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:*

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

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:

$$\sup_B \Big| \Pi_n\big(\frac{\vartheta - \theta^*}{\delta_n} \in B | \boldsymbol{X}^n\big) - N_{\Delta^n_{\theta^*}, V^{-1}_{\theta^*}}\big(B\big)\Big| \xrightarrow{P_0^n} 0.$$

Доказательство. Апостериорное распределение для величины $H = \frac{\vartheta - \theta^*}{\delta_n}$ полученное для выборки $\mathbf{X}^n$ обозначим Π_n. Также обозначим $N_{\Delta^n_{\theta^*}, V^{-1}_{\theta^*}}$ 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:

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

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

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

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

$$P_0^n||\Pi_n^K - \Phi_n^K||\mathbb{I}_{\Xi_n} \leq P_0^n||\Pi_n^K - \Phi_n^K||\mathbb{I}_{\Xi_n\cap\Omega_n} + 2P_0^n||\Pi_n^K - \Phi_n^K||\mathbb{I}_{\Xi_n\setminus\Omega_n}, \tag{1.5}$$

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:

$$\begin{aligned}
\frac{1}{2}P_0^n||\Pi_n^K - \Phi_n^K||\mathbb{I}_{\Xi_n\cap\Omega_n} &= P_0^n\int_K \left(1 - \frac{d\Phi_n^K}{d\Pi_n^K}\right)_+ d\Pi_n^K\mathbb{I}_{\Xi_n\cap\Omega_n} = \\
&= P_0^n\int_K \left(1 - \int_K \frac{s_n(g)\pi_n(g)\phi_n^K(h)}{s_n(h)\pi_n(h)\phi_n^K(g)}d\Phi_n^K(g)\right)_+ d\Pi_n^K\mathbb{I}_{\Xi_n\cap\Omega_n}.
\end{aligned}$$

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

$$\frac{1}{2}P_0^n||\Pi_n^K - \Phi_n^K||\mathbb{I}_{\Xi_n\cap\Omega_n} \leq P_0^n\int \left(1 - \frac{s_n(g)\pi_n(g)\phi_n^K(h)}{s_n(h)\pi_n(h)\phi_n^K(g)}\right)_+ d\Phi_n^K(g)d\Pi_n^K\mathbb{I}_{\Xi_n\cap\Omega_n} \leq \eta.$$

Подставляя в (1.5) получаем, что для любого компакта $K \subset \mathbb{R}^d$ выполняется, что $P_0^n||\Pi_n^K - \Phi_n^K||\mathbb{I}_{\Xi_n} \to 0$.

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 $P_0^n(\Xi_n) \to 0$. Также получаем, что $P_0^n||\Pi_n^{K_n} - \Phi_n^{K_n}|| \to 0$.

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

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

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

TODO: обозначения

Одним из фундаментальных способов определить сложность произвольного математического объекта является колмогоровская сложность. Ниже представлено формальеное определине колмогоровской сложности и основные ее соуйства.

Определение 2. *Способом описания назовем вычислимое частинчо определенное отображение из множества бинарных слов в себя:*

$$D : \{0,1\}^* \to \{0,1\}^*.$$

Определение 3. *Пусть задан некоторый способ описания D. Колмогоровской сложностью бинарной строки x назовем минимальную длину описания относительно D:*

$$K_D(x) = \min_{p \in \{0,1\}^*} \{|p| : D(p) = x\},$$

Перечислим некоторые свойства колмогоровской сложности [1].

Независимости от способа написания.

Теорема (без доказательств) 1. *Пусть заданы отображения D_1, D_2, такие что существуют константы c_1, c_2 такие что для любого другого отображения D' и для любой строки x:*

$$K_{D_1}(x) \leq K_{D'}(x) + c_1, \quad K_{D_2}(x) \leq K_{D'}(x) + c_2.$$

Тогда $K_{D_1}(x) = K_{D_2}(x) + O(1)$.

Т.к. колмогоровская сложность независима от способа написания, зафиксируем некоторый способ описания D и положим $K(x) = K_D(x)$.

Невычислимость

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

Из теоремы следует, что колмогровская сложность в общем случае невычислима: любая оценка сложности будет ограничена, и потому тривиальна.

Условная сложность Обобщим понятие колмогоровскорй сложности на случай двух бинарных строк.

Определение 4. *Пусть задано вычисилмое и частично определенное отображение из декартового произведения двух множеств бинарных слов в себя:*

$$D : \{0,1\}^* \times \{0,1\}^* \to \{0,1\}^*.$$

Условной колмогоровской сложностью бинарной строки y при условии x назовем минимальную длину описания относительно D:

$$K_D(y|x) = \min_{p \in \{0,1\}^*} \{|p| : D(p, y) = x\},$$

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, определение которой дано ниже.

Определение 5. *Пусть задана дискретная случайная величина x с вероятностным распределением p, принимающая значения $x_1, \ldots, x_n$, Энтропией распределения случайной величины x назовем:*

$$H(x) = -\sum_{i=1}^{n} p(x = x_i) \log\ p(x = x_i).$$

Взаимной информацией I двух случайных величин x, y назовем следующее выражение:

$$I(x, y) = H(x) - H(x|y), \quad H(x) = -\sum_{i} p_x(x_i) \log p_x(x_i)$$

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

Таким образом, свойства количества информации $I(x : y)$ и взаимной информации, во многом совпадают. Докажем теорему о связи колмогоровской сложности и энтропии распределения, подытоживающую связь этих двух математических объектов.

Теорема 2. *[2] Пусть задано семейство частично-определенных отображений $\mathfrak{D} = \{D : \{0,1\}^* \to \{0,1\}^*\}$, такое что для любого отображения $D \in \mathfrak{D}$ и элемента из области определения D в области определения не содержится префиксов этого элемента. Пусть f — вычислимая функция вероятности на пространстве бинарных векторов произвольной длины. Тогда*

$$0 \le \big(\mathsf{E}_f K(X) - H(x)\big) \le K(f) + O(1). \tag{2.1}$$

Для доказательства предварительно приведем две теоремы из [2] без доказательства.

Теорема (без доказательств) 3. *Пусть задано семейство частично-определенных отображений $\mathfrak{D} = \{D : \{0,1\}^* \to \{0,1\}^*\}$, такое что для любого отображения $D \in \mathfrak{D}$ и элемента из области определения D в области определения не содержится префиксов этого элемента.*

Тогда для минимальной средней длины описания слова:

$$L = \min_{D \in \mathfrak{D}} \sum_{i} |D(x_i)| p(x = x_i)$$

справедливо неравенсвто:

$$H(x) \le L \le H(x) + 1.$$

Теорема (без доказательств) 4. *Пусть f — вычислимое распределение на бинарных словах. Тогда справедлива следующие оценки:*

$$2^{K(f) \pm O(1) - K(x)} \ge f(x),$$

где $O(1)$ — длина некоторой программы, не зависящей от f, x.

Перейдем к доказательству основной теоремы.

Доказательство. Т.к. $K(X)$ — это длина кода для x, то по теореме 3:

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

Таким образом левая часть неравенства (2.1) доказана.

По теореме 4:

$$f(x) \le 2^{K(f) \pm O(1) - K(x)}.$$

Тогда

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

Посчитаем матожидание данной величины по всем x:

$$H(x) \geq \sum_x f(x)K(f) - \sum_x O(1) - \sum_x K(x).$$

Пользуясь тем, что $\sum_x f(x) = 1$ получим итоговую формулу для правой части неравенства:

$$H(x) + O(1) + K(x) \geq \sum_x f(x)K(f),$$

что и т.д. □

2.2 Колмогоровская сложность и принцип минимальной длины описания

Рассмотрим задачу выбора модели для заданной выборки. Будем полагать что заданная выборка описывается в виде некоторй бинарной строки x. В дальнейшем будем отождествлять выборки и ее бинарное описание x.

Для этого рассмотрим частный случай колмогоровской сложности, называемый префиксной колмогоровской сложностью. Эта сложность задается машиной Тьюринга специального вида, имеющей две ленты: 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 запись p осталась слева от читающей каретки, x осталась слева от пишушщей каретки и T остановлена.

Определение 6. *Префиксная Колмогоровская сложность:*

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

где $|i|$ — длина описания i-й префиксной машины Тьюринга.

Задачу выбора модели для выборки можно рассматривать как задачу нахождения префиксной колмогоровской сложности для выборки. В случае, если модель является дискриминативной, то вместо колмогоровсокй сложности можно использовать условную колмогоровскую сложность.Т.к. колмогоровская сложность невычислима, рассмотрим упрощенный подход к выбору модели: вместо колмогоровской сложности строки x будем искать некоторое множество S, в которое входит x, и чья сложность описания при помощи машины Тьюринга невелика. Таким образом, мы сможем найти "хорошую" машину Тьюрингу не для конкретной строки, а для некоторого семейства строк (или выборок), обладающих некоторыми общими свойствами или регулярностью.

Определение 7. *Сложностью конечного множества S назовем следующей величину:*

$$K(S) = \min_{p \in \{0,1\}^*, i \in \mathcal{N}} \{|i| + |p| : T_i(p) \text{ перечисляет все элементы множества } S\}.$$

Вместо задачи нахождения минимальной сложности для выборки x будем искать множество S, которое описывается некоторой машиной Тьюринга, и в которое входит заданная строка x. Приведем формулу для оценки разности между сложность ю выборки x и множества S, в которое входит данная выборка.

Теорема (без доказательств) 5. *Для любого $x \in S$ справедливо неравенство [3]:*

$$K(x) - K(S) \geq +\log|S| + O(1).$$

На практике задача выбора модели подразумевает, что мы можем выбрать модель, которая описывает выборку (или множество выборок) S неидеально, а с некоторым допустимым уровнем потери информации. Тогда задача выбора модели для заданной выборки ставится следуюющим образом:

$$\arg\min_S \{\log |S| + K(S) : x \in S, K(s) \leq \alpha, \tag{2.2}$$

где α — максимально допустимая сложность множества S.

Заметим, что решение задачи выбора модели в приведенном выше виде является вычислимой, то есть можно предложить алгоритм, вычисляющий данную задачу. Приведем схему данного алгоритма:

1. Положим $\hat{p}, \hat{S}$ неопределенным.
2. Для всех $S, p : T(p) = S, K(S) \leq \alpha$:
3. Если $\hat{S}$ неопределен или $|p| + \log(S) \leq \hat{p} + \log \hat{S}$, то $\hat{p}, \hat{S} = p, S$.

Т.к. множество пар (S, p) конечно (TODO: почему?), то алгоритм остановится, а потому вычислим. По построению он также доставлячет решение оптимизационной задачи (2.2).

TODO: оценка ошибки

TODO: про то что это MDL

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.

[3] Vereshchagin N. K., Vitányi P. M. B. Kolmogorov's structure functions and model selection //IEEE Transactions on Information Theory. – 2004. – Т. 50. – №. 12. – С. 3265-3290.