Дата последней сборки документа:
11 марта 2021 г. в 06:22

Классические и продвинутые темы теории вероятностей и математической статистики

Подвойский А.О.

Здесь приводятся заметки по некоторым вопросам, касающимся машинного обучения, анализа данных, программирования на языках `Python`, `R` и прочим сопряженным вопросам так или иначе, затрагивающим работу с данными.

Краткое содержание

Содержание

1. Эмпирическая и теоретическая функции распределения

Построим по выборке $X_1, X_2, \ldots, X_n$ случайную ступенчатую функцию $\hat{F}_n(x)$, возрастающую скачками величины $1/n$ в точках $X_{(i)}$ (i-ая порядковая статистика). Эта функция называется *эмпирической функцией распределения.* Чтобы задать значения в точках разрывов, формально определим ее так, чтобы она была непрерывна справа

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_{(i)} \leqslant x\}} = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \leqslant x\}}.$$

В отличие от эмпирической функции распределения выборки, интегральную функцию $F(x)$ распределения генеральной совокупности называют *теоретической функцией распределения.*

Различие между эмпирической и теоретической функциями распределения $F(x)$ состоит в том, что теоретическая функция определяет *вероятность* события $X_i \leqslant x$, а эмпирическая функция $\hat{F}_n(x)$ определяет *относительную частоту* этого события. Из теоремы Бернулли следует, что относительная частота события $X_i \leqslant x$, т.е. $\hat{F}_n(x)$ *стремится по вероятности* к вероятности $F(x)$ этого события, т.е. $\hat{F}_n(x) \xrightarrow{\mathbf{P}} F(x)$. Другими словами числа $\hat{F}_n(x)$ и $F(x)$ мало отличаются одно от другого [1, 191].

2. Доверительные интервалы

Доверительный интервал – интервал, покрывающий неизвестный скалярный параметр θ с заданной *доверительной вероятностью* $(1-\alpha)$

$$P(\hat{\theta}_1(X_1, X_2, \ldots, X_n) < \theta < \hat{\theta}_2(X_1, X_2, \ldots, X_n)) \geqslant 1-\alpha,$$

где $\hat{\theta}_{1,2}$ – нижняя и верхняя граница доверительного интервала (*случайные величины*), α – уровень значимости (она же вероятность ошибки первого рода).

Наиболее часто уровень значимости принимают равным 0.05 или 0.01. Если, например, принят уровень значимости равный 0.05, то означает, что в пяти случаях из ста мы рискуем допустить ошибку первого рода (отвергнуть правильную гипотезу) [1, 284].

Границы доверительного интервала являются *случайными величинами* – функциями от выборки (или другими словами границы доверительного интервала являются *статистиками*) – поэтому правильнее говорить не о вероятности попадания θ в доверительный интервал, а о вероятности того, что доверительный интервал покроет неизвестный параметр θ [1, 216].

Интервалы в нормальной модели Допустим, что элементы выборки X_i распределены по закону $\mathcal{N}(\theta, \sigma^2)$, причем параметр масштаба σ известен, а параметр сдвига θ – нет. Эту модель часто применяют к данным, полученным при независимых измерениях некоторой величины θ с помощью прибора (или метода), имеющего известную среднюю погрешность (стандартную ошибку) σ.

Если случайная величина X распределена нормально $\mathcal{N}(\theta, \sigma^2)$, то выборочная средняя $\bar{X}$, найденная по независимым наблюдениям, также распределена нормально. Параметры распределения таковы [1]

$$\mathbf{E}(\bar{X}) = \theta, \sqrt{\mathbf{D}(\bar{X})} = \frac{\sigma}{\sqrt{n}} \quad \to \quad \bar{X} \sim \mathcal{N}(\theta, \sigma^2/n).$$

Для центрированной и нормированной случайной величины $\sqrt{n}(\bar{X}-\theta)/\sigma \sim \mathcal{N}(0,1)$ в качестве границ интервала с доверительной вероятности $1-\alpha$ можно взять

$$\hat{\theta}_1 = \bar{X} - \sigma/\sqrt{n}\, x_{1-\alpha/2}, \quad \hat{\theta}_2 = \bar{X} + \sigma/\sqrt{n}\, x_{1-\alpha/2}.$$

Таким образом, с вероятностью 0.95 истинное значение параметра сдвига θ находится в интервале $\bar{X} \pm 1.96\,\sigma/\sqrt{n} \approx \bar{X} \pm 2\,\sigma/\sqrt{n}$ (правило двух сигм) [2, 147].

На практике, если значение σ неизвестно, то его заменяют на *состоятельную оценку* $\hat{\sigma} = S$, где $S^2 = \dfrac{1}{2}\sum(X_i - \hat{X})^2$.

Оценка $\hat{\theta}$ параметра θ называется *состоятельной*, если для всех $\theta \in \Theta$ последовательность

$$\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n) \xrightarrow{\mathbf{P}} \theta, \quad n \to \infty.$$

Здесь $\xrightarrow{\mathbf{P}}$ обозначает *сходимость по вероятности*

$$\forall \varepsilon > 0,\ \mathbf{P}(|\hat{\theta} - \theta| > \varepsilon) \to 0, \quad n \to \infty.$$

Состоятельность оценки (а точнее – последовательности оценок $\{\hat{\theta}_n\}$) означает концентрацию вероятностной массы около истинного значения параметра θ с ростом размера выборки n [2, 75].

3. Центральная предельная теорема

Пусть $X_1, \ldots, X_n$ – независимые одинаково распределенные случайные величины. Положим $S_n = X_1 + X_2 + \cdots + X_n$.

Если $0 < \sigma^2 = \mathbf{D}X_1 < \infty$, то

$$S_n^* = \frac{S_n - \mathbf{E}S_n}{\sqrt{\mathbf{D}S_n}} = \frac{S_n - \mu n}{\sigma\sqrt{n}} \xrightarrow{d} Z, \quad n \to \infty,$$

где Z – стандартная нормальная случайная величина, $Z \sim \mathcal{N}(0,1)$.

Пример Пусть случайные величины $Z_1, \ldots, Z_k$ распределены по закону $\mathcal{N}(0,1)$ и независимы. Тогда распределение случайной величины $R_k^2 = Z_1^2 + \cdots + Z_k^2$ называют распределением χ^2 с k степенями свободы (кратко $R_k^2 \sim \chi_k^2$).

Отметим, что каждое слагаемое имеет гамма-распределение с параметрами $\alpha = \lambda = 1/2$, т.е. $Z_i^2 \sim \Gamma(1/2, 1/2)$.

Поскольку R_k^2 – это сумма независимых и одинакового распределенных случайных величин Z_i^2, то согласно *центральной предельной теореме* имеет место *сходимость по распределению*

$$(R_k^2 - \mathbf{E}R_k^2)/\sqrt{\mathbf{D}R_k^2} = (R_k^2 - k)/\sqrt{2k} \xrightarrow{d} Z \sim \mathcal{N}(0,1), \quad k \to \infty.$$

Нормальное приближение является довольно точным уже при $k > 30$.

4. Фактический (достигаемый) уровень значимости

При проверке статистических гипотез в общем случае задается малое число α – вероятность, с которой мы можем позволить себе отвергнуть верную гипотезу (скажем, 0.05). Это число называют *уровнем значимости.*

Исходя из предположения, что гипотеза H верна, определяется *наименьшее* (самое крайнее левое) значение $x_{1-\alpha}$, удовлетворяющее условию

$$\mathbf{P}(T(X_1, \dots, X_n) \geqslant x_{1-\alpha} \,|\, H) = \int\limits_{x_{1-\alpha}}^{+\infty} p_T(x)\, dx \leqslant \alpha.$$

Другими словами, вероятность события, состоящего в том, что статистика примет значение большее $(1-\alpha)$-квантиля (вероятность маловероятного события) должна быть не больше заранее заданного уровня значимости α.

Если функция распределения статистики T непрерывна, то $x_{1-\alpha}$ является, очевидно, ее $(1-\alpha)$-квантилью. Такое $x_{1-\alpha}$ называют *критическим значением*: гипотеза H отвергается, если

$$t_0 = T(x_1, \dots, x_n) \geqslant x_{1-\alpha}$$

(произошло маловероятное событие), и принимается – в противном случае.

При этом величина

$$\alpha_0 = \mathbf{P}(T(X_1, \dots, X_n) \geqslant t_0 \,|\, H) = \int\limits_{t_0}^{+\infty} p_T(x)\, dx$$

задает *фактический (достигаемый) уровень значимости.* Он равен вероятности того, что статистика T (измеряющая степень отклонения полученной реализации от наиболее типичной) за счет случайности примет значение t_0 или даже больше. Другими словами, фактический (достигаемый) уровень значимости оценивает вероятность того, что случайная величина $T(X_1, \dots, X_n)$ попадет в область $[\,t_0, +\infty\,)$, где t_0 – это значение статистики, найденное по выборке.

Фактический (достигаемый) уровень значимости – наименьший уровень, на котором проверяемая гипотеза принимается[1] [2, 161].

Фактический (достигаемый) уровень значимости – это вероятность получить значение статистики как в эксперименте или более экстремальное ее значение при условии справедливости нулевой гипотезы.

Подытожив сказанное выше, можно получить следующее правило: если фактический (достигаемый) уровень значимости α_0 меньше заранее заданного уровня значимости α, то говорят, что данные свидетельствуют против нулевой гипотезы H_0 в пользу альтернативной и у нас есть основания отвергнуть нулевую гипотезу

$$\boxed{\text{если} \quad \alpha_0 < \alpha \quad \text{тогда} \quad \cancel{H_0}}$$

Критическое значение $x_{1-\alpha}$ допускается интерпретировать как квантиль уровня $(1-\alpha)$ только для статистик с непрерывной функцией распределения

[1] Наверное, правильнее говорить *не отвергается*

Вычисление фактического (достигаемого) уровня значимости нередко позволяет избежать категоричных (и при этом ошибочных) выводов, сделанных только на основе сравнения наблюдаемого значения статистики t_0 с критическим значением $x_{1-\alpha}$, найденным для формально заданного α.

5. Теоретические и выборочные квантили

Пусть $\alpha \in (0,1)$. Для *непрерывной* функции распределения F *теоретической α-квантилью* x_α (или квантилью уровня α) называется решение уравнения $F(x_\alpha) = \alpha$, т.е. $x_\alpha = F^{-1}(\alpha)$.

Так же, как и в случае медианы ($\alpha = 1/2$) это решение может быть не единственным.

Оценить x_α можно с помощью порядковой статистики $X_{([\alpha n]+1)}$, где $[\cdot]$ – обозначает целую часть. Эту оценку называют *выборочной α-квантилью.*

6. Ошибки I и II рода

Пример рассмотрим модель $X_i \sim \mathcal{N}(\theta, \sigma^2)$, где дисперсия известна, а математическое ожидание нет. Для проверки гипотезы $H_0 : \theta = \theta_0$ можно применить критерий, основанный на статистике $T(X_1, \cdots, X_n) = \bar{X}$.

Если H_0 верна, то $\bar{X} \sim \mathcal{N}(\theta_0, \sigma^2/n)$. Найдем *критическое значение* t_α из условия

$$\alpha = \mathbf{P}_{\theta_0}(\bar{X} \geqslant t_\alpha).$$

Тогда (центрируем и нормируем случайную величину $\bar{X}$)

$$\alpha = \mathbf{P}\left(\frac{\sqrt{n}(\bar{X}-\theta_0)}{\sigma} \geqslant \frac{\sqrt{n}(t_\alpha-\theta_0)}{\sigma}\right) = 1 - \Phi\left(\frac{\sqrt{n}(t_\alpha-\theta_0)}{\sigma}\right), \text{ так как } \frac{\sqrt{n}(\bar{X}-\theta_0)}{\sigma} \sim \mathcal{N}(0,1),$$

где $\Phi(x)$ – функция распределения закона $\mathcal{N}(0,1)$.

Из последнего соотношения получаем *критическое значение*

$$t_\alpha = \theta_0 + \sigma\, x_{1-\alpha}/\sqrt{n}.$$

Если значение выборочного среднего $\bar{x} \geqslant t_\alpha$, то гипотеза H_0 отвергается. Если нулевая гипотеза верна, то неравенство $\bar{X} \geqslant t_\alpha$ выполняется с вероятностью α. Отвергая в этом случае верную гипотезу H_0, мы совершаем *ошибку I рода.*

С другой стороны, может оказаться, что на самом деле верна не гипотеза H_0, а ее альтернатива $H_1 : \theta = \theta_1$. Если при этом случится, что $\bar{x} < t_\alpha$, то мы примем ошибочную гипотезу H_0 вместо H_1, тем самым допустив *ошибку II рода.*

Найдем вероятность β ошибки II рода для рассматриваемой модели. Когда верна альтернативная гипотеза, выборочное среднее распределено по закону $\mathcal{N}(\theta_1, \sigma^2/n)$, поэтому

$$\beta = \mathbf{P}_{\theta_1}(\bar{X} < t_\alpha) = \Phi\left(\frac{\sqrt{n}(t_\alpha-\theta_1)}{\sigma}\right) = \Phi\left(x_{1-\alpha} - \frac{\sqrt{n}(\theta_1-\theta_0)}{\sigma}\right).$$

Обобщить сказанное выше можно так

$$\alpha = \mathbf{P}(\,\mathrm{Rej}\,H_0^+\,),\ \text{ошибка I рода},$$
$$\beta = \mathbf{P}(\,\mathrm{Rej}\,H_1^+\,) = \mathbf{P}(\,\neg\mathrm{Rej}\,H_0^-\,),\ \text{ошибка II рода}.$$

Гипотеза H_0 заключается в том, что $\theta \in \Theta_0$, а альтернатива H_1 – в том, что $\theta \in \Theta_1$. Когда множество $\Theta_0(\Theta_1)$ состоит из единственной точки, гипотеза H_0 (альтернатива H_1) называется *простой*, иначе – *сложной*.

7. Критерий Холлендера-Прошана

В задачах теории надежности экспоненциальное распределение наработки на отказ $f(x) = \lambda e^{-\lambda x}$ характеризуется значением параметра $\lambda = const$, т.е. постоянством интенсивности отказов изделия во времени.

Отсюда следует, что вероятность безотказной работы изделия за время Δt определяется только промежутком времени Δt и не зависит от того, работало изделие раньше или нет.

Другими словами, вероятность безотказной работы нового изделия и изделия, проработавшего часть времени, должна быть одинакова. Проверка этого обстоятельства и является целью *критерия Холлендера-Прошана* [3, 295].

Статистикой Холлендера-Прошана является величина [2, 182]

$$T_n = \sum_{i>j>k} \psi(X_{(i)}, X_{(j)} + X_{(k)}),$$

где

$$\psi(a,b) = \begin{cases} 1, & \text{если } a > b, \\ 1/2, & \text{если } a = b, \\ 0, & \text{если } a < b. \end{cases}$$

Суммирование здесь производится по всем $n(n-1)(n-2)/6$ упорядоченным тройками (i,j,k), для которых $i > j > k$.

Для достаточно большой выборки можно воспользоваться нормальным приближением (на основании центральной предельной теоремы)

$$\frac{T_n - \mathbf{E}T_n}{\sqrt{\mathbf{D}T_n}} \xrightarrow{d} \xi \sim \mathcal{N}(0,1),$$

где

$$\mathbf{E}T_n = n(n-1)(n-2)/8,\ \mathbf{D}T_n = \frac{3}{2}n(n-1)(n-2)\left[\frac{5}{2592}(n-3)(n-4) + \frac{7}{432}(n-3) + \frac{1}{48}\right].$$

Список литературы

1. *Гмурман В.Е.* Теория вероятностей и математическая статистика. – М.: Высшая школа, 1972. – 368 с.
2. *Лагутин М.Б.* Наглядная математическая статистика. – М.: БИНОМ, 2009. – 472 с.

3. *Кобзарь А.И.* Прикладная математическая статистика. Для инженеров и научных работников. – М.: ФИЗМАТЛИТ, 2012. – 816 с.