

Фундаментальные теоремы машинного обучения

2020

Содержание

1	Теорема о связи распределений в экспонентном семействе (Бернштейн фон Мизес)	2
1.1	Асимптотическая нормальность	2
1.2	Теорема Бернштейна фон Мизеса	2
2	Сложность моделей	5
2.1	Колмогоровская сложность моделей	5
2.2	Колмогоровская сложность и принцип минимальной длины описания	7
2.3	Вероятностная интерпретация минимальной длины описания	8

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 $\Theta \in \mathbb{R}^d$. Подмножество Θ задает множество статистических моделей $\mathcal{P}^n = \{P_\theta^n | \theta \in \Theta\}$. Пусть для каждого n существует мера P_θ^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 $\theta^* \in \Theta$ и последовательность $\delta_n \rightarrow 0$. Пусть существует вектор $\Delta_{\theta^*}^n$ и невырожденная матрица V_{θ^*} , такие, что последовательность $\{\Delta_{\theta^*}^n\}$ ограничена по вероятностной мере, а также для любого компакта $K \subset \mathbb{R}^d$ выполняется:

$$\sup_{h \in K} \left| \log \frac{p_{\theta^* + \delta_n h}^n(\mathbf{X}^n)}{p_{\theta^*}^n(\mathbf{X}^n)} - h^\top V_{\theta^*} \Delta_{\theta^*}^n - \frac{1}{2} h^\top V_{\theta^*} h \right| \xrightarrow{P_\theta^n} 0.$$

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

Априорное распределение заданное на множестве Θ 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 $\Pi_n(A|\mathbf{X}^n)$, 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^* выполнено условия локальной асимптотической нормальности (Опр.1). Пусть задано априорное распределение Π . Пусть для некоторой последовательности чисел $M_n \rightarrow \infty$ выполняется следующее условие:

$$P_0^n \Pi_n(\|\vartheta - \theta^*\| > \delta_n M_n | \mathbf{X}^n) \rightarrow 0. \quad (1.1)$$

Тогда последовательность апостериорных распределений сходится к последовательности нормальных:

$$\sup_B \left| \Pi_n\left(\frac{\vartheta - \theta^*}{\delta_n} \in B | \mathbf{X}^n\right) - N_{\Delta_{\theta^*}^n, V_{\theta^*}^{-1}}(B) \right| \xrightarrow{P_\theta^n} 0.$$

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

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

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

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

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

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

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

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

$$\begin{aligned} \log \frac{\phi_n(h_n)s_n(g_n)\pi_n(g_n)}{\phi_n(g_n)s_n(h_n)\pi_n(h_n)} &= \\ &= (g_n - h_n)^\top \mathbf{V}_{\theta^*} \Delta_{\theta^*}^n + \frac{1}{2} h^\top \mathbf{V}_{\theta^*} h_n - \frac{1}{2} g_n^\top \mathbf{V}_{\theta^*} g_n + o(1) - \\ &= -\frac{1}{2} (h_n - \Delta_{\theta^*}^n)^\top \mathbf{V}_{\theta^*} (h_n - \Delta_{\theta^*}^n) + \frac{1}{2} (g_n - \Delta_{\theta^*}^n)^\top \mathbf{V}_{\theta^*} (g_n - \Delta_{\theta^*}^n) = o(1), \end{aligned} \quad (1.2)$$

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

$$\sup_{g, h \in K} f_n(g, h) \xrightarrow{P_0} 0. \quad (1.3)$$

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

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

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

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

где \mathbb{I}_{Ξ_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 $P_0^n \|\Pi_n^K - \Phi_n^K\|_{\mathbb{I}_{\Xi_n}} \rightarrow 0$.

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

Рассмотрим множество $\{\Xi_n | \Xi_n = \{\Pi_n(K_n) > 0\}\}$, по условию теоремы (1.1) получим, что $P_0^n(\Xi_n) \rightarrow 0$. Также получаем, что $P_0^n \|\Pi_n^{K_n} - \Phi_n^{K_n}\| \rightarrow 0$.

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

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

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

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 формальное определение колмогоровской сложности и основные ее свойства.

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

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

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 колмогоровская сложность в общем случае невычислима: любая оценка сложности будет ограничена, и потому тривиальна.

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

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

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

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, \dots, 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 $\mathfrak{D} = \{D : \{0, 1\}^* \rightarrow \{0, 1\}^*\}$, 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

$$0 \leq (\mathbb{E}_f K(X) - H(x)) \leq K(f) + O(1). \quad (2.1)$$

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

Теорема (без доказательств) 3. Пусть задано семейство частично-определенных отображений $\mathfrak{D} = \{D : \{0, 1\}^* \rightarrow \{0, 1\}^*\}$, 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 неравенство:

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

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

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

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:

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

Таким образом левая часть неравенства (2.1) доказана.

По теореме 4:

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

Тогда

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

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 некоторой 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 пишущей 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 колмогоровской 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 сложностью 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