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

Оценки параметров распределения

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

Что дано?

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

Что надо найти?
Оценку $\hat{\theta}(x_1, x_2, \ldots, x_n)$ параметра θ.

Статистика — функция выборки

Пусть $X_1, X_2 \ldots, X_n$ — случайная выборка из генеральной совокупонсти X. Распределение X известно с точность до числового параметра θ.

Статистикой называется произвольная измеримая функция выборки $T\colon X^n \to \mathbb{R}$, 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.

Точечная оценка параметра

Определение

Пусть $X_1, X_2, \ldots, X_n$ — случайная выборка для распределения, зависящего от параметра $\theta \in \Theta$. Тогда статистику $\hat{\theta}(X_1, \ldots, X_n)$, 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

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

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

Свойства точечных оценок: несмещенность

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

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

где $\mathbb{E}_\theta$ 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

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

Выборочная дисперсия является смещенной оценкой дисперсии D :

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

Эффективность оценки

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 оозначает, что дисперсия уменьшается с увеличением объема выборки.

Состоятельность оценки

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

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

Как правило несостоятельные оценки не востребованы на практике.

Моменты случайной величины

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

Метод моментов: оценка одного параметра

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

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

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

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x, \theta)\, dx = \phi(\theta)$$

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*.

Метод моментов: оценка одного параметра

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

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

Метод моментов: оценка двух параметров

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 второго порядка центральному эмпирическому моменту второго порядка.

Метод моментов: оценка двух параметров

Найти методом моментов по выборке $x_1, x_2, \ldots, x_n$ точечную оценку неизвестных 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

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

Учитывая, что $\mu_2 = D(X)$, а эмпирический центральный момент второго порядка равен $D_{\text{выб}}$, получим

$$D(X) = D_{\text{выб}}(X)$$

Математическое ожидание $\mathbb{E}(X) = a$. Дисперсия $D(X) = \sigma^2$.

Интервальная оценка

Это почти очевидно, что (в случае непрерывного распределения) точечная оценка не будет совпадать с истинным неизвестным значением параметра.
Решение состоит в том, чтобы вместо точечной оценки построить интервал, который покрывает неизвестный параметр. Но... на самом деле, даже в этом случае мы не можем гарантировать, что интервал обязательно покроет неизвестный параметр.
Действительно, границы интервала, будучи функциями от случайной выборки, сами являются случайными величинами, а потому факт накрытия истинного параметра является случайным событием.
Итак, мы должны поставить задачу: построить интервал

$$[\hat{\theta}_1 = \hat{\theta}_1(X_1, \ldots, X_n);\ \hat{\theta}_2 = \hat{\theta}_2(X_1, \ldots, X_n)],$$

покрывающий неизвестный параметр с заданной вероятностью (надежностью) γ.

Интервальная оценка

Необходимо на основании выборки найти статистики $\hat{\theta}_1 = \hat{\theta}_1(X_1, \ldots, X_n)$ и $\hat{\theta}_2 = \hat{\theta}_2(X_1, \ldots, X_n)$, которые с достоверностью γ удовлетворяют неравенству

$$\mathrm{P}(\hat{\theta}_1 < \theta < \hat{\theta}_2) = \gamma.$$

Обозначим вероятность того, что интервал не покроет истинный параметр, как $\alpha = 1 - \gamma$.

$$\mathrm{P}(\theta \notin [\hat{\theta}_1; \hat{\theta}_2]) = 1 - \gamma = \alpha.$$

Интервал $[\hat{\theta}_1; \hat{\theta}_2]$ называется доверительным интервалом, покрывающим неизвестный параметр θ с заданной достоверностью γ.

Доверительный интервал для математического ожидания нормальной выборки

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

Определим произвольное $\gamma \in (0, 1)$ 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 $\mathrm{N}(0, 1)$.

Пусть $z_{\frac{1+\gamma}{2}}$ — это $\frac{1+\gamma}{2}$-квантиль стандартного нормального распределения, тогда в силу симметрии имеем:

$$\mathrm{P}\left(-z_{\frac{1+\gamma}{2}} \leq Z \leq z_{\frac{1+\gamma}{2}}\right) = \gamma.$$

После подстановки выражения для Z получаем:

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

Доверительный интервал для математического ожидания нормальной выборки

Пусть $X_1, \ldots, X_n$ — случайная выборка из нормально распределенной генералной совокупности $X \sim \mathrm{N}(m, \sigma^2)$, где m, σ^2 — неизвестные константы.

Определим произвольное $\gamma \in (0, 1)$ 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 $\mathrm{t}(n-1)$, где S — несмещенное выборочное стандартное отклонение.

Пусть $t_{\frac{1+\gamma}{2},n-1}$ — это $\frac{1+\gamma}{2}$-квантиль распределения Стьюдента, тогда в силу симметрии имеем:

$$\mathrm{P}\left(-t_{\frac{1+\gamma}{2},n-1} \leq T \leq t_{\frac{1+\gamma}{2},n-1}\right) = \gamma.$$

После подстановки выражения для T получаем:

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

Квантиль

Квантиль в математической статистике — значение, которое заданная случайная величина не превышает с фиксированной вероятностью.

Например, фраза «90-й процентиль массы тела у новорожденных мальчиков составляет 4 кг» означает, что 90% мальчиков рождаются с весом, меньшим либо равным 4 кг, а 10% мальчиков рождаются с весом, большим либо равным 4 кг.

Рассмотрим вероятностное пространство $(\Omega,\ \mathcal{F},\ \mathbb{P})$, на котором задана случайная величина X. Пусть фиксировано $\alpha \in (0,\ 1)$. Тогда α-квантилем (или квантилем уровня α) распределения $\mathbb{P}^X$ называется число $x_\alpha \in \mathbb{R}$, такое что

$$\mathbb{P}(X \leqslant x_\alpha) \leqslant \alpha,$$

$$\mathbb{P}(X \geqslant x_\alpha) \geqslant 1 - \alpha.$$

Задания

1. Из нормально распределенной генеральной совокупности с неизвестным мат. ожиданием и известной дисперсией извлекли выборку. По этой выборке построили 95% доверительный интервал для мат. ожидания. Во сколько раз надо увеличить объем выборки, чтобы 95% доверительный интервал стал в два раза уже?