперсию выборочного среднего

При большом количестве бутстрепных выборок:



Особенности

- Число B стоит брать как можно больше.
- Размер бутстрепной выборки **всегда тот же**, что и у исходной.

  При генерации выборок иного размера распределение статистики T, вообще говоря, может быть другим.

  Например, дисперсия выборочного среднего зависит от размера выборки.
- Генерация бутстрепной выборки проводится независимо с повторами. Иначе полученный набор даже не является выборкой.

Бутстрепные доверительные интервалы

1. **Нормальный интервал**

   Пусть $\widehat{\theta}$ — а.н.о. θ с ас. дисп. $\sigma^2(\theta)$.

   $\widehat{v}_{boot}$ — бутстрепная оценка дисперсии.

   Бутстрепный дов. интервал для параметра θ имеет вид

   $$\left(\widehat{\theta} - z_{(1+\alpha)/2}\sqrt{\widehat{v}_{boot}}, \quad \widehat{\theta} + z_{(1+\alpha)/2}\sqrt{\widehat{v}_{boot}}\right)$$

2. **Центральный интервал**

   $\theta = G(\mathsf{P})$ и $\widehat{\theta} = G(\widehat{\mathsf{P}}_n)$ — оценка методом подстановки.

   $\theta_1^*, ..., \theta_B^*$ — оценки по бутстрепным выборкам.

   Бутстрепный доверительный интервал имеет вид

   $$C^* = \left(2\widehat{\theta} - \theta^*_{(\lceil B(1+\alpha)/2 \rceil)}, \quad 2\widehat{\theta} - \theta^*_{(\lfloor B(1-\alpha)/2 \rfloor)}\right).$$

Бутстрепные доверительные интервалы

3. **Квантильный интервал**

   $\widehat{\theta}$ — некоторая оценка θ.

   $\theta_1^*, ..., \theta_B^*$ — оценки по бутстрепным выборкам.

   Бутстрепный доверительный интервал имеет вид

   $$C^* = \left(\theta^*_{(\lfloor B(1-\alpha)/2 \rfloor)}, \quad \theta^*_{(\lceil B(1+\alpha)/2 \rceil)}\right).$$

   Утверждение. Если существует монотонное преобразование φ, для которого $\varphi(\widehat{\theta}) \sim \mathcal{N}(\varphi(\theta), \sigma^2)$, то $\mathsf{P}\left(\theta \in C^*\right) = \alpha$.

   На практике такое преобразование существует редко, но при этом часто может существовать приближенное преобразование.

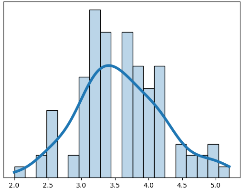
Пример: построение дов. интервалов для θ

$x = (5, 1, 3, 6, 4)$ — реализация выборки

$\theta = \mathsf{E}X_1$ — параметр, $\widehat{\theta} = \overline{X}$ — оценка, $\widehat{\theta} = 3.8$ — реализация оценки

Реализации оценки параметра по бутстрепным выборкам ($B = 100$):

4.2, 4.2, 2.6, 3.2, 4.2, 3.8, 3.2, 3.6, 3.6, 3.4,
3.8, 4.4, 3.6, 3.2, 4.6, 4.2, 3.0, 3.2, 4.0, 3.0,
4.0, 2.4, 3.4, 3.8, 2.0, 3.0, 4.6, 3.2, 3.6, 3.6, ...



1. **Нормальный интервал**

   $\widehat{\theta} = 3.8,\ v_{boot} = 0.394,\ z_{0.975} = 1.96$

   $(3.8 \pm 1.96 \cdot \sqrt{0.394}) = (2.57,\ 5.03)$

2. **Центральный интервал**

   $B(1+\alpha)/2 = 100 \cdot 0.975 = 97.5,\ B(1-\alpha)/2 = 100 \cdot 0.025 = 2.5$

   $\theta^*_{(\lceil 97.5 \rceil)} = 5, \quad \theta^*_{(\lfloor 2.5 \rfloor)} = 2.4$

   $(2 \cdot 3.8 - 5,\ 2 \cdot 3.8 - 2.4) = (2.6,\ 5.2)$

3. **Квантильный интервал**

   $(2.4,\ 5)$

Оценка доли покрытия интервалом

Задача:

Оценить $\mathsf{P}_\theta\Big(\theta \in \big(T_1(X), T_2(X)\big)\Big) = \mathsf{E}\mathsf{I}\{\theta \in \big(T_1(X), T_2(X)\big)\}$

Решение:

1. Генерируем B выборок из распределения P_θ размера n
2. Считаем $T_1(X_b)$, $T_2(X_b)$ и $\mathsf{I}\{\theta \in \big(T_1(X_b), T_2(X_b)\big)\} = I_b \in \{0, 1\}$
3. $\widehat{\mathsf{P}}(\cdot) = \overline{I_b} = \frac{1}{B}\sum\limits_{b=1}^{B} I_b$


ВСЁ!