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 PAC-обучаемость и сжатие 5

2.1 PAC-learning . . . 5

2.2 Sample Compression scheme . . . 5

2.3 Compression implais learning . . . 5

3 Сложность моделей 7

3.1 Колмогоровская сложность моделей . . . 7

3.2 Колмогоровская сложность и принцип минимальной длины описания . . . 9

3.3 Вероятностная интерпретация минимальной длины описания . . . 10

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 $\mathcal{P}^n = \{P^n_\theta | \theta \in \mathbf{\Theta}\}$. Пусть для каждого n существует мера P^n_0 которая доминирует все меры из множества $\mathcal{P}^n$. Пусть также все меры задаются своей плотностью p^n_θ.

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

$$\sup_{h \in K} \Big| \log \frac{p^n_{\theta^* + \delta_n h}}{p^n_{\theta^*}} \big(\boldsymbol{X}^n\big) - h^\mathsf{T} \boldsymbol{V}_{\theta^*} \Delta^n_{\theta^*} - \frac{1}{2} h^\mathsf{T} \boldsymbol{V}_{\theta^*} h \Big| \xrightarrow{P^n_0} 0.$$

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:*

$$P^n_0 \Pi_n\big(||\vartheta - \theta^*|| > \delta_n M_n | \boldsymbol{X}^n\big) \to 0. \tag{1.1}$$

Тогда последовательность апостериорных распределений сходится к последовательности нормальных:

$$\sup_B \Big|\Pi_n\big(\frac{\vartheta - \theta^*}{\delta_n} \in B | \boldsymbol{X}^n\big) - N_{\Delta^n_{\theta^*}, V^{-1}_{\theta^*}}\big(B\big)\Big| \xrightarrow{P^n_0} 0.$$

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:

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

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 (2.1) и (2.2) 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 (2.1). 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 (2.1) 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 (2.3) 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 $P_0^n||\Pi_n - \Phi_n||$:

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

так как увеличивая радиус компакта в бесконечность мы покроем все множество $\mathbb{R}^d$. Выражение (2.4) 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 PAC-обучаемость и сжатие

2.1 PAC-learning

Определение 2. *Класс гипотез $\mathcal{H}$ является PAC-обучаемым над множеством объектов $Z = z_1, \cdots z_d$ для функции потерь $l : \mathcal{H} \times Z \to \mathbb{R}_+$, если существует функция $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ и алгоритм $\mathcal{A}$ со свойством: для всех $\varepsilon, \delta \in (0,1)$, для любого распределения $\mathcal{D}$ над множеством объектов Z алгоритм $\mathcal{A}$ возвращает такое $h \in \mathcal{H}$, что с вероятностью $1-\delta$ выполняется:*

$$\mathsf{E}_{z\sim\mathcal{D}}\left[l\big(h,z\big)\right] \leq \min_{h'\in\mathcal{H}} \mathsf{E}_{z\sim\mathcal{D}}\left[l\big(h',z\big)\right] + \varepsilon.$$

2.2 Sample Compression scheme

Схема сжатия данных с параметром k состоит из двух отображений (κ, ρ) :

1. κ получает на вход выборку S, а на выходе получаем пару (S', I), где $|S'| = k$;
2. ρ получает на вход пару (S', I) на выходе выдает гипотезу h.

Причем выполняется следующее условие:

1. $\kappa\big(Y, y\big) = ((Z, z), I)$;
2. $\rho\big(\kappa\big(Y, y\big)\big)|_Y = y$.

2.3 Compression implais learning

Любую схему сжатия с параметром k можно рассматривать как алгоритм обучения $A = \rho \circ \kappa$. То что данный алгоритм PAC-обучаем доказывает следующая теорема.

Теорема 2. *Алгоритм обучения $A = \rho \circ \kappa$ является PAC-обучаемым, то есть*

$$P\big[p\big(\left\{h\big(z\big) \neq f\big(z\big)\right\}\big) > \varepsilon\big] \leq |I| \sum_{j=1}^{k} \binom{d}{j} (1-\varepsilon)^{m-j},$$

где p распределение над Z.

Доказательство. Сначала заметим, что всего существует

$$\sum_{j=1}^{k} \binom{d}{j}$$

подмножеств T множества Z размера не более k. С другой стороны всего есть $|I|$ вариантов выбрать информацию сжатия $i \in I$. Из выше описанного получаем, что каждой паре (T, i) соответствует своя функция

$$h_{T,i} = \rho\big((T, i), i\big).$$

С построения $h_{T,i}$ следует, что $h_{T,i}$ не зависит от $Z \setminus T$, тогда получаем, что если

$$p\big(\left\{h_{T,i}\big(x\big) \neq f(x)\right\}\big) \geq \varepsilon,$$

то для всех $m - |T|$ выполняется получаем, что

$$\prod_{t=1}^{m-|T|} p\big(\left\{h_{T,i}\big(x\big) \neq f(x)\right\}\big) \leq (1-\varepsilon)^{m-|T|}. \tag{2.1}$$

Получаем, что для любого $h_{T,i}$ выполняется неравенство (2.1).

И того получаем, что для произвольной $h_{T,i}$ выполняется неравенство:

$$P\big[p\big(\,\{h_{T,i}(z) \neq f(z)\}\big) > \varepsilon\big] \leq (1-\varepsilon)^{m-|T|}\,,$$

Рассмотрим множество функций при фиксированном $i \in I$:

$$\mathcal{H}_i = \{h_{T,i} : |T| \leq k\}\,, \tag{2.2}$$

тогда для алгоритма A для подмножества функций, которые получены при помощи сжатой информации i получаем:

$$P\big[p\big(\,\{h_{T,i}(z) \neq f(z)\}\big) > \varepsilon\big] \leq \sum_{j=1}^{k} \binom{d}{j} (1-\varepsilon)^{m-j}\,, \tag{2.3}$$

где $h_{T,i}$ это лучший алгоритм из множества $\mathcal{H}_i$.

Теперь заметим, что финальная функция h принадлежит множеству:

$$\mathcal{H}_{\kappa,\rho} = \{h_{T,i} : |T| \leq k, i \in I\}\,. \tag{2.4}$$

Вспомним, что для каждого T таких функций $|I|$, из чего уже для произвольного h используя выражение (2.3) имеем следующее неравенство:

$$P\big[p\big(\,\{h(z) \neq f(z)\}\big) > \varepsilon\big] \leq |I| \sum_{j=1}^{k} \binom{d}{j} (1-\varepsilon)^{m-j}\,,$$

что и доказывает исходную теорему. $\square$

Список литературы

[1] *Floyd, S., Warmuth, M.* (1995) Sample Compression, Learnability, and the Vapnik-Chervonenkis Dimension. // Machine Learning 21, 269–304. `https://doi.org/10.1023/A:1022660318680`

[2] *В.В.Вьюгин* КОЛМОГОРОВСКАЯ СЛОЖНОСТЬ И АЛГОРИТМИЧЕСКАЯ СЛУЧАЙНОСТЬ (2012) // МФТИ.

[3] *Shay Moran, Amir Yehudayoff* Sample Compression Schemes for VC Classes (2015) // `https://www.cs.bgu.ac.il/~adsmb182/wiki.files/meni-lecture.pdf`

$$T(f) \in \{0,1\}^+ \xrightarrow{\arg\min|f|} x \in \{0,1\}^+$$

Рис. 1

3 Сложность моделей

3.1 Колмогоровская сложность моделей

Одним из фундаментальных способов определить сложность произвольного математического объекта является колмогоровская сложность. Ниже представлено формальное определение колмогоровской сложности и основные ее свойства.

Определение 3. *Пусть задано вычислимое частично определенное отображение из множества бинарных слов в себя:*

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

Колмогоровской сложностью бинарной строки x назовем минимальную длину описания относительно T:

$$K_T(x) = \min_{f \in \{0,1\}^*} \{|f| : T(f) = x\},$$

Заметим, что колмогоровская сложность зависит от отображения T. В [1] доказано, что колмогоровсксая сложность $K_T(x)$ при двух отображениях T_1, T_2 отличается лишь на некоторую константу, не зависящих от строки x. Поэтому для дальнейшего изложения зафиксируем некоторое отображение T и положим $K(x) = K_T(x)$.

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:*

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

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

$$K_T(y|x) = \min_{f \in \{0,1\}^*} \{|f| : T(f, y) = x\},$$

Аналогично простой колмогоровской сложности, зафиксируем некоторое отображение T и положим $K_T(y|x) = K(y|x)$.

Рассмотрим некоторые свойства условной колмогоровской сложности.

Оценка условной Колмогоровской сложности [1]

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

Количество информации в паре x,y симметрично с точностью до константы:

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

где величина $I(x : y) = K(y) - K(y|x)$ задает количество информации в x об объекте y.

Отметим, что схожими свойствами обладает взаимная информация и энтропия, определения которых даны 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:*

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

Оценка условной энтропии

$$H(x, y) = H(x) + K(y|x).$$

Определение 6. *Взаимной информацией I двух случайных величин x, y назовем следующее выражение:*

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

Взаимная информация симметрична:

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

Таким образом, свойства энтропии и колмогровской сложности, а также 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 3. *[1] Пусть задана некоторая строка x длины n с частотами $p = (p_0, 1 - p_0)$ появлений нулей и единиц в строке. Тогда*

$$K(x) \leq H(x) + O(\log m).$$

Неравенство обращается в равенство для большинства строк x длины n.

Доказательство. Всего слов, которые можно получить с использованием заданных частот:

$$C = \frac{m!}{(p_0 m)!((1 - p_0)m)!}.$$

Т.к. количество таких слов конечно, то их можно пронумеровать и построить отображение, выдающее строку x по ее порядковому номеру. Таким образом, условная колмогоровская сложность ограничена сверху:

$$K(x|C, p_0) \leq \log C + O(1).$$

Воспользуемся формулой Стирлинга:

$$n! = \sqrt{(2\pi + o(1))n}\left(\frac{n}{e}^n\right).$$

И получим оценку:

$$C \leq 2^{nH(x)+O(\log n)}, \quad K(x|C, p_0) \leq nH(x) + O(\log n).$$

Для того, чтобы избавиться от условия в $K(x|C, p_0)$ потребуется $O(\log n)$ бит для описания чисел $p_0 n, n$.

TODO: Поскольку слов с более короткими описаниями меньше, чем C, то для большинства слов будет достигаться предложенная оценка. $\square$

Частным случаем колмогоровской сложности является префиксная колмогоровская сложность. 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. Колмогоровскую сложность относительно префисных машин Тьюринга назовем префиксной колмогоровской сложностью KP. Отметим, что префиксная колмогоровская сложность является частным случаем колмогоровской сложности, а потому:

$$K(x) \leq KP(x).$$

Теорема (без доказательств) 1. *Пусть задана вычислимая функция p_x вероятности на множестве бинарных строк. Тогда*

$$0 \leq \mathsf{E}_{p_x} KP(x) - H(x) \leq KP(p_x) + O(1),$$

где $K(p_x)$ определяется как минимальная длина программы для префиксной машины Тьюринга, вычисляющей p_x.

Таким образом, существует связь между энтропией и колмогровоской сложностью (как для обычного варианта сложности, так и для префиксной колмогоровской сложности): для простых распределений в смысле сложности фукнкции p_x энтропия будет приближать матожиданием префиксной колмогровской сложности.

3.2 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 выборку и ее бинарное описание x.

Задачу выбора модели для выборки можно рассматривать как задачу нахождения колмогоровской сложности для выборки. В случае, если модель является дискриминативной, то вместо колмогоровсокй сложности можно использовать условную колмогоровскую сложность. В общем виде колмогоровская сложность и префиксная сложность невычислимы [1], поэтому рассмотрим упрощенный подход к выбору модели: вместо сложности строки x будем искать некоторое множество S, в которое входит x, и чья префиксная сложность описания невелика. Таким образом, мы сможем найти "хорошую" машину Тьюрингу не для конкретной строки, а для некоторого семейства строк (или выборок), обладающих общими свойствами или регулярностью.

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

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

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) 2. *Для любого $x \in S$ справедливо неравенство [3]:*

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

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 следующим образом:

$$\arg\min_S \{\log |S| + K(S) : x \in S, K(s) \leq \alpha, \tag{3.1}$$

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:

Положим $\hat{p}, \hat{S}$ неопределенными;
for all $S, p : T(p) = S, |p| \leq \alpha$ **do**
  if $\hat{S}$ неопределен или $|p| + \log(S) \leq \hat{p} + \log \hat{S}$, **then**
    $\hat{p}, \hat{S} = p, S$

$$\begin{tikzcd} & S \arrow[dr, "\arg\min_S\{\log|S| + K(S) : x \in S, K(s) \leq \alpha"] & \\ f \arrow[ur, "\arg\min|f|"] & & x \end{tikzcd}$$

Рис. 2

end if
end for
return $\hat{S}$

Т.к. множество всех программ с длиной менее α является конечным, то алгоритм остановится, а потому вычислим. По построению он также доставлячет решение оптимизационной задачи (3.1).

Теорема (без доказательств) 3. *Обзначим за $S^*, K(S^*) \leq \alpha$ множество, доставляющее минимум следующей функции:*

$$\delta(S) = \log|S| - K(x|S).$$

Тогда $\delta(S^*) - \delta\hat{S} = O(\log|x|)$.

Таким образом, пара $(\hat{p}, \hat{S})$ доставляет минимуму суммы $|p| + \log(S)$ при ограничениях на длину программы $|p| \leq \alpha$. Первое слагаемое данной суммы — это длина программы, то есть сложность описания множества $\hat{S}$. Сама сумма $|p| + \log(S)$ характеризует длину кода, требуемого для двухэтапного описания строки x, где на первом этапе мы описываем множество S, а на втором этапе находим x во множестве S.

Описанный метод получения кода для описания выборки x соответствует *Принципу минимальной длины описания* или MDL: для заданной выборки требуется найти минимальный код, который описывает выборку (возможно, с некоторой наперед заданной допустимой ошибкой).

3.3 Вероятностная интерпретация минимальной длины описания

Рассмотрим подробнее задачу кодирования выборки. Задачу можно рассматривать как проблему передачу информации от кодировщика декодировщику. Задана выборка $\mathbf{X}, x \in \mathbf{X}$. Кодировщик кодирует информацию о выборке $\mathbf{X}$ с помощью некоторого кода $\mathbf{f}$ и передает ее декодировщику. Декодировщик декодирует код $\mathbf{f}(\mathbf{X})$, полученный от кодировщика и восстанавливает исходную выборку $\mathbf{X}$ (возможно, с некоторой потерей информации, если это оговорено заранее). Допустим, для кодирования информации о выборке доступно несколько методов кодирования, при этом для разных объектов выборки минимальный по количеству информации метод будет отличаться. Тогда кодировщик должен передать не только информацию о выборке, но и информацию о самом методе кодировния. Проблему выбора оптимального способа кодирования из нескольких можно рассмотреть при помощи вероятностной интерпретации минимальной длины описания.

Теорема (без доказательств) 4. *Пусть задан конечное или счетное множество Z с введенной на нем вероятностью P. Тогда существует код C, такой что для любого $x \in Z$ длина его описания L будет равняться $L(C(x)) = -\log P(x)$.*

Таким образом, вместо методов кодирования можно рассматривать задачу выбора функции вероятности (правдоподобия?), введенной на признаковом описании выборки. Задача формулируется следующим образом:

$$\log p(\mathbf{X}|\mathbf{f}) + L(\mathbf{f}),$$

где $L(\mathbf{f})$ — длина описания вероятностной модели $\mathbf{f}$.

Одним из критериев качества вероятностного кодирования с помощью смеси кодов относительно фиксированного кодирования является регрет (TODO):

$$R(x) = -\log P(x) + \min_{\mathbf{f}}(\log P(x|\mathbf{f})).$$

Регрет характеризует разницу между длиной рассматриваемого $\log P(x)$ кода для x в сравнении с наилучшим кодом из некоторого множества $\mathbf{f}$.

Пусть модель $\mathbf{f}$ зависит от вектора параметра $\mathbf{w}$:

$$\log p(x|\mathbf{f}) = \log p(x|\mathbf{w}(x)),$$

где $\mathbf{w}(x)$ — оптимальные параметры для объекта x.

Тогда регрет выглядит следующим образом:

$$R(P, x) = -\log P(x) + \min_{\mathbf{w}}(\log P(x|\mathbf{w}(x)).$$

Для всей выборки регрет определяется следующим образом:

$$R(\mathbf{X}) = \max_{x \in \mathbf{X}}(-\log P(x) + \min_{\mathbf{w}}(\log P(x|\mathbf{w}(w))).$$

Теорема 4 ((Штарьков, 1987) **??**,**??**)**.** *Пусть величина*

$$\log \sum_{x} p(x|\mathbf{w}(x))$$

конечна. Тогда следующее выражение доставляет единственный минимум регрета $R(\mathbf{X})$:

$$\frac{p(x|\mathbf{w}(x))}{\sum_{x' \in \mathbf{X}} p(x'|\mathbf{w}(x'))}.$$

Доказательство. Рассмотрим выражение регрета для данной вероятностной меры:

$$-\log P(x) + \min_{\mathbf{w}}(\log P(x|\mathbf{w}(w))) = \log \sum_{x} p(x|\mathbf{w}(x)).$$

Выражение регрета не зависит от объекта выборки x. Заметим, что для любых двух отличных распределений p_1, p_2 существует хотя бы один элемент, такой что

$$p_1(x) < p_2(x).$$

Действительно, пусть это неверно: $\forall x p_1(x) \geq p_2(x), \exists x' : p_1(x) > p_2(x)$. Просуммируем вероятности p_1 и p_2 и воспользуемся тем, что сумма вероятностей по всем объектам будет равна единице:

$$1 = \sum p_1 > \sum p_2 = 1,$$

приходим к противоречию.

Тогда для любого распределения на выборке существует элемент x':

$$\frac{p(x'|\mathbf{w}(x'))}{\sum_x p(\mathbf{w}|w(x))} > p(x').$$

Тогда

$$R(P, x) \geq -\log p(x') + \log p(x'|\mathbf{w}(x')) > \frac{p(x'|\mathbf{w}(x'))}{\sum_x p(x|\mathbf{w}(x))} = R(x).$$

□

Альтернативным методом выбора модели является двусвязный байесовский вывод. На *первом уровне* байесовского вывода находится апостериорное распределение параметров:

$$\hat{\mathbf{w}} = \arg\max \frac{p(\mathbf{X}|\mathbf{w},\mathbf{f})p(\mathbf{w}|\mathbf{f})}{p(\mathbf{X}|\mathbf{f}}),$$

где $p(\mathbf{X}|\mathbf{w},\mathbf{f})$ — правдоподобие выборки при условии модели $\mathbf{f}$ с параметрами $\mathbf{w}$, $p(\mathbf{w}|\mathbf{f})$ — априорное распределение параметров при условии модели $\mathbf{f}$. Оно задается на основе наших предположений о природе выборки и о способах ее порождения. Знаменатель $p(\mathbf{w}|\mathbf{f})p(\mathbf{X}|\mathbf{f})$ называется *evidence* или *обоснованность модели* и определяет, насколько хорошо модель способна описать выборку в среднем по всем возможным значениям параметров.

На *втором уровне* байесовского вывода осуществляется выбор модели на основе обоснованности модели:

$$\mathbf{f} = \arg\max p(\mathbf{w}|\mathbf{f})p(\mathbf{X}|\mathbf{f}) = \int_{\mathbf{w}} p(\mathbf{X}|\mathbf{w},\mathbf{f})p(\mathbf{w}|\mathbf{f})d\mathbf{w}$$

Эта величина также выступает байесовской интерпертацией минимальной длины. Для подтверждения этого факта докажем следующую теорему.

Теорема 5. *Пусть правдоподобие $p(\mathbf{X}|\mathbf{w},\mathbf{f})$ соответствует экспоненциальному семейству распределений, т.е.*

$$p(x|\mathbf{w},\mathbf{f}) = h(x)g(\boldsymbol{\eta})exp\left(\boldsymbol{\eta}\cdot\mathbf{T}(x)\right),$$

где $h, g, \mathbf{T}$ — некоторые функции, $\boldsymbol{\eta}$ — некоторый параметр распределения.

Пусть в качестве априорного распределения выступает распределение Джефри:

$$p(\mathbf{w}|\mathbf{f}) = \frac{\sqrt{I}}{\int_w \sqrt{I(w)}},\ \textit{где } I\textit{—определить матрицы Фишера:}$$

$$I(\mathbf{w}) = det\{\mathsf{E}_w \frac{-\partial^2}{\partial w_i \partial w_j}\log p(\mathbf{X}|\mathbf{w},\mathbf{f})\}_{ij}.$$

Тогда регрет можно аппроксимировать следующим выражением:

$$\frac{k}{2}\log\frac{n}{2\pi} - \log p(\mathbf{X}|\mathbf{w},\mathbf{f}) + \log\sqrt{I(\mathbf{w})}.$$

Доказательство. Воспользуемся аппроксимацией Лапласа для упрощения формулы обоснованности модели. Разложим $\log p(\mathbf{X},\mathbf{w}|\mathbf{f}) = \log p(\mathbf{X}|\mathbf{w},\mathbf{f})p(\mathbf{w}|\mathbf{f})$ в точке локального максимума $\mathbf{w}_0$ в ряд Тейлора:

$$\log p(\mathbf{X},\mathbf{w}|\mathbf{f}) \approx \log p(\mathbf{X},\mathbf{w}_0|\mathbf{f}) - \frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^{\mathsf{T}}\mathbf{A}(\mathbf{w}-\mathbf{w}_0),$$

$$A_{ij} = -\frac{\partial^2}{\partial w_i \partial w_j}\log p(\mathbf{X},\mathbf{w}|\mathbf{f})|_{\mathbf{w}=\mathbf{w}_0}.$$

Полученное распределение представимо в виде ненормированного гауссового распределения. для такой аппроксимации плотности вероятности запишем нормирующий коэффициент:

$$\log p(\mathbf{X},\mathbf{w}|\mathbf{f}) \approx \log p(\mathbf{X}|\mathbf{w}_0,\mathbf{f}) + p(\mathbf{w}_0|\mathbf{f}) - \log\sqrt{\frac{(2\pi)^k}{-}\det\mathbf{A}}.$$

Подставляя в полученную формулу распределение Джеффри получим:

$$p(\mathbf{X},\mathbf{w}|\mathbf{f}) \approx p(\mathbf{X}|\mathbf{w}_0,\mathbf{f}) - \log\sqrt{\frac{(2\pi)^k}{2}}.$$

TODO: там еще информация Фишера + экспоненциальное семейство. Тогда регрет будет равен:

$$R(P,x) \approx \log\sqrt{\frac{(2\pi)^k}{2}}.$$

□

Теорема (без доказательств) 5. *При количество параметров, стремящемся к бесконечности, регрет оптимальной модели отличается от байесовской оценки на константу:*

$$R(P, x) = \frac{k}{2} \log \frac{n}{2\pi} + \log \int_{\mathbf{w}} \sqrt{I(\mathbf{w}} d\mathbf{w} + o(1).$$

TODO: вывод, что эти вещи совпадают.

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. – T. 50. – №. 12. – С. 3265-3290.

[4] Grunwald P. A tutorial introduction to the minimum description length principle //arXiv preprint math/0406077. – 2004.

[5] Штарьков Ю. М. Универсальное последовательное кодирование отдельных сообщений //Проблемы передачи информации. – 1987. – Т. 23. – №. 3. – С. 3-17.