Chaining и энтропийное неравенство Dudley

Голубев Г.К.
CNRS, Université de Provence
Институт Проблем Передачи Информации им. Харкевича
PreMoLab

Лекция посвящена двум темам:

1. Chaining, который изобрел А. Н. Колмогоров.
2. Статистическая теория автоматического обучения Вапника-Червоненкиса.

Цель лекции - показать, что chaining очень простой и мощный прием, который позволяет получать рекордные результаты в теории автоматического обучения.

Задача, которую мы сегодня обсуждаем исключительно просто формулируется: пусть имеется набор случайных величин

$$\{\xi_t,\ t \in T\}, \quad \xi_t \in \mathbf{R}^1,$$

T — конечное множество.

Наша задача — вычислить

$$\mathbf{E} \max_{t \in T} \xi_t.$$

Попробуем понять, насколько сложна эта задача. Именно, попробуем ответить на вопрос: насколько эта задача сложнее вычисления

$$\mathbf{E}\sum_{t\in T}\xi'_t = \sum_{t\in T}\mathbf{E}\xi'_t,$$

где ξ'_t некоторый набор других с. в.

ЛЕММА. *Пусть ξ_t — гауссовские случайные величины с нулевым средним и $\mathbf{E}\xi_t^2 = \sigma_t^2 \le \sigma$. Тогда*

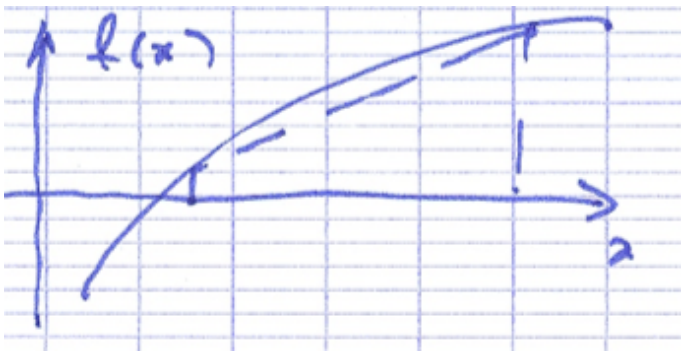
$$\mathbf{E}\max_{t\in T}\xi_t \le \sqrt{2\sigma^2\log(n)},$$

здесь $n = \#T$ — число элементов T.

Доказательство. Для любого $\lambda > 0$

$$\max_{t\in T} \xi_t = \lambda^{-1} \log \Big[\max_{t\in T} \mathrm{e}^{\lambda \xi_t} \Big] \leq \lambda^{-1} \log \Big[\sum_{t\in T} \mathrm{e}^{\lambda \xi_t} \Big].$$

Воспользуемся далее неравенством Йенсена: если $f(\cdot)$ вогнутая функция, то $\mathbf{E} f(\xi) \leq f(\mathbf{E}\xi)$.



Имеем

$$\mathbf{E} \max_{t\in T} \xi_t \leq \lambda^{-1} \log \Big[\sum_{t\in T} \mathbf{E} \exp(\lambda \xi_t) \Big]$$

$$\leq \lambda^{-1} \log \left[n \exp(\lambda^2 \sigma^2 / 2) \right] = \frac{\log(n)}{\lambda} + \frac{\sigma^2 \lambda}{2}.$$

Поскольку это неравенство справедливо для любого λ, то чтобы улучшить верхнюю границу, минимизируем ее по λ. Находим

$$-\frac{\log(n)}{\lambda^2} + \frac{\sigma^2}{2} = 0, \quad \lambda = \sqrt{\frac{2\log(n)}{\sigma^2}}.$$

Поэтому

$$\mathbf{E} \max_{t \in T} \xi_t \leq \log(n) \sqrt{\frac{\sigma^2}{2\log(n)}} + \sqrt{\sigma^2 \log(n)/2} = \sqrt{2\log(n)\sigma^2}.$$

▲

Основная идея доказательства: если случайные величины $\xi_t' > 0$ имеют тяжелые хвосты, то

$$\max_{t \in T} \xi_t' \asymp \sum_{t \in T} \xi_t'.$$

Заметим также, что на самом деле не важно, что ξ_t это гауссовские с. в. Мы использовали только то, что

$$\mathbf{E}\exp(\xi_t\lambda) \leq \exp(\lambda^2\sigma^2/2).$$

Величины, удовлетворяющие этому неравенству часто называют суб-гауссовскими.
Заметим, что сумма независимых суб-гауссовских величин также суб-гауссовская величина.
Помимо гауссовских случайных величин, какие еще величины являются суб-гауссовскими? Кажется, что и ограниченные случайные величины должны также быть суб-гауссовскими.

ЛЕММА. (Хёфдинг) *Пусть ξ — с. в. с нулевым средним* $\mathbf{E}\xi = 0$ *и* $\xi \in [a, b]$. *Тогда*

$$\mathbf{E}\exp(\lambda\xi) \leq \exp\left\{\frac{\lambda^2(b-a)^2}{8}\right\}.$$

Доказательство. Рассмотрим функцию

$$\phi(\lambda) = \log\left[\int_a^b e^{\lambda x} p(x)\,dx\right] = \log\left[\mathbf{E}\exp(\lambda\xi)\right]$$

и заметим, что

$$\phi'(\lambda) = \int_a^b x \frac{e^{\lambda x} p(x)}{\int_a^b e^{\lambda u} p(u)\,du}\,dx.$$

$$\phi''(\lambda) = \int_a^b x^2 \frac{e^{\lambda x} p(x)}{\int_a^b e^{\lambda u} p(u)\, du}\, dx$$
$$- \left\{ \int_a^b x \frac{e^{\lambda x} p(x)}{\int_a^b e^{\lambda u} p(u)\, du}\, dx \right\}^2 = \mathrm{Var}(Z),$$

где Z — с. в. с плотностью

$$\frac{e^{\lambda x} p(x)}{\int_a^b e^{\lambda u} p(u)\, du}.$$

Заметим далее, что $Z \in [a, b]$ и что

$$\left[Z - \frac{a+b}{2} \right]^2 \leq \left(\frac{b-a}{2} \right)^2.$$

Поэтому

$$\mathrm{Var}(Z) = \min_x \mathbf{E}\big[Z - x\big]^2 \le \left(\frac{b-a}{2}\right)^2.$$

Чтобы завершить доказательство, проинтегрируем неравенство

$$\phi''(\lambda) \le \left(\frac{b-a}{2}\right)^2,$$

учитывая, что $\phi(0) = 0$ и $\phi'(0) = 0$. Тогда находим

$$\phi(\lambda) \le \frac{\lambda^2}{2}\left(\frac{b-a}{2}\right)^2. \quad \blacktriangle$$

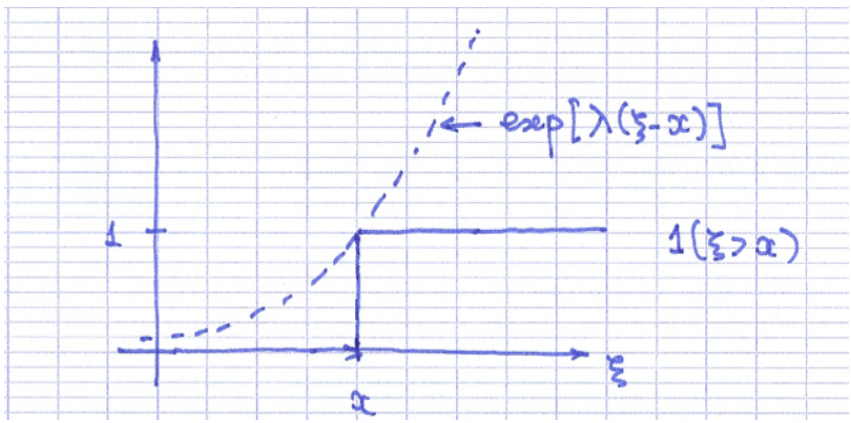
ТЕОРЕМА. (Хёфдинг) *Пусть ξ_t, $t \in T$ — независимые с. в. и такие, что $\xi_t \in [a, b]$. Тогда*

$$\mathbf{P}\left\{\left|\frac{1}{n}\sum_{t\in T}\left(\xi_t - \mathbf{E}\xi_t\right)\right| \geq x\right\} \leq 2\exp\left\{-\frac{2nx^2}{(b-a)^2}\right\}.$$

Доказательство. Воспользуемся неравеством Чернова

$$\mathbf{P}\{\xi > x\} \leq \exp(-\lambda x)\mathbf{E}\exp(\lambda\xi),$$

которое справеливо для любой с. в. ξ и любого $\lambda > 0$.



Положим

$$\xi = \frac{1}{n} \sum_{t \in T} \left[\xi_t - \mathbf{E} \xi_t \right].$$

Тогда из неравенства Чернова и леммы Хёфдинга находим

$$\mathbf{P}\{\xi > x\} \leq \exp \left\{ \min_{\lambda} \left[-\lambda x + \frac{\lambda^2}{8} \frac{(b-a)^2}{n} \right] \right\}.$$

Оптимальное λ:

$$\lambda \frac{(b-a)^2}{4n} = x, \quad \lambda = \frac{4nx}{(b-a)^2}.$$

Поэтому

$$\mathbf{P}\{\xi > x\} \leq \exp\left\{ - \frac{2nx^2}{(b-a)^2} \right\}.$$

Аналогичное неравенство справедливо для с. в. $-\xi$.
Поэтому

$$\mathbf{P}\{|\xi| > x\} \leq 2\exp\left\{ - \frac{2nx^2}{(b-a)^2} \right\}. \quad \blacktriangle$$

Принципиально важный метод в этом доказательстве — неравенство Чернова.

$$\mathbf{P}\{\xi > x\} \leq \exp\left\{ \min_{\lambda} \left[- \lambda x + \log \mathbf{E} e^{\lambda \xi} \right] \right\}$$

Наш первый метод получения верхних границ для $\mathbf{E} \max_{t \in T} \xi_t$ дает хорошие результаты только в том случае, когда ξ_t близки к независимым с. в.

Задача, которую мы рассмотрим далее — это поиск метода, который позволял бы получать разумные результаты для зависимых с. в.

До сих пор мы имели дело с суб-гауссовскими с. в. Теперь рассмотрим несколько более общие семейства с. в., а именно, величины с ограниченной нормой Орлича.

Определение. *Функция $\psi(x)$: $\mathbf{R}^+ \to \mathbf{R}^+$ называется функцией Юнга, если $\psi(0) = 0$, $\psi(x)$ не убывает и выпукла.*

Определение. *Норма Орлича с. в. Z относительно функции Юнга ψ определяется как*

$$\|Z\|_\psi = \inf\left\{c > 0 : \mathbf{E}\psi\left(\frac{|Z|}{c}\right) \leq 1\right\}.$$

Примеры:

для суб-гауссовских с. в. $\psi(x) = \exp(x^2/2) - 1$;

для суб-экспоненциальных с. в. $\psi(x) = \exp(x) - 1$;

для с. в. с ограниченным p-ым моментом $\psi(x) = x^p$.

ЛЕММА. *Пусть* ξ_t, $t \in T$, $(\#T = n)$ *такова, что* $\|\xi_t\|_\psi \le 1$. *Тогда*

$$\mathbf{E} \max_{t \in T} |\xi_t| \le \psi^{-1}(n).$$

Здесь $\psi^{-1}(\cdot)$ *— функция, обратная к* $\psi(\cdot)$.
Доказательство аналогично доказательству первой леммы в лекции. Имеем

$$\max_{t \in T} |\xi_t| \le \psi^{-1}\left(\max_{t \in T} \psi\big(|\xi_t|\big) \right) \le \psi^{-1}\left(\sum_{t \in T} \psi\big(|\xi_t|\big) \right).$$

Так как ψ^{-1} вогнута и $\mathbf{E}\psi\big(|\xi_t|\big) \le 1$, то применив неравенство Йенсена, завершаем доказательство. ▲

ТЕОРЕМА. *Пусть на T задана полунорма $d(\cdot,\cdot)$ и ξ_t удовлетворяет условию Липшица относительно нормы Орлича $\|\cdot\|_\psi$:*

$$\|\xi_t - \xi_s\|_\psi \le d(t,s).$$

Обозначим:
D — диаметр T: $D \triangleq \max_{t,s\in T} d(t,s)$,
$N(T,d;\epsilon)$ — минимальное число шаров, необходимых для покрытия T.
Тогда

$$\mathbf{E}\max_{t,s\in T}|\xi_t - \xi_s| \le 8\int_0^D \psi^{-1}\left[N(T,d;\epsilon)\right]d\epsilon.$$

Доказательство. Будем ипользовать следующие обозначения:

$T_l \in T$ — 2^{-l}– сеть, содержащая $N(T, d; 2^{-l})$ точек,

l_0 — максимальное целое число, такое что

$$N(T, d; 2^{-l_0}) = 1,$$

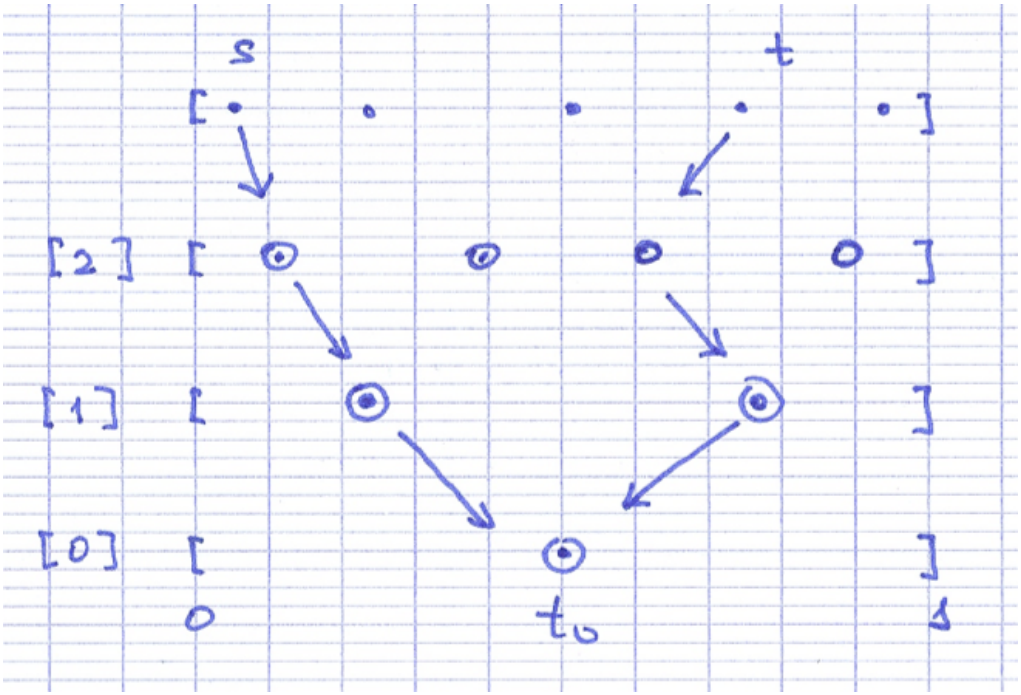
l_1 — минимальное целое число, такое что

$$N(T, d; 2^{-l_1}) = n.$$

Для любой точки $t \in T$ определим ее проекцию $\pi_l(t)$ на сеть T_l:

$$\pi_l(t) = \arg\min_{s \in T_l} d(t, s).$$

Пример: сети на $[0, 1]$.



Chaining: пусть $t \in T_{l_1}$, тогда

$$\xi_t - \xi_{t_0} = \xi_t - \xi_{\pi_{l_1-1}(t)} + \xi_{\pi_{l_1-1}(t)} - \xi_{\pi_{l_1-2}(\pi_{l_1-1}(t))} + \cdots$$

Поэтому, воспользовавшись леммой о максимуме случайных величин, у которых норма Орлича ограничена сверху 1, для любых $t, s \in T_{l_1}$ имеем

$$\mathbf{E} \max_{t,s \in T} |\xi_t - \xi_s| \le 2 \sum_{j=l_1}^{l_0} \mathbf{E} \max_{t \in T_j} |\xi_t - \xi_{\pi_{j-1}(t)}|$$

$$\le 2 \sum_{j=l_1}^{l_0} \max_{t \in T_j} \|\xi_t - \xi_{\pi_{j-1}(t)}\|_\psi \times \mathbf{E} \max_{t \in T_j} \frac{|\xi_t - \xi_{\pi_{j-1}(t)}|}{\|\xi_t - \xi_{\pi_{j-1}(t)}\|_\psi}$$

$$\le 2 \sum_{j=l_1}^{l_0} 2^{-(j-1)} \times \psi^{-1}\big[N(T, d; 2^{-j})\big].$$

Чтобы завершить доказательство заметим, что

$$2^{-j}\psi^{-1}\left[N(T,d;2^{-j})\right] \leq 2\int_{2^{-j}}^{2^{-j+1}} \psi^{-1}\left[N(T,d;\epsilon)\right]d\epsilon.$$

Поэтому,

$$\begin{aligned}\mathbf{E}\max_{t,s\in T}|\xi_t-\xi_s| \leq & 8\int_0^{2^{-l_0}} \psi^{-1}\left[N(T,d;\epsilon)\right]d\epsilon \\ = & 8\int_0^{D} \psi^{-1}\left[N(T,d;\epsilon)\right]d\epsilon. \quad \blacktriangle\end{aligned}$$

Замечание. Если случайные величины ξ_t суб-гауссовские, то поскольку

$$\psi(x) = \exp(x^2/2) - 1, \quad \psi^{-1}(x) = \sqrt{2\log(x+1)}$$

и

$$d(t,s) = \sqrt{\mathbf{E}[\xi_t - \xi_s]^2},$$

приходим к энтропийному неравенству Дадли

$$\mathbf{E} \max_{t,s \in T} |\xi_t - \xi_s| \leq 8 \int_0^D \sqrt{2 \log N(T, d; \epsilon)} d\epsilon;$$

здесь $\log N(T, d; \epsilon)$ — метрическая энтропия.

Замечание. Более точную верхнюю границу для $\max_{t\in T}\xi_t$ дает

ТЕОРЕМА о мажорирующей мере. *Пусть $B(t,\epsilon)$ — шар радиуса ϵ центром в t, μ — некоторая вероятностная мера на T.*

$$\mathbf{E}\max_{t,s\in T}|\xi_t-\xi_s|\le K_\psi\max_{t\in T}\int_0^D\psi^{-1}\left[\frac{1}{\mu[B(t,s)]}\right]d\epsilon;$$

здесь K_ψ — постоянная, зависящая только от функции Юнга $\psi(\cdot)$.

Статистическая теория обучения

Предположим, что у нас имеется n пар независимых одинаково распределенных случайных величин

$$(X_1, Y_1), \quad \dots \quad , (X_n, Y_n),$$

где $X_i \in \mathbf{R}^d$, $Y_i \in \{1, -1\}$.
Предположим, что мы знаем величину X_{n+1}.

Задача состоит в том, чтобы по имеющимся данным

$$\left\{X_{n+1}, (X_i, Y_i), i = 1, \dots, n\right\}$$

предсказать Y_{n+1}.

Пример: обнаружение спама в электронной почте.
Имеется n сообщений $\varepsilon_1, \ldots, \varepsilon_n$, каждое из которых мы классифицировали либо как спам $\{\psi_i = -1\}$ либо как полезное сообщение $\{\psi_i = 1\}$.
Необходимо придумать метод (алгоритм), который без нашей помощи решал бы, является ли сообщение ε_{n+1} спамом или нет, т.е. предсказал бы Y_{n+1}.
С каждым сообщением ε_i свяжем вектор $X_i \in \mathbf{R}^d$, который содержит частоту появления слов в сообщении ε_i.
Заметим, что величина d в рассматриваемой задаче должна быть велика $d \approx 1000$.

Формальное решение. Предположим, что нам известно распределение случайной величины X при $Y = 1$ и $Y = -1$, которые мы будем обозначать как

$$\mathbf{P}_1(X) \quad \text{и} \quad \mathbf{P}_{-1}(X).$$

Будем считать, что эти меры имеют плотности относительно некоторой σ-конечной меры μ. Обозначим их

$$p_1(X) \quad \text{и} \quad p_{-1}(X).$$

Наша цель найти функцию $g(x)$, которая минимизирует вероятность ошибки

$$e(g) = \pi \mathbf{P}_1\big(g(X) = -1\big) + (1 - \pi)\mathbf{P}_{-1}\big(g(x) = 1\big),$$

здесь $\pi = \mathbf{P}\big\{Y = 1\big\}$.

ЛЕММА Неймана-Пирсона. *Наилучшее решающее правило имеет вид*

$$g^*(X) = \arg\min_g \{e(g)\} = \begin{cases} 1, & \frac{\pi p_1(x)}{\pi p_1(x)+(1-\pi)p_{-1}(x)} \geq \frac{1}{2} \\ -1, & \frac{\pi p_1(x)}{\pi p_1(x)+(1-\pi)p_{-1}(x)} \leq \frac{1}{2} \end{cases}$$

Оценить $p_1(\cdot)$ и $p_{-1}(\cdot)$ можно только при небольших размерностях d.

Если оценивание $p_{\pm 1}(x)$ типичная задача, которая решается в статистике, то в теории автоматического обучения используется принципиально другой подход. А именно, задается некоторый класс функций $G = \{g(x) : \mathbf{R}^n \to \{-1, +1\}\}$ и прогноз ищется в этом классе.

Оптимальное решение этой задачи

$$g^* = \arg\min_g \underbrace{\mathbf{P}\{g(X) \neq Y\}}_{R(g)}$$

это прогноз, который минимизирует среднюю вероятность ошибки.

Для его вычисления нам опять же нужны плотности $p_{\pm 1}(x)$.
Поэтому, чтобы избежать их оценивания, мы определяем

$$g_n^* = \arg\min_{g \in G} \underbrace{R_n(g)}_{},$$

где

$$R_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{g(X_i) \neq Y_i\}$$

эмпирический риск.
Например,

$$G = \left\{ g(x) = \text{sing}(\langle \theta, X - X_0 \rangle),\ X_0,\ \theta \in \mathbf{R}^d \right\}.$$

Нас интересует верхняя граница для $R(g_n^*)$. Заметим, что

$$R(g_n^*) = \big[R(g_n^*) - R_n(g_n^*)\big] + [R_n(g_n^*) - R_n(g^*)] + R_n(g^*).$$

Ясно, что $R_n(g_n^*) \le R_n(g^*)$. Поэтому

$$R(g_n^*) \le R(g^*) + \big[R_n(g^*) - R(g^*)\big] + \big[R(g_n^*) - R_n(g_n^*)\big].$$

Оценивание выражений в квадратных скобках в этом неравенстве – типичная задача в теории эмпирических процессов.
Воспользовавшись теоремой Хёфдинга, несложно оценить

$$\mathbf{P}\big\{|R_n(g^*) - R(g^*)| > \delta\big\}.$$

Определение. *Эмпирический процесс — это набор с. в., индексируемых некоторым функциональным классом $\mathcal{F}$ и таких, что каждая с. в. есть сумма независимых одинаково распределенных величин.*

Для эмпирического процесса как правило используется обозначение

$$\left\{Pf - P_n f\right\}_{f \in \mathcal{F}},$$

где

$$Pf = \mathbf{E} f(Z_i), \quad P_n f = \frac{1}{n} \sum_{i=1}^{n} f(Z_i),$$

а Z_i — независимые, одинаково распределенные случайные величины.

В теории автоматического обучения класс функций $\mathcal{F}_G$ определеяется

$$f(X, Y) = \mathbf{1}\{g(X) \neq Y\}, \quad g \in G.$$

Поскольку $f(\cdot, \cdot) \in [0, 1]$, мы можем применить теорему Хёфдинга

$$\mathbf{P}\left\{ |R_n(g^*) - R(g^*)| > \epsilon \right\} \leq 2 \exp\left\{ -2n\epsilon^2 \right\}$$

или, что эквивалентно, сказать, что с вероятностью не меньшей, чем $1 - \delta$

$$|R(g^*) - R_n(g^*)| \leq \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}.$$

Для $R(g_n^*) - R_n(g_n^*)$ неравенство Хёфдинга напрямую применить нельзя поскольку g_n^* зависит от обучающей выборки.

Идея: использовать простое неравенство

$$R(g_n^*) - R_n(g_n^*) \le \max_{g \in G} \big[R(g) - R_n(g)\big] = \max_{f \in \mathcal{F}_G} \big[Pf - P_n f\big].$$

Предположим, что G состоит из конечного числа элементов. Тогда очевидно, что и $\mathcal{F}_G$ конечно и мы можем опять использовать неравенство Хёфдинга

$$\begin{aligned}\mathbf{P}\Big\{ \max_{f \in \mathcal{F}_G} |Pf - P_n f| \ge \epsilon \Big\} &\le \sum_{f \in \mathcal{F}_G} \mathbf{P}\Big\{ |Pf - P_n f| > \epsilon \Big\} \\ &\le 2N \exp\Big\{ -2n\epsilon^2 \Big\}.\end{aligned}$$

И, следовательно, с вероятностью не меньшей, чем $1 - \delta$

$$R(g_n^*) - R_n(g_n^*) \le \sqrt{\frac{1}{2n} \log\left(\frac{2N}{\delta}\right)}.$$

Средние Радемахера

Радемахеровские случайные величины это н. о. р. с. в., принимающие значения $\{1, -1\}$ с равными вероятностями. Будем обозначать для краткости

$$R_n^\sigma f = \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i),$$

σ_i — радемахеровские с. в.

Z_i – н.о.р.с.в. (В нашей задаче $Z_i = \left(X_i, Y_i\right)$ и $f(Z_i) = \mathbf{1}\{g(X_i) \neq Y_i\}$)

$\mathbf{E}_\sigma$ — усреднение по радемахеровским с. в.

$\mathbf{E}$ — усреднение по всем с. в.

Определение. *Радемахеровское среднее класса $\mathcal{F}$ определяется как*

$$\mathcal{R}(\mathcal{F}_G) = \mathbf{E} \max_{f \in \mathcal{F}_G} R_n^\sigma(f)$$

и условное радемахеровское среднее определяется следующим образом:

$$\mathcal{R}_n(\mathcal{F}_G) = \mathbf{E}_\sigma \max_{f \in \mathcal{F}_G} R_n^\sigma(f).$$

ТЕОРЕМА *Для любой функции $f \in \mathcal{F}_G$ с вероятностью не меньшей, чем $1 - \delta$*

$$P(f) \leq P_n(f) + 2\mathcal{R}(\mathcal{F}_G) + \sqrt{\frac{\log(1/\delta)}{n}},$$

$$P(f) \leq P_n(f) + 2\mathcal{R}_n(\mathcal{F}_G) + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Доказательство. Ключевую роль в нем играет следующий результат о концентрации меры.
ТЕОРЕМА(Mc. Diarmid) *Пусть* Z_i — *н. о. р. с. в. Предположим, что функция* $F(\cdot) : \mathbb{R}^n \to \mathbb{R}$ *такова, что для любого* $i = 1, \ldots, n$ *и любых* $x_1, \ldots, x_n,\ x_i' \in \mathbb{R}$

$$\left|F(x_1, \cdot\cdot, x_{i-1}, x_i, x_{i+1}, \cdot\cdot, x_n) - F(x_1, \cdot\cdot, x_{i-1}, x_i', x_{i+1}, \cdot\cdot, x_n)\right| \leq C.$$

Тогда

$$\mathbf{P}\left\{\left|F(Z_1, \ldots, Z_n) - \mathbf{E}F(Z_1, \ldots, Z_n)\right| > \epsilon\right\} \leq 2\exp\left\{-\frac{2\epsilon^2}{nC^2}\right\}.$$

Доказательство нашей теоремы будет следовать следующему плану:

1. с помощью теоремы Mc Diarmid свяжем

$$\max_{f \in \mathcal{F}_G} \left[Pf - P_n f \right] \quad \text{и} \quad \mathbf{E} \max_{f \in \mathcal{F}_G} \left[Pf - P_n f \right];$$

2. используем симметризацию, чтобы связать $\mathbf{E} \max_{f \in \mathcal{F}_G} \left[Pf - P_n f \right]$ и радемахеровское среднее класса $\mathcal{F}_G$.

I. Пусть P_n^i — эмпирическая мера, в которой модифицирован i-ый элемент

$$Z_i \to Z_i'.$$

Тогда мы имеем, учитывая, что $f \in \{0, 1\}$

$$\left| \max_{f \in \mathcal{F}_G} (Pf - P_n f) - \max_{f \in \mathcal{F}_G} (Pf - P_n^i f) \right| \leq \max_{f \in \mathcal{F}_G} |P_n^i f - P_n f| \leq \frac{1}{n}.$$

Поэтому согласно теореме Mc. Diarmid с вероятностью не меньшей, чем $1 - \delta$

$$\max_{f \in \mathcal{F}_G} (Pf - P_n f) \leq \mathbf{E} \max_{f \in \mathcal{F}_G} (Pf - P_n f) + \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}.$$

II. Лемма о симметризации

$$\mathbf{E} \max_{f \in \mathcal{F}_G} (Pf - P_n f) \leq 2\mathbf{E} \max_{f \in \mathcal{F}_G} R_n^\sigma(f).$$

Доказательство. Введем независимую от Z выборку Z', имеющую тоже самое вероятностное распределение, что и Z, и обозначим

$$P'_n f = \frac{1}{n} \sum_{i=1}^{n} f(Z'_i), \quad P'f = \mathbf{E} P'_n f.$$

Тогда, используя неравенство Йенсена, имеем

$$\mathbf{E}\max_{f\in\mathcal{F}_G}\left[Pf - P_n f\right] = \mathbf{E}\max_{f\in\mathcal{F}_G}\left[\mathbf{E}P_n' f - P_n f\right]$$

$$\leq \mathbf{E}\max_{f\in\mathcal{F}_G}\left[P_n' f - P_n f\right] = \mathbf{E}_\sigma\mathbf{E}\max_{f\in\mathcal{F}_G}\frac{1}{n}\sum_{i=1}^n \sigma_i\left[f(Z_i) - f(Z_i')\right]$$

$$\leq \mathbf{E}_\sigma\mathbf{E}\max_{f\in\mathcal{F}_G}\frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i) + \mathbf{E}_\sigma\mathbf{E}\max_{f\in\mathcal{F}_G}\left\{-\frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i')\right\}$$

$$= 2\mathbf{E}\max_{f\in\mathcal{F}_G} R_n^\sigma f.$$

Это неравенство завершает доказательство леммы и первого неравенства в теореме. Второе неравенство доказывается тем же самым способом, но неравенство концентрации Mc. Diarmid применяется для функции $F(Z_1,\ldots,Z_n) = \mathcal{R}_n(\mathcal{F}_G)$.

Связь с исходным классом G.

Из предыдущей теоремы вытекает, что с вероятностью не меньшей, чем $1 - \delta$ для любого $g \in G$ выполняется следующее неравенство:

$$R(g) \leq R_n(g) + 2\mathcal{R}_n(\mathcal{F}_G) + \sqrt{\frac{2\log(2/\delta)}{n}};$$

здесь

$$\mathcal{R}_n(\mathcal{F}_G) = \mathbf{E}_\sigma \max_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \mathbf{1}\{g(X_i) \neq Y_i\}.$$

Рademахеровское среднее $\mathcal{R}_n(\mathcal{F}_G)$ и рademахеровское исходного класса G

$$\mathcal{R}_n(G) = \mathbf{E}_\sigma \max_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(X_i)$$

очень тесно связаны. Действительно,

$$\mathcal{R}_n(\mathcal{F}_G) = \mathbf{E}_\sigma \max_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \mathbf{1}\{g(X_i) \neq Y_i\}$$

$$= \mathbf{E}_\sigma \max_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \frac{1 - Y_i g(X_i)}{2} = \frac{1}{2} \mathbf{E}_\sigma \max_{g \in G} \sum_{i=1}^{n} \frac{1}{n} \sigma_i Y_i g(X_i)$$

$$= \frac{1}{2} \mathbf{E}_\sigma \max_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(X_i).$$

Поэтому для любого $g \in G$ с вероятностью не меньшей, чем $1 - \delta$

$$R(g) \leq R_n(g) + \mathcal{R}_n(G) + \sqrt{\frac{2\log(2/\delta)}{n}}$$

и, в частности, отсюда вытекает, что

$$R(g_n) \leq \min_{g \in G} R(g) + \mathcal{R}_n(G) + 2\sqrt{\frac{2\log(2/\delta)}{n}}.$$

Напомним, что

$$g_n = \arg\min_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\Big\{ g(X_i) \neq Y_i \Big\}.$$

Для оценки $\mathcal{R}_n(\mathcal{F}_G)$ мы используем энтропийную границу Дадли .

ТЕОРЕМА.(Дадли)

$$\begin{aligned}\mathcal{R}_n(\mathcal{F}_G) =& \mathbf{E}_\sigma \max_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\\ \leq& \frac{12}{\sqrt{n}}\int_0^\infty \sqrt{\log N(\mathcal{F}_G, d;\epsilon)}d\epsilon.\end{aligned}$$

Здесь

$$d^2(f_1,f_2) = \frac{1}{n}\sum_{i=1}^{n}\left[f_1(Z_i)-f_2(Z_i)\right]^2.$$

Во многих случаях вычислить энтропию помогает следующий результат.

ТЕОРЕМА (Haussler (1995)) *Если размерность Вапника-Червоненкиса класса $\mathcal{F}_G$ не превосходит q, то*

$$N(\mathcal{F}_G, d; \epsilon) \leq Cq\left(\frac{4\mathrm{e}}{\epsilon}\right)^q;$$

здесь и далее C обозначает универсальные постоянные, значения которых могут меняться в зависимости от неравенства, в котором они появляются.

Таким образом, из этого результата и теоремы Дадли получаем верхнюю границу

$$\mathcal{R}_n(\mathcal{F}_G) \leq C\sqrt{\frac{q}{n}}.$$

Поэтому, объединяя полученные неравенства, приходим к следующему результату:
при выполнении условий предыдущей теоремы с вероятностью не меньшей, чем $1 - \delta$ *справедлива следующая верхняя граница для риска* g_n^**:*

$$R(g_n^*) \leq \min_{g \in G} R(g) + \sqrt{\frac{Cq + 4\log(2/\delta)}{n}};$$

здесь q *– размерность Вапника-Червоненкиса класса* G *и*

$$g_n^* = \arg\min_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{g(X_i) \neq Y_i\}.$$