Фундаментальные теоремы машинного обучения

2020

Содержание

1 Теорема о связи распределений в экспонентном семействе (Бернштейн фон Мизес)

1.1 Асимптотическая нормальность

Пусть заданы объекты из некоторого распределения:

$$\mathbf{X}^n = \{X_i\}_{i=1}^n,$$

где n число объектов.

Пусть задано некоторое открытое подмножество $\mathbf{\Theta} \in \mathbb{R}^d$. Подмножество $\mathbf{\Theta}$ задает множество статистических моделей $\mathcal{P}^n = \{P_\theta^n | \theta \in \mathbf{\Theta}\}$. Пусть для каждого n существует мера P_0^n которая доминирует все меры из множества $\mathcal{P}^n$. Пусть также все меры задаются своей плотностью p_θ^n.

Определение 1. *Рассмотрим некоторую внутреннюю точку $\theta^* \in \mathbf{\Theta}$ и последовательность $\delta_n \to 0$. Пусть существует вектор $\Delta_{\theta^*}^n$ и невырожденная матрица $\boldsymbol{V}_{\theta^*}$, такие, что последовательность $\{\Delta_{\theta^*}^n\}$ ограничена по вероятностной мере, а также для любого компакта $K \subset \mathbb{R}^d$ выполняется:*

$$\sup_{h \in K} \Big| \log \frac{p_{\theta^* + \delta_n h}^n}{p_{\theta^*}^n}\big(\boldsymbol{X}^n\big) - h^\mathsf{T} \boldsymbol{V}_{\theta^*} \Delta_{\theta^*}^n - \frac{1}{2} h^\mathsf{T} \boldsymbol{V}_{\theta^*} h \Big| \overset{P_0^n}{\to} 0.$$

Тогда модель $\mathcal{P}^n$ удовлетворяет условия локальной асимптотической нормальности в точке θ^* (local asymptotic normality).

Априорное распределение заданное на множестве $\mathbf{\Theta}$ обозначим Π, а его плотность π. Предположим, что π положительно в некоторой окрестности точки θ^*.

Апостериорное распределение построенное на основе множестве объектов $\mathbf{X}^n$ обозначим $\Pi_n\big(A|\mathbf{X}^n\big)$, где A некоторое борелевское множество. Будем обозначать случайную величину из апостериорного распределения как ϑ.

1.2 Теорема Бернштейна фон Мизеса

Теорема 1. *Пусть для некоторой точки θ^* выполено условия локальной асимптотической нормальности (Опр.1). Пусть задано априорное распределение Π. Пусть для некоторой последовательности чисел $M_n \to \infty$ выполняется следующее условие:*

$$P_0^n \Pi_n\big(||\vartheta - \theta^*|| > \delta_n M_n | \boldsymbol{X}^n\big) \to 0. \tag{1.1}$$

Тогда последовательность апостериорных распределений сходится к последовательности нормальных:

$$\sup_B \big| \Pi_n\big(\frac{\vartheta - \theta^*}{\delta_n} \in B | \boldsymbol{X}^n\big) - N_{\Delta_{\theta^*}^n, V_{\theta^*}^{-1}}\big(B\big) \big| \overset{P_0^n}{\to} 0.$$

Доказательство. Апостериорное распределение для величины $H = \frac{\vartheta - \theta^*}{\delta_n}$ полученное для выборки $\mathbf{X}^n$ обозначим Π_n. Также обозначим $N_{\Delta_{\theta^*}^n, V_{\theta^*}^{-1}}$ как Φ_n. Рассмотрим некоторый компакт $K \subset \mathbb{R}^d$. Рассмотрим условное апостериорное распределение:

$$\Pi_n^K\big(B|\mathbf{X}^n\big) = \Pi_n\big(B \cap K|\mathbf{X}^n\big) / \Pi_n\big(K|\mathbf{X}^n\big),$$
$$\Phi_n^K\big(B\big) = \Phi_n\big(B \cap K\big) / \Phi_n\big(K\big).$$

Рассмотрим некоторый компакт $K \subset \mathbb{R}^d$. Для любой окрестности $U\big(\theta^*\big) \subset \mathbf{\Theta}$ существует некоторый номер n, такой, что $\theta^* + K\delta_n \subset U\big(\theta^*\big)$.

Рассмотрим функцию $f_n : K \times K \to \mathbb{R}$:

$$f_n(g,h) = \left(1 - \frac{\phi_n(h)s_n(g)\pi_n(g)}{\phi_n(g)s_n(h)\pi_n(h)}\right)_+,$$

где ϕ_n, π_n — распределение Φ_n и Π_n соответственно, s_n является отношением правдоподобия:

$$s_n(h) = \frac{p^n_{\theta^*+h\delta_n}}{p^n_{\theta^*}}.$$

Рассмотрим две произвольные последовательности $\{h_n\}, \{g_n\} \subset K$:

$$\begin{aligned}
&\log \frac{\phi_n(h_n)s_n(g_n)\pi_n(g_n)}{\phi_n(g_n)s_n(h_n)\pi_n(h_n)} = \\
&= (g_n - h_n)^\mathsf{T}\mathbf{V}_{\theta^*}\Delta^n_{\theta^*} + \frac{1}{2}h^\mathsf{T}\mathbf{V}_{\theta^*}h_n - \frac{1}{2}g_n^\mathsf{T}\mathbf{V}_{\theta^*}g_n + o(1) - \\
&= -\frac{1}{2}(h_n - \Delta^n_{\theta^*})^\mathsf{T}\mathbf{V}_{\theta^*}(h_n - \Delta^n_{\theta^*}) + \frac{1}{2}(g_n - \Delta^n_{\theta^*})^\mathsf{T}\mathbf{V}_{\theta^*}(g_n - \Delta^n_{\theta^*}) = o(1),
\end{aligned} \tag{1.2}$$

где первое слагаемое получено используя локальную асимптотическую нормальность (Опр.1), а второе с плотности нормального распределения. Тогда из (1.2) получаем, что:

$$\sup_{g,h\in K} f_n(g,h) \overset{P_0}{\to}_{n\to\infty} 0. \tag{1.3}$$

Обозначим за Ξ_n событие, что $\Pi_n(K) > 0$. Рассмотрим некоторое $\eta > 0$, которое задает следующее множество:

$$\Omega_n = \{\sup_{g,h\in K} f_n(g,h) \le \eta\}_*, \tag{1.4}$$

где $*$ обозначает измеримое покрытие множества. Из (1.3) и (1.4) получаем следующее неравенство:

$$P^n_0||\Pi^K_n - \Phi^K_n||\mathbb{I}_{\Xi_n} \le P^n_0||\Pi^K_n - \Phi^K_n||\mathbb{I}_{\Xi_n\cap\Omega_n} + 2P^n_0||\Pi^K_n - \Phi^K_n||\mathbb{I}_{\Xi_n\setminus\Omega_n}, \tag{1.5}$$

где $\mathbb{I}_{\Xi_n}$ — индикаторная функция, $||\cdot||$ является вариационной нормой (total-variational norm). Второе слагаемое равняется нулю в силу (1.3). Используя свойство данной нормы первое слагаемое принимает следующий вид:

$$\begin{aligned}
&\frac{1}{2}P^n_0||\Pi^K_n - \Phi^K_n||\mathbb{I}_{\Xi_n\cap\Omega_n} = P^n_0\int_K \left(1 - \frac{d\Phi^K_n}{d\Pi^K_n}\right)_+ d\Pi^K_n\mathbb{I}_{\Xi_n\cap\Omega_n} = \\
&= P^n_0\int_K \left(1 - \int_K \frac{s_n(g)\pi_n(g)\phi^K_n(h)}{s_n(h)\pi_n(h)\phi^K_n(g)}d\Phi^K_n(g)\right)_+ d\Pi^K_n\mathbb{I}_{\Xi_n\cap\Omega_n}.
\end{aligned}$$

Используя неравенство Йенсена, а также (1.3) получаем следующее:

$$\frac{1}{2}P^n_0||\Pi^K_n - \Phi^K_n||\mathbb{I}_{\Xi_n\cap\Omega_n} \le P^n_0\int\left(1 - \frac{s_n(g)\pi_n(g)\phi^K_n(h)}{s_n(h)\pi_n(h)\phi^K_n(g)}\right)_+ d\Phi^K_n(g)d\Pi^K_n\mathbb{I}_{\Xi_n\cap\Omega_n} \le \eta.$$

Подставляя в (1.5) получаем, что для любого компакта $K \subset \mathbb{R}^d$ выполняется, что $P^n_0||\Pi^K_n - \Phi^K_n||\mathbb{I}_{\Xi_n} \to 0$.

Рассмотрим последовательность шаров $\{K_m\}$ с центом в нуле с радиусом M_m, причем $M_m \to \infty$.

Рассмотрим множество $\{\Xi_n|\Xi_n = \{\Pi_n(K_n) > 0\}\}$, по условию теоремы (1.1) получим, что $P^n_0(\Xi_n) \to 0$. Также получаем, что $P^n_0||\Pi^{K_n}_n - \Phi^{K_n}_n|| \to 0$.

Теперь рассмотрим $P_0^n||\Pi_n - \Phi_n||$:

$$\begin{aligned} P_0^n||\Pi_n - \Phi_n|| &\leq P_0^n||\Pi_n - \Pi_n^{K_n}|| + P_0^n||\Phi_n - \Phi_n^{K_n}|| \\ &\leq 2\left(\Pi\left(\mathbb{R}^d \setminus K_n\right)\right) + 2\left(\Phi\left(\mathbb{R}^d \setminus K_n\right)\right) \to 0, \end{aligned} \tag{1.6}$$

так как увеличивая радиус компакта в бесконечность мы покроем все множество $\mathbb{R}^d$. Выражение (1.6) заканчивает доказательство данной теоремы. □

Список литературы

[1] *Kleijn, B. J. K., and van der Vaart, A. W.* (2012). The Bernstein-Von-Mises theorem under misspecification. Electronic Journal of Statistics, 6, 354-381. `https://doi.org/10.1214/12-EJS675`

2 Сложность моделей

2.1 Колмогоровская сложность моделей

Нужна мотивация и пара слов о префиксной сложности.

Определение 2. *Колмогоровская сложность:*

$$K(x) = \min_{p\in\{0,1\}^*, i\in\mathcal{N}} \{|p| : T(p) = x\},$$

Перечислим некоторые свойства колмогоровской сложности.
Невычислимость

Теорема 2. *[1] Пусть k — произвольная вычислимая функция. Если $k(x) \leq K(x)$ для всех x, для которых определена k, то k — ограничена.*

Оценка условной Колмогоровской сложности [1]

$$K(x, y) \leq K(x) + K(y|x) + O(\log K(x, y)).$$

Разность $I(x : y) = K(y) - K(y|x)$ задает количество информации в x об объекте y. **Количество информации в паре x,y симметрично с точностью до константы**:

$$I(x : y) = I(y : x) + O(\log K(x, y)).$$

Отметим, что схожими свойствами обладает взаимная информация:

$$I(x, y) = H(x) - H(x|y), \quad H(x) = -\sum_i p_x(x_i) \log p_x(x_i);$$

$$I(x, y) = I(y, x).$$

Теорема 3. *[2] Пусть f — вычислимая функция вероятности на пространстве бинарных векторов произвольной длины. Тогда*

$$0 \leq \left(\mathsf{E}_f K(X) - H(x)\right) \leq K(f) + O(1).$$

Доказательство. K(X) — это длина префиксного кода для x, по noiseless coding theorem:

$$H(X) \leq \mathsf{E}_f K(X).$$

$$f(x) \leq 2^{K(f)+O(1)}(K(x) + O(1)) :$$

$$\log \frac{1}{f(x)} \geq K(x) - K(f) - O(1) :$$

$$\sum_x f(x)K(X) \leq H(X) + K(f) + O(1).$$

□

2.2 Принцип минимальной длины описания

Рассмотрим префиксную сложность, которая описывается машиной Тьюринга специального вида, имеющих однонаправленную ленту для чтения и двунаправленную рабочую ленту. Будем полагать что машина Тьюринга T останавливается на p с выводом x: $T(p) = x$, если вся зхапись p осталась слева от читающей каретки, x осталась слева от пишушщей каретки и T остановлена.

Определение 3. *Префиксная Колмогоровская сложность:*

$$K(x) = \min_{p \in \{0,1\}^*, i \in \mathcal{N}} \{|i| + |p| : T_i(p) = x\},$$

2.3 Вероятностная интерпретация минимальной длины описания

Список литературы

[1] Успенский В., Шень А., Верещагин Н. Колмогоровская сложность и алгоритмическая случайность. – Litres, 2017

[2] Grunwald P., Vitányi P. Shannon information and Kolmogorov complexity //arXiv preprint cs/0410002. – 2004.