4 Занятие четвертое. Факультатив: Непараметрический дельта-метод и VC-размерность

4.1 Непараметрический дельта-метод

4.1.1 Функция влияния и производная Гато

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

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

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

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

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 $p \to 0$.

Вопрос 1. Пусть $f(F) = \int_{\mathbb{R}} a(x)dF(x)$. Чему равно $L_{f,F}$?
Нам понадобится более общее понятие дифференцируемости по Адамару. Пусть $\mathcal{D} = \{F - G\}$, где F, G — функции распределения.

Определение 1. Функционал f дифференцируем по Адамару, если производная Гато это линейный непрерывный функционал направления и для любых $D_n \in \mathcal{D}$, $\varepsilon_n \to 0$ выполнено соотношение

$$\frac{f(F + \varepsilon_n D_n) - f(F)}{\varepsilon_n} - f'_{D_n}(F) \to 0, \ n \to \infty.$$

Это что-то в духе равномерности предела Гато по всем направлениям.

Вопрос 2. Правда ли функционал $\int_{\mathbb{R}} a(x)dF(x)$ является дифференцируемым по Адамару для любой F, для которой он конечен?

4.1.2 Непараметрический дельта-метод

Как мы уже говорили, функция $\widehat{F}_n(x)$ является асимптотически нормальной оценкой функции $F(x)$ при каждом x. А вдруг нам повезет и функционалы $f(\widehat{F}_n)$ 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:

Теорема 1. *Если f дифференцируем по Адамару, то*

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

где $\tau^2 = \int_{\mathbb{R}} L_{f,F}(x)^2 dF(x)$. Более того, если $\tau(F)$ непрерывный функционал, то

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

где $\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n} L^2_{f,\widehat{F}_n}(X_i)$.

Это непараметрический Дельта-метод — мы видим, что при применении функционала f, асимптотическая нормальность сохраняется и можем пересчитать дисперсию.
Дельта-метод позволяет доказывать асимптотическую нормальность самых разнообразных функционалов от ЭФР, что, в свою очередь, позволяет строить на их основе доверительные интервалы и проверять гипотезы.

Задача 1. Получить следующую теорему: если распределение имеет непрерывную плотность и $p(x_{1/2}) > 0$, где $x_{1/2}$ — корень уравнения $F(x_{1/2}) = 1/2$, то

$$\sqrt{n}(X_{([(n+1)/2])} - x_{1/2}) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{4p(x_{1/2})^2}\right).$$

Здесь $X_{([(n+1)/2])}$ — это $[(n+1)/2]$-й из X, упорядоченных по возрастанию.

4.2 Оценки для эмпирического распределения и размерность Вапника-Червоненкиса

4.2.1 Эмпирическое распределение и его близость к теоретическому

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

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

Погрешность приближения ЭФР мы умеем оценивать с помощью неравенства Дворецкого-Кифера-Вольфовица, а можно ли оценить погрешность приближения эмирическим распределением теоретического? При фиксированном A можно показать, что

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

Это неравенство основано на том, что величина $n\widehat{\mathbf{P}}_n(A)$ имеет биномиальное распределение $Binom(n, \mathbf{P}_X(A))$ 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 $\mathbf{P}(\sup_{A\in\mathcal{A}} |\widehat{\mathbf{P}}_n(A) - \mathbf{P}_X(A)| > \varepsilon)$.

Пример 1. При одноточечном множестве $\mathcal{A}$ у нас уже есть оценка Хёфдинга.
При k-точечном множестве мы можем записать

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

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 $A = \{X_1(\omega), ..., X_n(\omega)\}$, для которого $\widehat{\mathbf{P}}_n(A) = 1$, а $\mathbf{P}_X(A) = 0$. Соответственно, при каждом ω найдется такое множество A, что $\widehat{\mathbf{P}}_n(A) - \mathbf{P}_X(A) = 1$. Значит наша вероятность будет равна 0 при любом $\varepsilon < 1$.

4.2.2 Неравенство Вапника-Червоненкиса

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

Пример 2. 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,...,m, то мы сможем получить те же $m+1$ элемент что и прежде.

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

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

Пример 3. Если $\mathcal{A} = \mathcal{B}(\mathbb{R})$, то неравенство приобретает вид

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

Это малоинтересная оценка, поскольку она больше единицы.

Пример 4. При $\mathcal{A}$, состоящем из k множеств A неравенство приобретает вид

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

Это ухудшенная версия неравенства Хеффдинга.

Пример 5. При $\mathcal{A}$, состоящем из множества $(-\infty, x]$ неравенство приобретает вид

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

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.

4.2.3 VC-размерность

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 6.

- 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, пересекающейся с ним по A,C, но не пересекающейся 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 3. А чему равна VC-размерность множества всех прямоугольников с осями параллельными осям координат?

Вопрос 4. Тот же вопрос для параллелипипедов в $\mathbb{R}^d$.

4.2.4 Неравенство Вапника-Червоненкиса в терминах VC-размерность

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

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

Это математическое ожидание разбивается на две части:
а) $X_1, ..., X_n : \mathbf{P}(X_{n+1} \notin I(X_1, ..., X_n) | X_1, ..., X_n) \leq 1 - k/n + \varepsilon$,
б) $X_1, ..., X_n : \mathbf{P}(X_{n+1} \notin I(X_1, ..., X_n) | X_1, ..., X_n) > 1 - k/n + \varepsilon$.
В случае а) математическое ожидание не превосходит $1 - k/n + \varepsilon$, в случае б) не превосходит

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

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.

4.2.5 Улучшения неравенства

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

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

Наиболее удачная из известных мне оценок принадлежит Lugosi (1995)

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

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.

Задача 2. Построить непараметрический доверительный прямоугольник вероятности не менее 95% на основе $X_1,...,X_n$ 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.

5 Ответы на вопросы

Ответ 1. Поскольку функционал линеен, то

$$\frac{f(F+\varepsilon D)-f(F)}{\varepsilon}=f(D)$$

Следовательно, производная Гато есть

$$\int_{\mathbb{R}}a(x)dD(x).$$

Ответ 2. Поскольку рассматриваемый функционал линеен и $f(F+\varepsilon D)=f(F)+\varepsilon f(D)$, то выражение, фигурирующее в определении дифференцируемости по Адамару, тождественно равно 0. Однако, сама производная не является непрерывным функционалом.если не потребовать, чтобы $a(x)$ была непрерывной ограниченной функцией.

Ответ 3. 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 $VC(\mathcal{A})=4$.

Ответ 4. Ответ 2d получается совершенно аналогично.