

Занятие четвертое. О непараметрическом оценивании, о том как хорошо оценивать и как оценивать еще лучше, о том, как бутстрэпить без знания распределения

Общие слова

Мы привыкли рассматривать именно параметрическую модель $X_i \sim F_\theta$. С другой стороны, начиная исследование, мы часто не имеем никакой предварительной информации о распределении. В таком случае более подходящей является непараметрическая модель $X_i \sim F$, а оценки мы будем строить для $F(x)$, $\mathbf{E}X_1$ или другого показателя, связанного с моделью. При этом мы сохраняем определения состоятельности, несмещенности и асимптотической нормальности, просто не апеллируем в них к параметризации модели.

Пример 1. Так $\bar{X} = (X_1 + \dots + X_n)/n$ будет несмещенной состоятельной оценкой $\mathbf{E}X_1$ и в непараметрической модели, а при $\mathbf{E}X_1^2 < \infty$ еще и асимптотически нормальной. Аналогично $S_0^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$ будет несмещенной состоятельной оценкой дисперсии, а при $\mathbf{E}X_1^4 < \infty$ асимптотически нормальной.

Как оценивать функции распределения и как с помощью этого строить непараметрические оценки?

Итак, давайте начнем с того, что оценим функцию распределения и плотность нашей выборки X_i . Достаточно хорошей (несмещенной состоятельной и асимптотически нормальной) оценкой функции распределения в конкретной точке x является эмпирическая функция распределения

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I_{X_i \leq x}$$

Вопрос 1. Почему она при каждом x обладает свойствами несмещенности, состоятельности и асимптотической нормальности?

Вопрос 2. Показать, что \hat{F}_n — ОМП для ф.р. в непараметрической модели (считая, что рассматриваемые распределения являются всеми дискретными распределениями с какими-либо значениями и вероятностями).

В R ЭФР может быть легко подсчитана по выборке с помощью команды `ecdf(x)`. Обратите внимания, что при построении `plot` от этой функции, нужно будет написать `plot(Vectorize(ecdf))`.

Теорема Гливленко-Кантелли утверждает, что ЭФР состоятельна как оценка всей функции распределения, даже более, $\hat{F}_n \rightarrow F$ п.н. по равномерной норме.

Существует и явная оценка сверху: неравенство Дворецкого-Кифера-Вольфовица

$$\mathbf{P}(\|\hat{F}_n - F\| > \varepsilon) \leq 2e^{-2n\varepsilon^2}, \quad \varepsilon > 0.$$

Упражнение 1. Построить доверительное множество для ф.р. распределения Коши с помощью неравенства Д.-К.-В.

Из сходимости $\hat{F}_n \rightarrow F$ при п.в. ω вытекает сходимость при тех же ω $f(\hat{F}_n) \rightarrow f(F)$ для любого непрерывного (по равномерной норме) функционала f . Вполне естественно, таким образом, исследуя функционалы $f(F)$, оценивать их $f(\hat{F}_n)$.

Пример 2. Для оценки математического ожидания $\mathbf{E}X_1 = \int_{\mathbb{R}} x dF(x)$ и дисперсии $\mathbf{D}X_1 = \int_{\mathbb{R}} x^2 dF(x) - (\int_{\mathbb{R}} x dF(x))^2$ возьмем

$$\hat{\theta}_1 = \int_{\mathbb{R}} x d\hat{F}_n(x), \quad \hat{\theta}_2 = \int_{\mathbb{R}} x^2 d\hat{F}_n(x) - \hat{\theta}_1^2.$$

Вопрос 3. Почему $\hat{\theta}_1 = \bar{X}$, $\hat{\theta}_2 = S^2$?

Упражнение 2. Построить оценку для асимметрии: $\mathbf{E}(X - \mathbf{E}X)^3 (\mathbf{D}X)^{-3/2}$.

Как оценить плотность?

С той же целью полезно бывает оценить плотность, если мы знаем, что распределение абсолютно-

непрерывно.

Простейшей оценкой плотности является гистограмма, с которой вы, вероятно, и так знакомы. Гистограмма предлагает рассмотреть некоторое разбиение области значений нашей величины $(a_i, a_{i+1}]$, $i = 1, 2, \dots$. Тогда гистограмму на $(a_i, a_{i+1}]$ положим равной $N_i/N = \#\{x_j \in (a_i, a_{i+1}]\}/N$, то есть частоте попадания в полуинтервал наших наблюдений. В целом, гистограмма неплохая оценка, стремящаяся к росту числа наблюдений и уменьшению ширины интервалов разбиения к истинному значению плотности, но она кусочно-постоянна.

Хорошим методом получения непрерывной оценки для плотности является так называемая *ядерная оценка*. Назовем *ядерной* неотрицательную функцию $K(x)$, т.ч. $\int_{\mathbb{R}} K(x)dx = 1$, $\int_{\mathbb{R}} xK(x)dx = 0$, $\int_{\mathbb{R}} x^2K(x)dx > 0$ (то есть $K(x)$ — некоторая плотность с нулевым матожиданием и конечной дисперсией). Назовем *ядерной оценкой* плотности

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_n} K\left(\frac{x - X_i}{h_n}\right),$$

где K — ядерная функция, а h_n — некий параметр, называющийся шириной окна сглаживания. Если f — непрерывна, $h_n \rightarrow 0$, $nh_n \rightarrow \infty$, то $\hat{f}_n(x)$ будет состоятельной для $f(x)$. При этом $\mathbf{E}(\hat{f}_n(x) - f(x))^2$ есть $O(h_n^4) + O\left(\frac{1}{nh_n}\right) + O\left(\frac{1}{n}\right)$. Таким образом, наилучший порядок h есть $O(n^{-1/5})$. Более точная формула для h , дающего минимум квадратичного отклонения

$$h^* = \left(\frac{\int K(x)^2 dx}{n \left(\int x^2 K(x) dx \right)^2 \int (f''(x))^2 dx} \right)^{1/5}.$$

Эта формула не вполне удовлетворительна для применения, поскольку использует неизвестное нам $f''(x)$, но дает представить порядок h^* .

Функция K , как мы видим, не влияет на порядок сходимости. Часто берут в роли $K(x)$ равномерную на $[-1, 1]$ плотность (прямоугольное ядро), стандартную нормальную плотность (гауссово ядро) или $3(1 - x^2)I_{|x| \leq 1}/4$ (ядро Епанчикова).

В R ядерная оценка плотности встроена, например, в стандартный график `qplot` с геометрией `geom=density`, по умолчанию устанавливается гауссовое ядро, но можно поменять его с помощью параметра `kernel`.

Непосредственно ядерную оценку для плотности можно рассчитать с помощью функции `density`. Параметр `kernel` отвечает за вид ядра, параметр `bw` отвечает за ширину окна.

Упражнение 3. Оцените а) стандартную нормальную плотность на выборке размера 100. б) смесь из трех нормальных плотностей $\mathcal{N}(0, 1)$, $\mathcal{N}(6, 1)$, $\mathcal{N}(-3, 1)$ с равными вероятностями. Что происходит при уменьшении ширины окна h ?

Существуют и другие подходы к оценке плотности, например, оценить ее коэффициенты разложения в ряд Фурье. Мы не будем их касаться в рамках нашего семинара.

Поговорим о том, как оценить качество полученной оценки для плотности. Хорошим показателем для нас являлась бы величина

$$\int_{\mathbb{R}} (\hat{f}_n(x) - f(x))^2 dx = \int_{\mathbb{R}} \hat{f}_n(x)^2 dx - 2 \int_{\mathbb{R}} \hat{f}_n(x) f(x) dx + \int_{\mathbb{R}} f(x)^2 dx,$$

которую мы не можем подсчитать, поскольку не знаем $f(x)$.

Оценкой кросс-валидации для оценки плотности \hat{f}_n называют

$$\hat{J}(h) = \int_{\mathbb{R}} \hat{f}_n(x)^2 dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{n,i}(X_i),$$

где $\hat{f}_{n,i}$ — оценка, построенная по выборке с исключенным X_i .

Вопрос 4. Показать, что $\mathbf{E}\hat{J}(h) = \mathbf{E}\left(\int_{\mathbb{R}} \hat{f}_n(x)^2 dx - 2 \int_{\mathbb{R}} \hat{f}_n(x)f(x)dx\right)$.

Тем самым, маленькие значения $\hat{J}(h)$ указывают на то, что $\hat{f}_n(x)$ близко к $f(x)$.

Для ядерной оценки оценку кросс-валидации можно найти по формуле

$$\hat{J}(h) = \frac{1}{hn^2} \sum_i \sum_j K^* \left(\frac{X_i - X_j}{h} \right) + \frac{2}{nh} K(0) + O\left(\frac{1}{n^2}\right).$$

$K^* = K * K - 2K$, $*$ — свертка. С помощью оценки кросс-валидации удобно измерять качество приближения. В частности, с ее помощью можно откалибровать ширину h .

Осуществить выбор h с помощью кросс-валидации в R можно, указав `bw=bw.ucv(x)`.

С помощью оценки для плотности можно строить оценки $\int g(x)\hat{f}_n(x)dx$ для функционалов вида $\int g(x)f(x)dx$. Например, для математического ожидания мы получим оценку

$$\int_{\mathbb{R}} x \hat{f}_n(x) dx = \bar{X}$$

Вопрос 5. Почему верно последнее тождество?

Бутстреп или как прикрывать свои недостатки в непараметрическом случае

Если мы захотим получать несмещенные оценки, то нам, как и прежде, будет полезна процедура бутстреппинга. Однако, раньше мы брали выборку из распределения с оцениваемым параметром, то теперь мы будем брать выборку из распределения с оцениваемой функцией распределения. Если в качестве оценки для функции распределения используется эмпирическая функция распределения, то это равносильно рассмотрению выборок X_1^*, \dots, X_n^* , взятых из нашей выборки с возвращением.

Таким образом, если мы рассматриваем статистику $f(\hat{F}_n)$ в качестве оценки $f(F)$, мы можем брать выборки из распределения \hat{F}_n и на основе этих выборок изучать распределение $f(\hat{F}_n)$, ожидая, что оно близко к распределению $f(F)$. В частности, мы можем исследовать смещение $f(\hat{F}_n)$ по сравнению с $f(F)$, приближая его смещением $f(\hat{F}_n) - f(F)$, $\hat{F}_n - \mathcal{E}\hat{F}_n$ выборки из \hat{F}_n . Аналогичным образом мы можем изучать и другие параметры распределения $f(\hat{F}_n)$, например, дисперсию.

Для бутстрэпа в пакете `bootstrap` есть одноименная функция, хотя реализация этого алгоритма совсем проста и не требует специальных функций.

Упражнение 4. Сгенерировать выборку $R[0, 1]$ размера 50 и оценить дисперсию оценок а) \bar{X} , б) *MED*.

Упражнение 5. Исправить смещение статистики S^2 как оценки дисперсии с помощью `bootstrap` по выборке размера 100 из распределения $\mathcal{N}(0, 1)$.

Предположим, что $\hat{\theta} = f(\hat{F}_n)$ — оценка какого-то параметра распределения $f(F)$. Тогда мы можем построить для нее `pivotal`-интервал тем же методом, что и прежде — сгенерируем m выборок из функции распределения \hat{F}_n (то есть m выборок того же размера с возвращением), отранжировать значение нашей оценки $\hat{\theta}$ на этих выборках и взять в качестве левой границы доверительного интервала $[\alpha m]$ по возрастанию из значений $\hat{\theta}$, а в качестве правой границы — $[(1 - \alpha)m]$.

Можно действовать более хитро. Оценим с помощью непараметрического `bootstrap`-а дисперсию нашей оценки $f(F)$, извлечем из этой оценки корень и получим оценку $S(f)$ для стандартного отклонения нашей оценки. Рассмотрим теперь величину $(f(\hat{F}_n) - f(F))/S(f)$. Построим для нее `bootstrap`-интервал ширины 0.95, откуда получим доверительный интервал для $f(F)$. Этот интервал называется `studentized pivotal`.

Упражнение 6. Испытайте метод `pivotal` для интервалов для а) среднего $R[0, 1]$, б) дисперсии $\exp(1)$. Рассмотренные интервалы не точные, их уровень доверия близок к $1 - \alpha$ с ростом n (хотя и не равен ему). Стьюдентовский интервал имеет более высокую скорость сходимости к $1 - \alpha$, равную $O(1/n)$, обычный `pivotal` интервал имеет скорость $O(1/\sqrt{n})$.

Функции влияния и Дельта-метод в непараметрическом случае

Этот параграф требуется только если вы сдаете на продвинутом уровне

Пусть f — функционал от ф.р. Рассмотрим так называемую производную Гато нашего функционала

$$f'_G(F) = \lim_{p \rightarrow 0} \frac{f((1-p)F + pG) - f(F)}{p},$$

где $G(x)$ — некоторая ф.р. Это аналог обычной производной по направлению, только для случая, когда аргументом у нас служат функции.

Мы будем рассматривать частный случай $G = \delta_x$, где δ_x — ф.р. константы x , то есть $\delta_x(y) = 1$ при $x \leq y$ и 0 при $x > y$. Тогда

$$L_{f,F}(x) = f'_{\delta_x}(F) = \lim_{p \rightarrow 0} (f((1-p)F + p\delta_x) - f(F))/p$$

называют функцией влияния.

Давайте поймем физический смысл функции влияния. Функция $(1-p)F(x) + p\delta_x$ соответствует функции распределения следующей величины: с вероятностью p она равна x , а с вероятностью $1-p$ она равна $X \sim F$. Таким образом, функция влияния предлагает добавить в данные из распределения F небольшой процент p данных тождественно равных x , сравнить, насколько при этом изменился функционал и поделить на p , устремив $p \rightarrow 0$.

Вопрос 6. Пусть $f(F) = \int_{\mathbb{R}} a(x)dF(x)$. Чему равно $L_{f,F}$?

Как мы уже говорили, функция $\hat{F}_n(x)$ является асимптотически нормальной оценкой функции $F(x)$ при каждом x . А вдруг нам повезет и функционалы $f(\hat{F}_n)$ тоже будут асимптотически нормальными оценками для $f(F)$? В параметрическом случае у нас функции от оценок оказывались асимптотически нормальными в силу дельта-метода при дифференцируемых f . В непараметрическом случае верна аналогичная теорема: если f дифференцируема по Гато

$$\sqrt{n} \frac{f(\hat{F}_n) - f(F)}{\tau} \rightarrow Z \sim \mathcal{N}(0, 1),$$

где $\tau^2 = \int_{\mathbb{R}} L_{f,F}(x)^2 dF(x)$. Кроме того,

$$\sqrt{n} \frac{f(\hat{F}_n) - f(F)}{\hat{\tau}} \rightarrow Z \sim \mathcal{N}(0, 1),$$

где $\hat{\tau} = \frac{1}{n} \sum_{i=1}^n L_{f,\hat{F}_n}^2(X_i)$. Это некоторый аналог Дельта-метода — мы видим, как меняется дисперсия величины при применении функционала f .

Нетрудно заметить, что этот метод упрощает построение Стюдентовских интервалов, поскольку позволяет сократить расходы на подсчет дисперсии.

Упражнение 7. Рассматривая $f(F) = F(1/3) - F(1/4)$, найти для $R[0, 1]$ с помощью функции влияния доверительный интервал для $f(F)$ а) обычный б) бутстрэповский стьюдентовский.

Размерность Валника-Червоненкиса

Этот параграф требуется только если вы сдаете на продвинутом уровне

Зачастую нам требуется оценить не функцию распределения, а само распределение $\mathbf{P}(X \in A)$. Естественно делать это с помощью эмпирического распределения

$$\mathbf{P}_X(A) := \mathbf{P}(X \in A) \approx \hat{\mathbf{P}}_n(A) := \frac{1}{n} \sum_{i=1}^n I_{X_i \in A}.$$

Погрешность приближения ЭФР мы умеем оценивать с помощью неравенства Дворецкого-Кифера-Вольфовица, а можно ли оценить погрешность приближения эмпирическим распределением теоретиче-

ского? При фиксированном A можно показать, что

$$\mathbf{P}(|\hat{\mathbf{P}}_n(A) - \mathbf{P}_X(A)| > \varepsilon) \leq 2e^{-2n\varepsilon^2}.$$

Это неравенство основано на том, что величина $n\hat{\mathbf{P}}_n(A)$ имеет биномиальное распределение $\text{Binom}(n, \mathbf{P}_X(A))$ и является частным случаем так называемого неравенства Хёфдинга.

Мы бы хотели получить оценку сверху по всем A из некоторого множества \mathcal{A} , т.е. оценить $\mathbf{P}(\sup_{A \in \mathcal{A}} |\hat{\mathbf{P}}_n(A) - \mathbf{P}_X(A)| > \varepsilon)$.

Пример 3. При одноточечном множестве \mathcal{A} у нас уже есть оценка Хёфдинга. При k -точечном множестве мы можем записать

$$\mathbf{P}(\sup_{A \in \mathcal{A}} |\hat{\mathbf{P}}_n(A) - \mathbf{P}_X(A)| > \varepsilon) \leq k \sup_A \mathbf{P}(|\hat{\mathbf{P}}_n(A) - \mathbf{P}_X(A)| > \varepsilon) \leq 2ke^{-2n\varepsilon^2}.$$

При $\mathcal{A} = \mathcal{B}(\mathbb{R})$ (то есть когда мы рассматриваем все A разом) для любого непрерывного распределения \mathbf{P}_X при любом ω найдется множество $A = \{X_1(\omega), \dots, X_n(\omega)\}$, для которого $\hat{\mathbf{P}}_n(A) = 1$, а $\mathbf{P}_X(A) = 0$. Соответственно, при каждом ω найдется такое множество A , что $\hat{\mathbf{P}}_n(A) - \mathbf{P}_X(A) = 1$. Значит наша вероятность будет равна 0 при любом $\varepsilon < 1$.

Тем не менее, для каких-то не очень больших множеств \mathcal{A} мы надеемся получить хорошую оценку сверху. Для этого нам понадобится понятие VC-размерности.

Рассмотрим множество $R = \{x_1, \dots, x_m\}$. Рассмотрим какую часть его подмножеств мы можем получить, пересекая R с $A \in \mathcal{A}$. Назовем их количество $N_{\mathcal{A}}(R)$.

Пример 4. Если $\mathcal{A} = \mathcal{B}(\mathbb{R}^n)$, то $N_{\mathcal{A}}(R)$ будет равно 2^m , потому что пересечением с борелевскими множествами можно получить любое подмножество R .

Если \mathcal{A} состоит из k элементов, то $N_{\mathcal{A}}$ не больше k по определению. При разных R эта величина будет принимать различные значения от 1 до k .

Если $\mathcal{A} = \{(-\infty, x], x \in \mathbb{R}\}$ — множество всех лучей, то $N_{\mathcal{A}}(R) = m + 1$, потому что мы можем получить либо пустое множество, либо множество из 1 самого маленького элемента R , либо из двух наименьших и так далее.

Если $\mathcal{A} = \{(-\infty, x], x \in \mathbb{Z}\}$, то $N_{\mathcal{A}}(R)$ будет зависеть от множества R . Скажем, если все элементы R лежат от 0 до 1, то $N_{\mathcal{A}}(R) = 2$, поскольку можно будет получить только пустое множество и R , а если они принимают значения $1, 2, \dots, m$, то мы сможем получить те же $m + 1$ элемент что и прежде.

Величину $s(\mathcal{A}, m) = \sup_R N_{\mathcal{A}}(R)$ называют shatter coefficient. Оказывается, что справедливо следующее неравенство (Вапник, Червоненкис, 1971):

$$\mathbf{P}(\sup_{A \in \mathcal{A}} |\hat{\mathbf{P}}_n(A) - \mathbf{P}_X(A)| > \varepsilon) \leq 8s(\mathcal{A}, n)e^{-\frac{n\varepsilon^2}{32}}.$$

Пример 5. Если $\mathcal{A} = \mathcal{B}(\mathbb{R})$, то неравенство приобретает вид

$$\mathbf{P}(\sup_{A \in \mathcal{A}} |\hat{\mathbf{P}}_n(A) - \mathbf{P}_X(A)| > \varepsilon) \leq 2^{n+3}e^{-\frac{n\varepsilon^2}{32}}$$

Это малоинтересная оценка, поскольку она больше единицы.

Пример 6. При \mathcal{A} , состоящем из k множеств A неравенство приобретает вид

$$\mathbf{P}(\sup_{A \in \mathcal{A}} |\hat{\mathbf{P}}_n(A) - \mathbf{P}_X(A)| > \varepsilon) \leq 8ke^{-\frac{n\varepsilon^2}{32}},$$

Это ухудшенная версия неравенства Хёфдинга.

Пример 7. При \mathcal{A} , состоящем из множества $(-\infty, x]$ неравенство приобретает вид

$$\mathbf{P}(\sup |\hat{F}_n(x) - F_X(x)| > \varepsilon) \leq 8(n+1)e^{-\frac{n\varepsilon^2}{32}},$$

Это ухудшенная версия неравенства Дворецкого-Кифера-Вольфовица.

В отличие от неравенства ДКВ неравенство Вапника-Червоненкиса пригодно и для других случайных элементов, а не только случайных величин — случайных векторов, процессов и других.

Давайте получим более простой метод оценки, чем прямой подсчет $s(\mathcal{A}, n)$. Назовем размерностью Вапника-Червоненкиса $\mathcal{A} \subset \mathcal{B}(\mathbb{R})$ а) ∞ , если $s(\mathcal{A}, n) = 2^n$ при всех n , б) $\max\{k : s(\mathcal{A}, k) = 2^k\}$ иначе.

Будем обозначать размерность Вапника-Червоненкиса $VC(\mathcal{A})$. Легко понять, что если $s(\mathcal{A}, k) < 2^k$ при каком-то k , то $s(\mathcal{A}, l) < 2^l$ при всех $l \geq k$, поэтому достаточно найти первый номер k при котором неравенство нарушается.

Иначе говоря, размерность Вапника-Червоненкиса системы \mathcal{A} — это самое большое k при котором для какого-то k -точечного множества можно, пересекая его с элементами \mathcal{A} , получить все его подмножества.

Пример 8.

- Для $\mathcal{A} = \mathcal{B}(\mathbb{R}^n)$ мы для любого n -точечного множества B_n можем получить пересечением с \mathcal{A} все подмножества B_n . Значит, в силу а) имеем $VC(\mathcal{A}) = \infty$.
- Для \mathcal{A} , состоящего из $(-\infty, x]$ при всех x , $VC(\mathcal{A}) = 1$. Действительно, для любого двухточечного множества $\{a, b\}$, $a < b$, мы не сможем получить точку b пересечением нашего множества с каким-либо лучом $(-\infty, x]$.
- Для \mathcal{A} , состоящего из всех отрезков $[x, y]$ $VC(\mathcal{A}) = 2$. Действительно, для любого множества R из трех элементов $R = \{a, b, c\}$, $a < b < c$ нельзя получить множество $\{a, c\}$ пересечением $\{a, b, c\}$ с каким-либо отрезком. С другой стороны для двухточечного множества $\{a, b\}$ можно получить такими пересечениями и \emptyset , и $\{a\}$, и $\{b\}$ и $\{a, b\}$.

- Пусть \mathcal{A} — множество всех полуплоскостей на плоскости. Тогда $VC(\mathcal{A}) = 3$.

Докажем это. Если 3-точечное множество R состоит из точек, не лежащих на одной прямой, то можно отделить любую из этих точек от 2 других прямой. Тем самым можно получить в пересечении R с некоторой полуплоскостью любое подмножество из 1 и 2 точек. Получить \emptyset и R также не представляет труда.

С другой стороны, никакое 4-точечное множество R не имеет $N_{\mathcal{A}}(R) = 2^4$. Убедиться в этом нетрудно в каждой конфигурации:

Если они образуют выпуклый четырехугольник ABCD, то не существует полуплоскости, пересекающей с ним по A,C, но не пересекающей по B,D. Если четырехугольник невыпуклый и A лежит внутри треугольника BCD, то B,C,D не могут лежать в этом пересечении без A. И, наконец, если 3 точки лежат на одной прямой, то нельзя получить пересечение с полуплоскостью две крайних из них, не получив в том же пересечении среднюю. Поэтому $VC(\mathcal{A}) = 3$.

Вопрос 7. А чему равна VC-размерность множества всех прямоугольников с осями параллельными осям координат?

Вопрос 8. Тот же вопрос для параллелипипедов в \mathbb{R}^d

С помощью VC-размерности можно оценить $s(\mathcal{A}, n)$:

$$s(\mathcal{A}, n) \leq n^{VC(\mathcal{A})} + 1.$$

При этом подсчет размерности Вапника-Червоненкиса значительно более прост, чем подсчет $s(\mathcal{A}, n)$.

В этот момент у читателя, вероятно, возник вопрос — а зачем все эти сложности? У нас есть неравенство Хеффдинга, позволяющее нам оценить вероятность попадания в любое множество, зачем же нам возиться с супремумами?

Пример 9. Представим, что мы хотим построить 95% доверительный параллелипипед вида $[a_1, b_1] \times [a_2, b_2] \times \dots \times [a_d, b_d]$ для X_{n+1} на основе выборки $X_1, \dots, X_n \in \mathbb{R}^d$. Допустим, мы выберем какой-то параллелипипед I (зависящий от X_1, \dots, X_n), такой что в него попали k из наших X_i . Тогда $\hat{P}_n(I) = \frac{k}{n}$. Оценим максимальное возможное значение $\mathbf{P}(X_{n+1} \notin I(X_1, \dots, X_n))$:

$$\mathbf{P}(X_{n+1} \notin I(X_1, \dots, X_n)) = \mathbf{E}(\mathbf{P}(X_{n+1} \notin I(X_1, \dots, X_n) | X_1, \dots, X_n))$$

Это математическое ожидание разбивается на две части:

а) $X_1, \dots, X_n : \mathbf{P}(X_{n+1} \notin I(X_1, \dots, X_n) | X_1, \dots, X_n) \leq 1 - k/n + \varepsilon,$

б) $X_1, \dots, X_n : \mathbf{P}(X_{n+1} \notin I(X_1, \dots, X_n) | X_1, \dots, X_n) > 1 - k/n + \varepsilon.$

В случае а) математическое ожидание не превосходит $1 - k/n + \varepsilon$, в случае б) не превосходит

$$\mathbf{P}(\exists \tilde{I} : |\hat{P}_n(\tilde{I}) - \mathbf{P}(X \in \tilde{I})| > \varepsilon) = \mathbf{P}(\exists \tilde{I} : |\hat{P}_n(\tilde{I}) - \mathbf{P}(X \in \tilde{I})| > \varepsilon),$$

где \tilde{I} — некоторый параллелепипед. Неравенство Вапника-Червоненкиса позволяет нам дать нужную оценку для последнего выражения:

$$8(n^v + 1)e^{-\frac{n\varepsilon^2}{32}}$$

для последней вероятности, где v — VC-размерность множества всех параллелепипедов, равная $2d$ в силу вопроса выше. То есть вероятность непопадания в I не больше

$$1 - \frac{k}{n} + \varepsilon + 8(n^v + 1)e^{-\frac{n\varepsilon^2}{32}}$$

и если ε достаточно мало, n достаточно велико, а k/n достаточно близко к 1, то мы можем быть сделана сколь угодно близкой к 1.

Оценки, полученные Вапником и Червоненкисом, были позже значительно улучшены. Так Devroye (1982) показал, что правую часть можно заменить на

$$\mathbf{P}(\sup_{A \in \mathcal{A}} |\hat{P}_n(A) - \mathbf{P}_X(A)| > \varepsilon) \leq 4e^{4\varepsilon(1+\varepsilon)} s(\mathcal{A}, n) e^{-2n\varepsilon^2}.$$

Наиболее удачная из известных мне оценок принадлежит Lugosi (1995)

$$\mathbf{P}(\sup_{A \in \mathcal{A}} |\hat{P}_n(A) - \mathbf{P}_X(A)| > \varepsilon) \leq 4e(v+1) \left(\frac{32e^5 n^2 \varepsilon^3}{v^2} \right)^v e^{-2n\varepsilon^2}$$

и справедлива при $n\varepsilon^2 > v/2$, где $v = VC(\mathcal{A})$. Стоит отметить, что есть и асимптотически более удачное неравенство (Talagrand, 1994), в котором, увы, не указана явно константа.

Упражнение 8. Построить непараметрический доверительный прямоугольник вероятности не менее 95% на основе X_1, \dots, X_n а) для выборки из независимых $\mathcal{N}(0, 1)$ б) равномерной на единичном квадрате выборки при $n = 100000$. Насколько часто в него попадает выборка?

Конечно, указанный подход а) работает только при больших выборках (порядка 10^5) б) дает только нижнюю и верхнюю оценки уровня доверия. Зато он позволяет построить доверительное множество любой формы, лишь бы мы могли подсчитать в этом случае VC-размерность. Кроме того, мы получаем оценки для вероятностей попадания во все множества такой формы разом.

Ответы на вопросы

1. Несмещенность вытекает из линейности математического ожидания

$$\mathbf{E} \frac{1}{n} \sum_{i=1}^n I_{X_i \leq x} = \frac{1}{n} \sum_{i=1}^n \mathbf{E} I_{X_i \leq x} = \mathbf{P}(X_1 \leq x) = F_X(x).$$

Состоятельность — прямое следствие ЗБЧ, а асимптотическая нормальность — ЦПТ, примененных к $I_{X_i \leq x}$.

2. Если значения нашей случайной величины y_1, \dots, y_m с вероятностями p_1, \dots, p_m , а выборка x_1, \dots, x_n , то

а) если какой-то из x_i не встречается среди y_i , то правдоподобие 0.

б) иначе правдоподобие принимает вид $L = p_1^{N_1} \dots p_m^{N_m}$, где N_1, \dots, N_m — число x_i , равных y_1, \dots, y_m соот-

ветственно. Но тогда будем рассматривать параметры p_1, \dots, p_{m-1} (и $p_m = 1 - p_1 - \dots - p_{m-1}$):

$$\ln L(p_1, \dots, p_{m-1}) = \sum_{i=1}^{m-1} N_i \ln p_i + N_m \ln(1 - p_1 - \dots - p_{m-1}), \quad \frac{\partial}{\partial p_i} \ln L = \frac{N_i}{p_i} - \frac{N_m}{p_m},$$

То, что найденная точка максимум, можно показать напрямую или рассматривая пару p_i, p_j , фиксируя их сумму.

3. Если x_1, \dots, x_n (значения элементов выборки) различны, то $\hat{F}_n(x)$ — ф.р. дискретной случайной величины \hat{X} со значениями x_1, \dots, x_n и вероятностями $1/n$. Тогда

$$\mathbf{E}\hat{X} = x_1 \frac{1}{n} + x_2 \frac{1}{n} + \dots + x_n \frac{1}{n} = \bar{x}, \quad \mathbf{D}\hat{X} = (x_1 - \bar{x})^2 \frac{1}{n} + \dots + (x_n - \bar{x})^2 \frac{1}{n} = S^2.$$

Если же какие-то значения совпадают, то слагаемые, соответствующие одинаковых x_i , объединятся.

4. Первое слагаемое в обоих выражениях совпадает. Рассмотрим матожидание левой части второго из них

$$\mathbf{E} \frac{2}{n} \sum_{i=1}^n \hat{f}_{n,i}(X_i) = \frac{2}{n} \sum_{i=1}^n \mathbf{E} \hat{f}_{n,i}(X_i).$$

В силу линейности матожидания

$$\mathbf{E} \hat{f}_{n,i}(X_i) = \frac{1}{n-1} \sum_{j \neq i} \frac{1}{h_n} \mathbf{E} K \left(\frac{X_i - X_j}{h_n} \right) = \frac{1}{h_n} \int_{\mathbb{R}^2} K \left(\frac{x-y}{h_n} \right) f(x) f(y) dy dx.$$

С другой стороны,

$$\mathbf{E} \int_{\mathbb{R}} \hat{f}_n(x) f(x) dx = \int_{\mathbb{R}} \frac{1}{n} \frac{1}{h_n} \sum_{j=1}^n \mathbf{E} K \left(\frac{x - X_j}{h_n} \right) dx = \frac{1}{h_n} \int_{\mathbb{R}} K \left(\frac{x-y}{h_n} \right) f(y) f(x) dy dx.$$

Подставляя полученные выражения в правую и левую части искомого тождества, получаем требуемое.

5.

$$\int_{\mathbb{R}} x \hat{f}_n(x) dx = \frac{1}{n} \frac{1}{h_n} \sum_{i=1}^n \int_{\mathbb{R}} x K \left(\frac{x - X_i}{h_n} \right) dx = \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}} (y \cdot h_n + X_i) K(y) dy = \frac{1}{n} \sum_{i=1}^n X_i,$$

где в последнем тождестве мы воспользовались тем, что $\int_{\mathbb{R}} y K(y) dy = 0$, $\int_{\mathbb{R}} K(y) dy = 1$.

6.

$$\int_{\mathbb{R}} a(y) d((1-p)F_X(y) + p\delta_x(y)) = pa(x) + (1-p) \int_{\mathbb{R}} a(y) dF_X(y),$$

поскольку

$$\int_{\mathbb{R}} a(y) d\delta_x(y) = \mathbf{E}a(x) = a(x).$$

Следовательно,

$$L_{f,F}(x) = a(x) - \int_{\mathbb{R}} a(y) dF_X(y).$$

7. Рассмотрим произвольные 5 точек. Выберем среди них одну из самых левых, одну из самых правых, одну из самых верхних, одну из самых нижних. Тогда не найдется прямоугольника, содержащего их, но содержащего пятую точку.

Для 4 точек $(1,0)$, $(0,1)$, $(-1,0)$, $(0,-1)$ можно получить пересечением с прямоугольниками любой набор. Значит $VC(\mathcal{A}) = 4$.

8. Ответ 2d получается совершенно аналогично.