

Математические основы искусственного интеллекта.

Описательные статистики. Метод моментов. Доверительные интервалы

Солодушкин Святослав Игоревич

Кафедра вычислительной математики и компьютерных наук,
УрФУ имени первого Президента России Б.Н. Ельцина

Март 2022

Важной является задача оценивания параметров генеральной совокупности $X(\theta)$.

Что дано?

- 1 Априорные сведения о виде распределения генеральной совокупности.
- 2 Выборка из генеральной совокупности x_1, x_2, \dots, x_n .

Что надо найти?

Оценку $\hat{\theta}(x_1, x_2, \dots, x_n)$ параметра θ .

Пусть X_1, X_2, \dots, X_n — случайная выборка из генеральной совокупности X . Распределение X известно с точностью до числового параметра θ .

Статистикой называется произвольная измеримая функция выборки $T: X^n \rightarrow \mathbb{R}$, которая не зависит от неизвестных параметров распределения.

Условие измеримости статистики означает, что эта функция является случайной величиной, то есть определены вероятности ее попадания в интервалы.

От неизвестных параметров статистика не зависит, т. е. можно по имеющимся данным найти значение этой функции, а следовательно, основывать на этом значении оценки и прочие статистические выводы.

Определение

Пусть X_1, X_2, \dots, X_n — случайная выборка для распределения, зависящего от параметра $\theta \in \Theta$. Тогда статистику $\hat{\theta}(X_1, \dots, X_n)$, принимающую значения в Θ , называют точечной оценкой параметра θ .

Формально статистика $\hat{\theta}$ может не иметь ничего общего с интересующим нас значением параметра θ . Ее полезность для получения практически приемлемых оценок вытекает из дополнительных свойств, которыми она обладает (или не обладает).

Пример. Случайная величина

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i,$$

назваемая выборочным средним, является точечной оценкой среднего в генеральной совокупности.

Свойства точечных оценок: несмещенность

Оценка $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ называется несмещенной, если ее математическое ожидание равно оцениваемому параметру генеральной совокупности:

$$\mathbb{E}_{\theta} [\hat{\theta}] = \theta, \quad \forall \theta \in \Theta,$$

где \mathbb{E}_{θ} обозначает математическое ожидание,
 θ — истинное значение параметра.

Пример. Выборочное среднее \bar{X} является несмещенной оценкой среднего m в генеральной совокупности

$$\mathbb{E}[\bar{X}] = \mathbb{E} \left[\frac{\sum_{i=1}^n X_i}{n} \right] = \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n X_i \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E} X_i = \frac{nm}{n} = m.$$

Пример смещенной оценки

Выборочная дисперсия — это случайная величина вида

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2,$$

Выборочная дисперсия является смещенной оценкой дисперсии D :

$$\mathbb{E} [S_n^2] = \mathbb{E} \left[\frac{\sum_{i=1}^n \left((X_i - m) - (\bar{X} - m) \right)^2}{n} \right] = \frac{n}{n-1} D.$$

Оценка $\hat{\theta}$ называется эффективной, если она обладает минимальной дисперсией среди всех возможных несмещенных точечных оценок.

То, что оценка обладает минимальной дисперсией, не означает, что эта дисперсия мала; и тем более не означает, что дисперсия уменьшается с увеличением объема выборки.

Оценка $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ называется состоятельной, если $\forall \theta \in \Theta$ она с увеличением объема выборки n стремится по вероятности к параметру генеральной совокупности

$$\hat{\theta}_n \rightarrow \theta \text{ по вероятности при } n \rightarrow \infty.$$

Как правило несостоятельные оценки не востребованы на практике.

Пусть X — случайная величина.

k -м начальным моментом случайной величины X , где $k \in \mathbb{N}$, называется величина

$$\nu_k = \mathbb{E} \left[X^k \right],$$

k -м центральным моментом случайной величины X называется величина

$$\mu_k = \mathbb{E} \left[(X - \mathbb{E}X)^k \right],$$

если математические ожидания $\mathbb{E}[\cdot]$ в правых частях этих равенств определены.

Метод моментов — метод оценки неизвестных параметров распределений в математической статистике. Идея метода заключается в замене истинных соотношений выборочными аналогами.

Пусть задан вид плотности распределения $f(x, \theta)$, определяемый одним неизвестным параметром θ . Требуется найти точечную оценку параметра θ .

Для оценки одного параметра достаточно иметь одно уравнение относительно этого параметра. Следуя методу моментов, приравняем начальный теоретический момент первого порядка начальному эмпирическому моменту первого порядка.

Метод моментов: оценка одного параметра

Учитывая, что $\nu_1 = \mathbb{E}(X)$, а эмпирический начальный момент первого порядка равен \bar{x} , получим

$$\mathbb{E}(X) = \bar{x} \quad (*)$$

Математическое ожидание $\mathbb{E}(X)$

$$M(X) = \int_{-\infty}^{\infty} xf(x, \theta) dx = \phi(\theta)$$

есть функция от θ , поэтому (*) можно рассматривать как уравнение с одним неизвестным θ .

Решив это уравнение относительно параметра θ , найдем его точечную оценку $\hat{\theta}$, которая является функцией от выборочной средней.

Кстати, в данном случае выборочная средняя является *достаточной статистикой*.

Найти методом моментов по выборке x_1, x_2, \dots, x_n точечную оценку неизвестного параметра λ показательного распределения, плотность распределения которого $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$.

$$\mathbb{E}[X] = \frac{1}{\lambda} = \bar{x}.$$

Пусть задан вид плотности распределения $f(x, \theta_1, \theta_2)$, определяемый двумя неизвестными параметрами θ_1, θ_2 . Требуется найти точечные оценки параметров θ_1, θ_2 .

Для оценки двух параметров достаточно иметь два уравнения относительно этих параметров. Следуя методу моментов, приравняем

- 1 начальный теоретический момент первого порядка начальному эмпирическому моменту первого порядка,
- 2 центральный теоретический момент первого порядка центральному эмпирическому моменту первого порядка.

Метод моментов: оценка двух параметров

Найти методом моментов по выборке x_1, x_2, \dots, x_n точечную оценку неизвестного параметров a и σ нормального распределения, плотность распределения которого

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-a)^2/(2\sigma^2)}.$$

Учитывая, что $\nu_1 = \mathbb{E}(X)$, а эмпирический начальный момент первого порядка равен \bar{x} , получим

$$\mathbb{E}(X) = \bar{x} \quad (*)$$

Учитывая, что $\mu_2 = D(X)$, а эмпирический центральный момент первого порядка равен \bar{x} , получим

$$\mathbb{E}(X) = \bar{x} \quad (*)$$

Математическое ожидание $\mathbb{E}(X)$

$$M(X) = \int_{-\infty}^{\infty} xf(x, \theta) dx = \phi(\theta)$$

Проблема состоит в нахождении на основании выборки статистик $\hat{\theta}_1 = \hat{\theta}_1(X_1, \dots, X_n)$ и $\hat{\theta}_2 = \hat{\theta}_2(X_1, \dots, X_n)$, которые с достоверностью удовлетворяют неравенству $\hat{\theta}_1 < \theta < \hat{\theta}_2$.

Зададимся достаточно малым числом α — уровнем значимости. Интервал $[\hat{\theta}_1; \hat{\theta}_2]$ называется интервальной оценкой параметра θ , если $P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1 - \alpha$.

Интервал $I(X) = [\hat{\theta}_1; \hat{\theta}_2]$ называется доверительным интервалом параметра на уровне значимости α или с надежностью $1 - \alpha$.

Доверительный интервал для математического ожидания нормальной выборки

Пусть X_1, \dots, X_n — случайная выборка из нормально распределенной генеральной совокупности $X \sim N(m, \sigma^2)$, где σ^2 — известная дисперсия.

Определим произвольное $\alpha \in (0, 1)$ и построим доверительный интервал для неизвестного среднего m .

Доверительный интервал для математического ожидания нормальной выборки

Случайная величина

$$Z = \frac{\bar{X} - m}{\sigma/\sqrt{n}}$$

имеет стандартное нормальное распределение $N(0, 1)$.

Пусть z_α — это α -квантиль стандартного нормального распределения, тогда в силу симметрии имеем:

$$P\left(-z_{1-\frac{\alpha}{2}} \leq Z \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

После подстановки выражения для Z получаем:

$$P\left(\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Доверительный интервал для математического ожидания нормальной выборки

Пусть X_1, \dots, X_n — случайная выборка из нормально распределенной генеральной совокупности $X \sim N(m, \sigma^2)$, где m, σ^2 — неизвестные константы.

Определим произвольное $\alpha \in (0, 1)$ и построим доверительный интервал для неизвестного среднего m .

Доверительный интервал для математического ожидания нормальной выборки

Случайная величина

$$T = \frac{\bar{X} - m}{S/\sqrt{n}}$$

имеет распределение Стьюдента с $n - 1$ степенями свободы $t(n - 1)$, где S — несмещенное выборочное стандартное отклонение.

Пусть $t_{\alpha, n-1}$ — это α -квантиль распределения Стьюдента, тогда в силу симметрии имеем:

$$P\left(-t_{1-\frac{\alpha}{2}, n-1} \leq T \leq t_{1-\frac{\alpha}{2}, n-1}\right) = 1 - \alpha$$

.

После подстановки выражения для T получаем:

$$P\left(\bar{X} - t_{1-\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} \leq m \leq \bar{X} + t_{1-\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

- 1 Сформулируйте