Дата последней сборки документа:
9 марта 2021 г. в 09:36

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. 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 $\mathcal{N}(\theta, \sigma)$, то выборочная средняя $\bar{X}$, найденная по независимым наблюдениям, также распределена нормально. Параметры распределения таковы

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

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 $S^2 = \frac{1}{2}\sum(X_i - \hat{X})^2$.

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 $k > 30$.

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.