МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ

(Machine Learning and Data Mining)

Н. Ю. Золотых

http://www.uic.unn.ru/~zny/ml

Лекция 17

Теория машинного обучения

Истоки: В. П. Вапник, А. Я. Червоненкис [1971]

Лемма 17.1 *Пусть* $A_1, A_2, \ldots, A_q$ *— некоторые события (зависимые или независимые), заданные на одном вероятностном пространстве, тогда*

$$\mathsf{Pr}(A_1 \cup A_2 \cup \cdots \cup A_q) \leq \mathsf{Pr}\, A_1 + \mathsf{Pr}\, A_2 + \cdots + \mathsf{Pr}\, A_q.$$

Лемма 17.2 (С.Н. Бернштейн, H. Chernoff, W. Hoeffding) *Пусть* $Z^{(1)}, Z^{(2)}, \ldots, Z^{(N)}$ *— независимые одинаково распределенные случайные величины.*

$$\mathsf{Pr}\left\{Z^{(i)} = 1\right\} = \theta, \qquad \mathsf{Pr}\left\{Z^{(i)} = 0\right\} = 1 - \theta$$

(схема Бернулли). Тогда

$$\mathsf{Pr}(|\widehat{\theta} - \theta| > \gamma) \leq 2e^{-2\gamma^2 N},$$

где

$$\widehat{\theta} = \frac{1}{N}\sum_{i=1}^{N} Z^{(i)}.$$

Рассмотрим задачу бинарной классификации.

$\mathscr{F}$ — некоторое множество функций $f:\ \mathbf{R}^d \to \{0,1\}$.

(X,Y) — $(d+1)$-мерная с.в. с неизвестной функцией распределения $P(x,y)$

Ожидаемый риск

$$R(f) = \mathsf{E}\, I\big(f(X) \neq Y\big) = \mathsf{Pr}\ I\big(f(X) \neq Y\big).$$

$(x^{(1)}, y^{(1)}),\ (x^{(2)}, y^{(2)}),\ \ldots, (x^{(N)}, y^{(N)})$ — обучающая выборка.

Эмпирический риск (частота ошибки) на данной выборке есть

$$\widehat{R}(f) = \frac{1}{N}\sum_{i=1}^{N} I\big(f(x^{(i)}) \neq y^{(i)}\big).$$

Задача: оценить $R(f)$ в терминах $\widehat{R}(f)$.

Принцип минимизации эмпирического риска:

$$\widehat{f} = \operatorname*{argmin}_{f\in\mathscr{F}} \widehat{R}(f)$$

17.1. Случай конечного $\mathscr{F}$

Пусть $\mathscr{F} = \{f_1, f_2, \ldots, f_q\}$.

Рассмотрим $f \in \mathscr{F}$.

Введем семейство случайных величин

$$Z^{(i)} = I\big(f(X^{(i)}) \neq Y^{(i)}\big) \qquad (i = 1, 2, \ldots, N).$$

$Z^{(i)}$ распределены одинаково и независимо — схема испытаний Бернулли.

По лемме 2

$$\mathsf{Pr}\,(|\widehat{R}(f) - R(f)| > \gamma) \leq 2e^{-2\gamma^2 N}$$

Откуда

$$\mathsf{Pr}\,(|\widehat{R}(f) - R(f)| \leq \gamma) \geq 1 - 2e^{-2\gamma^2 N}$$

Итак, для *конкретной* f эмпирический риск *близок* к ожидаемому с *большой* вероятностью.

Нас же интересует более сильное условие:

$$\mathsf{Pr}\,\Big\{\forall\, f \in \mathscr{F} :\ |\widehat{R}(f) - R(f)| \leq \gamma\Big\}$$

$$\mathsf{Pr}\left\{\exists\, f \in \mathscr{F}:\ |\widehat{R}(f) - R(f)| > \gamma\right\} = \mathsf{Pr}(A_1 \cup \cdots \cup A_q) \le \sum_{i=1}^{q} \mathsf{Pr}(A_i) \le \sum_{i=1}^{q} 2e^{-2\gamma^2 N} = 2qe^{-2\gamma^2 N}$$

Мы доказали утверждение:

Утверждение 17.3 (О равномерной сходимости $\widehat{R}$ к R в случае конечного $\mathscr{F}$)

$$\mathsf{Pr}\left\{\forall\, f \in \mathscr{F}:\ |\widehat{R}(f) - R(f)| \le \gamma\right\} \ge 1 - 2qe^{-2\gamma^2 N}$$

Следствие 17.4 *Если* $N \ge \dfrac{1}{2\gamma^2} \ln \dfrac{2q}{\delta}$, *тогда*

$$\mathsf{Pr}\left\{\forall\, f \in \mathscr{F}:\ |\widehat{R}(f) - R(f)| \le \gamma\right\} \ge 1 - \delta.$$

Следствие 17.5

$$\mathsf{Pr}\left\{\forall\, f \in \mathscr{F}:\ |\widehat{R}(f) - R(f)| \le \sqrt{\frac{1}{N} \ln \frac{2q}{\delta}}\right\} \ge 1 - \delta$$

Следствие 17.6 (Принцип минимизации эмпирического риска для конечного $\mathscr{F}$) *Пусть*

$$f^* = \operatorname*{argmin}_{f \in \mathscr{F}} R(f), \qquad \widehat{f} = \operatorname*{argmin}_{f \in \mathscr{F}} \widehat{R}(f),$$

тогда

$$\mathsf{Pr}\left\{R(\widehat{f}) \le R(f^*) + 2\sqrt{\frac{1}{N}\ln\frac{2q}{\delta}}\right\} \ge 1 - \delta.$$

Доказательство. Если $\forall\, f \in \mathscr{F}$

$$|\widehat{R}(f) - R(f)| \le \gamma = \sqrt{\frac{1}{N}\ln\frac{2q}{\delta}},$$

то

$$0 \le R(\widehat{f}) - R(f^*) = R(\widehat{f}) - \widehat{R}(\widehat{f}) + \widehat{R}(\widehat{f}) - \widehat{R}(f^*) + \widehat{R}(f^*) - R(f^*) =$$
$$= \underbrace{R(\widehat{f}) - \widehat{R}(\widehat{f})}_{\le\gamma} + \underbrace{\widehat{R}(\widehat{f}) - \widehat{R}(f^*)}_{\le 0} + \underbrace{\widehat{R}(f^*) - R(f^*)}_{\le\gamma} \le 2\gamma.$$

Поэтому

$$\mathsf{Pr}\left\{0 \le R(\widehat{f}) - R(f^*) \le 2\gamma\right\} \ge \mathsf{Pr}\left\{R(\widehat{f}) - \widehat{R}(\widehat{f}) \le \gamma \text{ и } \widehat{R}(f^*) - R(f^*) \le \gamma\right\} \ge$$
$$\ge \mathsf{Pr}\left\{\forall\, f \in \mathscr{F} :\ |\widehat{R}(f) - R(f)| \le \gamma\right\} \ge 1 - \delta.$$

∎

17.2. Случай бесконечного $\mathscr{F}$

Коэффициентом разнообразия (*shatter coefficient*) $\Delta_0(\mathscr{F}, \mathbf{x})$ множества $\mathscr{F}$ на выборке $\mathbf{x} = \left\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\right\}$ – количество всевозможных способов разбить $\mathbf{x}$ на 2 подмножества, т. е. количество всевозможных бинарных векторов вида

$$\left(f(x^{(1)}), f(x^{(2)}), \ldots, f(x^{(N)})\right),$$

порождаемых всеми функциями f из $\mathscr{F}$.

Функция роста:

$$\Delta(\mathscr{F}, N) = \max_{|\mathbf{x}|=N} \Delta_0(\mathscr{F}, \mathbf{x}).$$

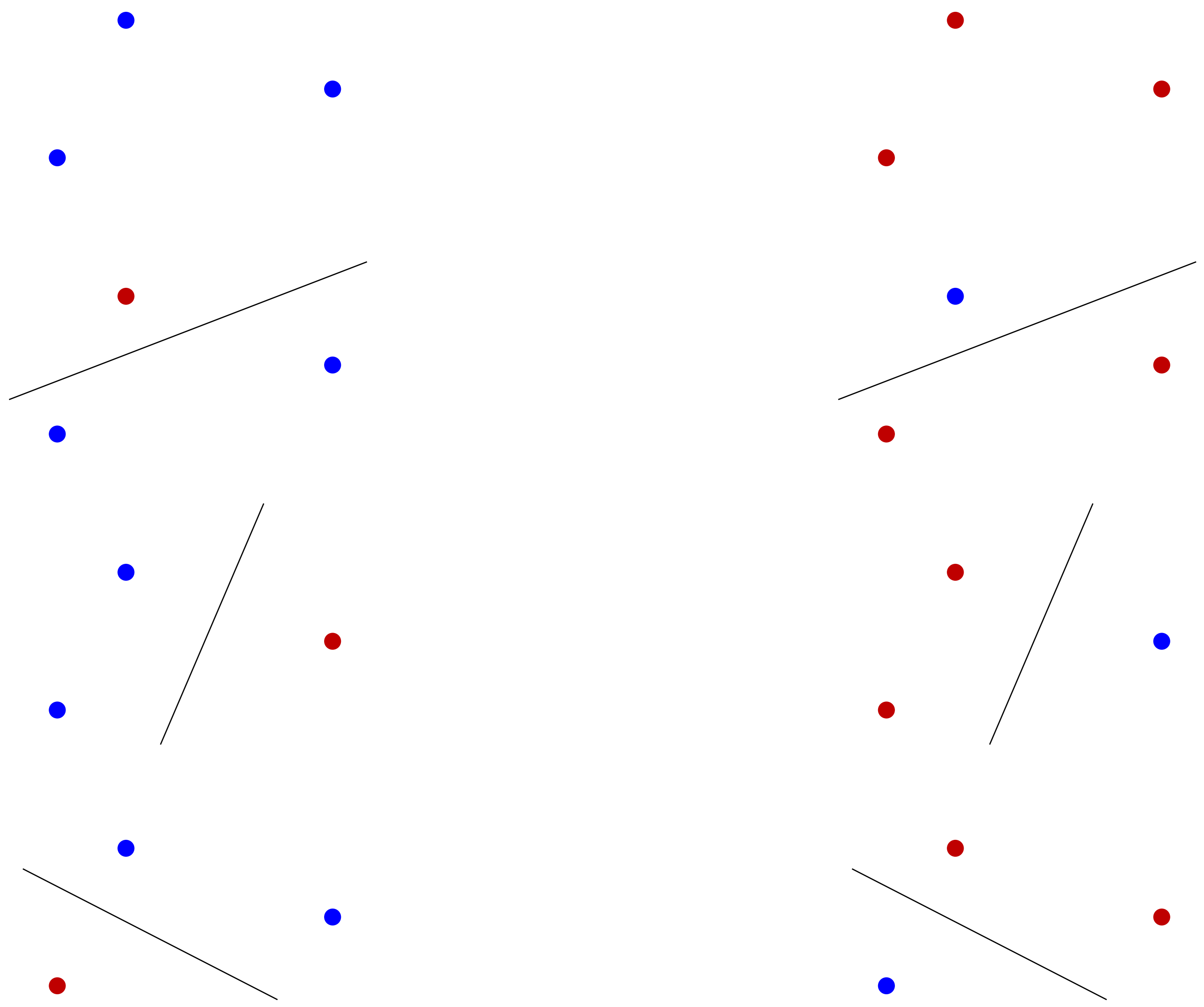
Пример.

В качестве $\mathscr{F}$ рассмотрим класс пороговых функций (линейных решающих правил)

$$\mathscr{F} = \mathrm{TF}_d = \left\{ f(x) = I(\beta_0 + \langle \beta, x \rangle \leq 0) : \ \beta_0 \in \mathbf{R}, \ \beta \in \mathbf{R}^d \right\}$$

$$\Delta_0(\mathrm{TF}_d, \, \mathbf{x}) = 8$$

(все возможные разбиения)



Если 3 точки на прямой, то $\Delta_0(\mathrm{TF}_d, \mathbf{x}) = 6$

Если N точек на прямой, то $\Delta_0(\mathrm{TF}_d, \mathbf{x}) = 2N$

Если 4 точки (в общем положении) в $\mathbf{R}^2$, то ...

$\Delta_0(\mathrm{TF}_d, \mathbf{x}) = 14$, т. е. все без двух

Утверждение 17.7 *Функция роста класса пороговых функций равна*

$$\Delta(\mathrm{TF}_d, N) = 2\left(\binom{N-1}{0} + \binom{N-1}{1} + \dots + \binom{N-1}{d}\right).$$

Например,

$\Delta(\mathrm{TF}_2, 2) = 4$, $\Delta(\mathrm{TF}_2, 3) = 8$, $\Delta(\mathrm{TF}_2, 4) = 14$, $\Delta(\mathrm{TF}_2, 5) = 22$

$\Delta(\mathrm{TF}_3, 2) = 4$, $\Delta(\mathrm{TF}_3, 3) = 8$, $\Delta(\mathrm{TF}_3, 4) = 16$, $\Delta(\mathrm{TF}_3, 5) = 30$

Размерность Вапника–Червоненкиса, или *емкость*, $\operatorname{VC}\mathscr{F}$ – наибольшая мощность множества в $\mathscr{X}$, *разбиваемая* (*shatter*) функциями из $\mathscr{F}$.

Говорят, что $\mathscr{F}$ *разбивает* множество $\left\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\right\}$, если функции из $\mathscr{F}$ позволяют разделить это множество на два подмножества всеми 2^N возможными способами, т. е. для любого двоичного набора $\mathbf{y} \in \{0,1\}^N$ найдется $f \in \mathscr{F}$, такое, что $\mathbf{y} = \left(f(x^{(1)}),\ f(x^{(2)}),\ \ldots,\ f(x^{(N)})\right)$.

Эквивалентно: $\operatorname{VC}\mathscr{F}$ – это максимальное N, при котором $\Delta(\mathscr{F}, N) = 2^N$

Если такого максимального N не существует, то $\operatorname{VC}\mathscr{F} = \infty$

Чтобы доказать, что $\operatorname{VC}\mathscr{F} = h$ достаточно проделать две вещи:

1) доказать $\operatorname{VC}\mathscr{F} \geq h$; для этого достаточно построить $\left\{x^{(1)}, x^{(2)}, \ldots, x^{(h)}\right\}$, разбиваемое функциями из $\mathscr{F}$;

2) доказать $\operatorname{VC}\mathscr{F} \leq h$; для этого достаточно показать, что никакое множество из $n + 1$ точек нельзя разбить функциями из $\mathscr{F}$.

Утверждение 17.8 *Размерность Вапника–Червоненкиса класса пороговых функций равна*

$$\mathrm{VC}(\mathrm{TF}_d) = d + 1.$$

Утверждение 17.9 *Если* $\mathscr{F}$ *конечно, то*

$$\mathrm{VC}(\mathscr{F}) \leq \log_2 |\mathscr{F}|.$$

Теорема 17.10 (Вапник–Червоненкис) *(О равномерной сходимости $\widehat{R}$ к R в случае конечной $\mathrm{VC}\,\mathscr{F}$)*
Пусть $\mathscr{F}$ задано и $\mathrm{VC}\,\mathscr{F}$ конечно, тогда

$$\mathsf{Pr}\left\{\forall\, f \in \mathscr{F}:\ |\widehat{R}(f) - R(f)| \leq \gamma\right\} \geq 1 - \delta, \tag{1}$$

$$\mathsf{Pr}\left\{R(\widehat{f}) \leq R(f^*) + 2\gamma\right\} \geq 1 - \delta. \tag{2}$$

где

$$\gamma = O\left(\sqrt{\frac{\mathrm{VC}\,\mathscr{F}}{N}\ln\frac{N}{\mathrm{VC}\,\mathscr{F}} + \frac{1}{N}\ln\frac{1}{\delta}}\right). \tag{3}$$

В П.Ч. неравенства $R(\widehat{f}) \leq R(f^*) + 2\gamma$ с ростом $\mathrm{VC}\,\mathscr{F}$ первое слагаемое убывает, а второе возрастает.

Сумма достигает своего минимума при некотором оптимальном значении $\mathrm{VC}\,\mathscr{F}$.

$\mathrm{VC}\,\mathscr{F}$ мало (модель с малой емкостью) – недообучение

$\mathrm{VC}\,\mathscr{F}$ велико (модель с крайне большой емкостью) – переобучение

Также можно доказать, что если $\mathrm{VC}\,\mathscr{F} = \infty$, то равномерной сходимости нет.

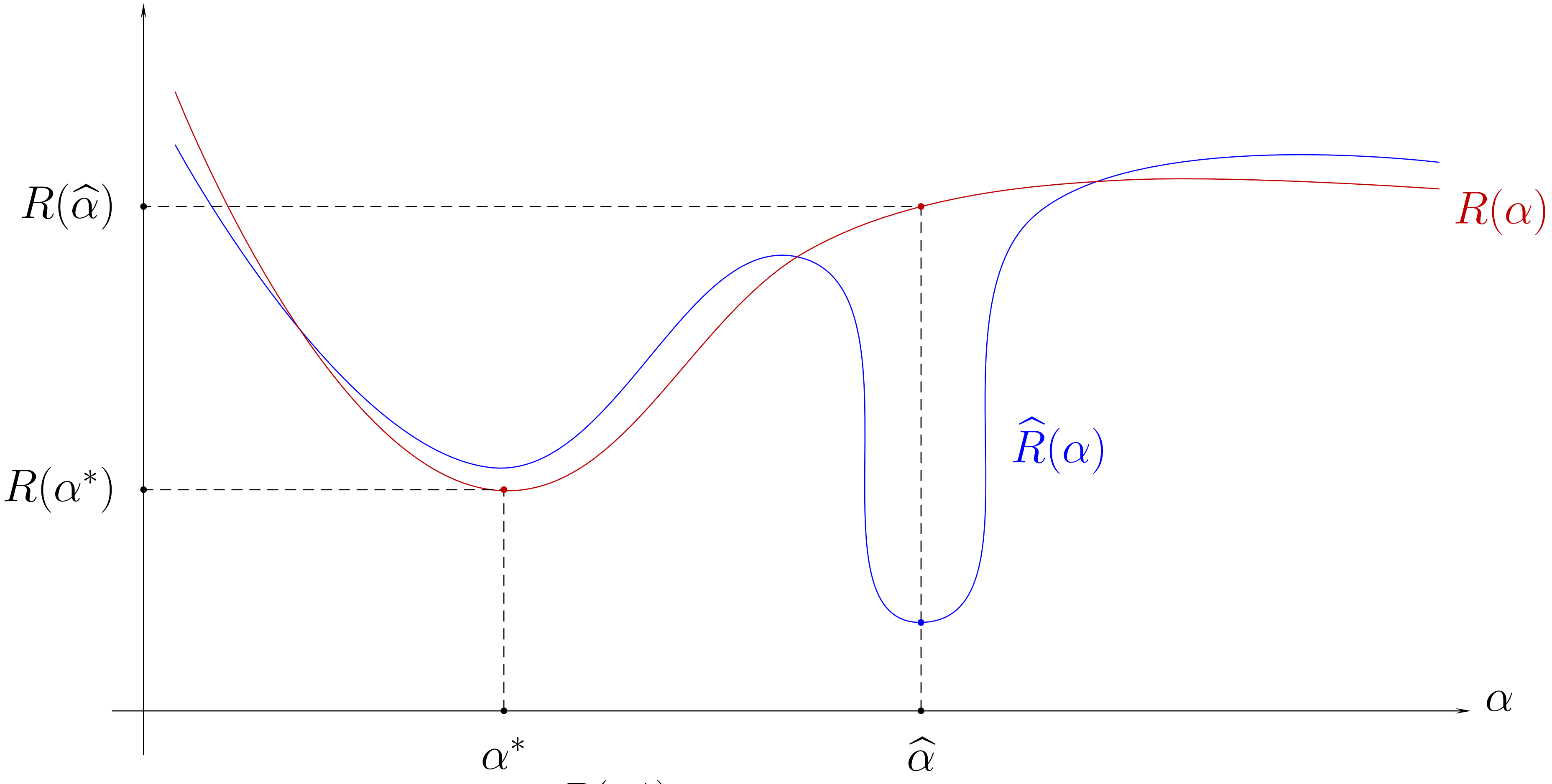
Как бороться с переобучением:

уменьшать $\mathrm{VC}\,\mathscr{F}$

увеличивать N

Пусть $\mathscr{F} = \{f : \; f(x, \alpha), \; \alpha \in [0, 1]\}$ — класс решающих правил

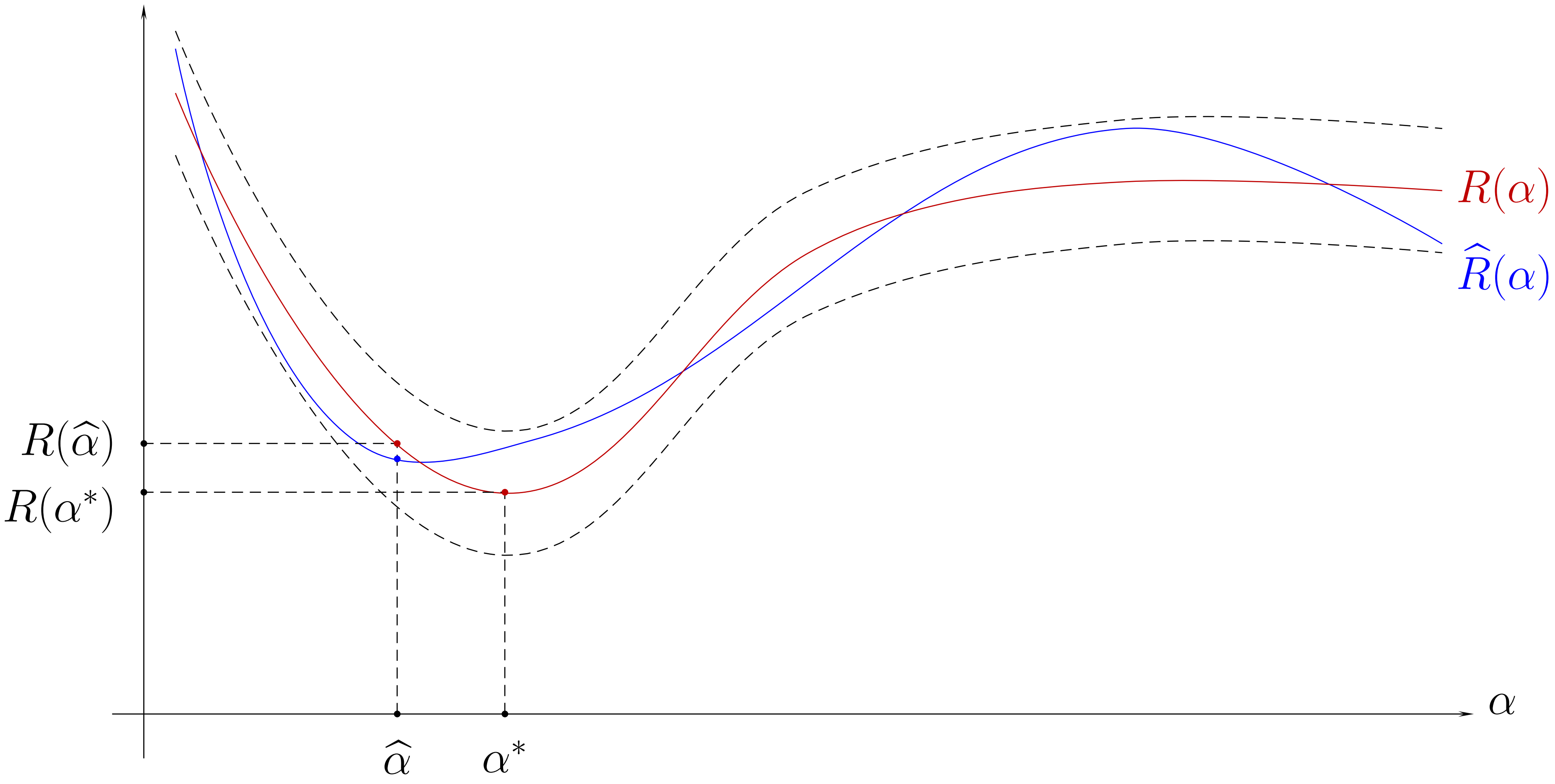
$R(\alpha)$ — средний риск, $\widehat{R}(\alpha)$ — эмпирический риск на функции $f(x, \alpha)$



$R(\widehat{\alpha})$ далеко от минимального значения $R(\alpha^*)$, хотя

$$\lim_{N \to \infty} \mathsf{Pr} \left\{ |\widehat{R}(f) - R(f)| \leq \gamma \right\} = 1.$$

Равномерная сходимость



$$\lim_{N\to\infty} \mathsf{Pr}\left\{\forall f :\in \mathscr{F} : \ |\widehat{R}(f) - R(f)| \le \gamma\right\} = 1.$$

Принцип *структурной минимизации риска* предназначен для отыскания оптимальной емкости.

Пусть в $\mathscr{F}$ выделена некоторая цепочка подсемейств

$$\mathscr{F}_1 \subset \mathscr{F}_2 \subset \cdots \subset \mathscr{F}_M = \mathscr{F}\,.$$

Необходимо выбрать семейство с минимальным значением верхней оценки в (2).

К сожалению, оценки (1)–(3) являются слишком завышенными, чтобы их можно было использовать для практических вычислений. Однако они представляют большой интерес с теоретической точки зрения и являются стимулом для дальнейших исследований в поиске подобных оценок, в частности в теории вероятностно–приближенно–корректного обучения (probably approximately correct learning, PAC–learning).

Изложенные здесь результаты распространяются на задачу классификации с бóльшим числом классов и на задачу восстановления регрессии.

Пусть $\mathscr{F}$ — некоторый класс функций $f:\ \mathscr{X} \to \mathscr{Y}$. Размерностью Вапника–Червоненкиса для класса $\mathscr{F}$ называется $\operatorname{VC}\mathscr{F} = \operatorname{VC}\mathscr{F}'$, где

$$\mathscr{F}' = \bigl\{I\bigl(f(x) - y\bigr):\ f \in \mathscr{F},\, y \in \mathscr{Y}\bigr\}\,.$$

17.2.1.* Функция роста для линейных классификаторов

Система из N неравенств с неизвестными $\beta_0, \beta_1, \ldots, \beta_d$:

$$\begin{cases} \beta_0 + \langle \beta, x \rangle \leq 0, \text{ если } f(x) = 0 \\ \beta_0 + \langle \beta, x \rangle > 0, \text{ если } f(x) = 1 \end{cases}$$

Не нарушая общности, можем считать, что точка 0 — один из объектов.

Множество всех функций разобьется на два равномощных класса:

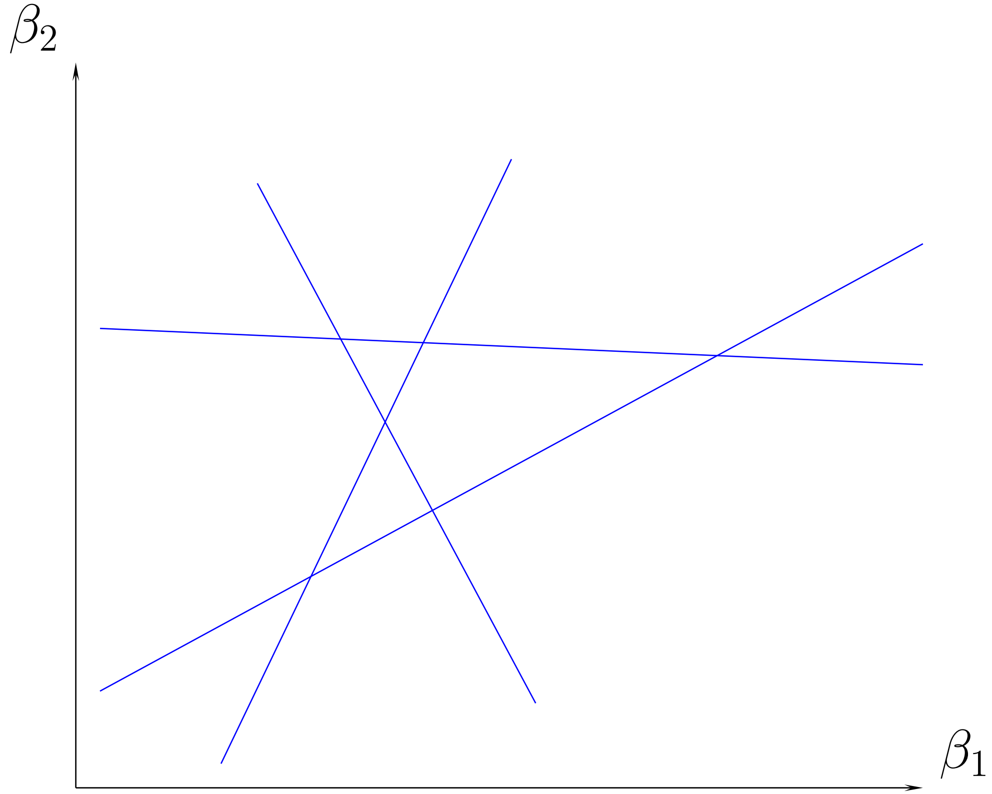
для которых $f(0) = 0$

и для которых $f(0) = 1$

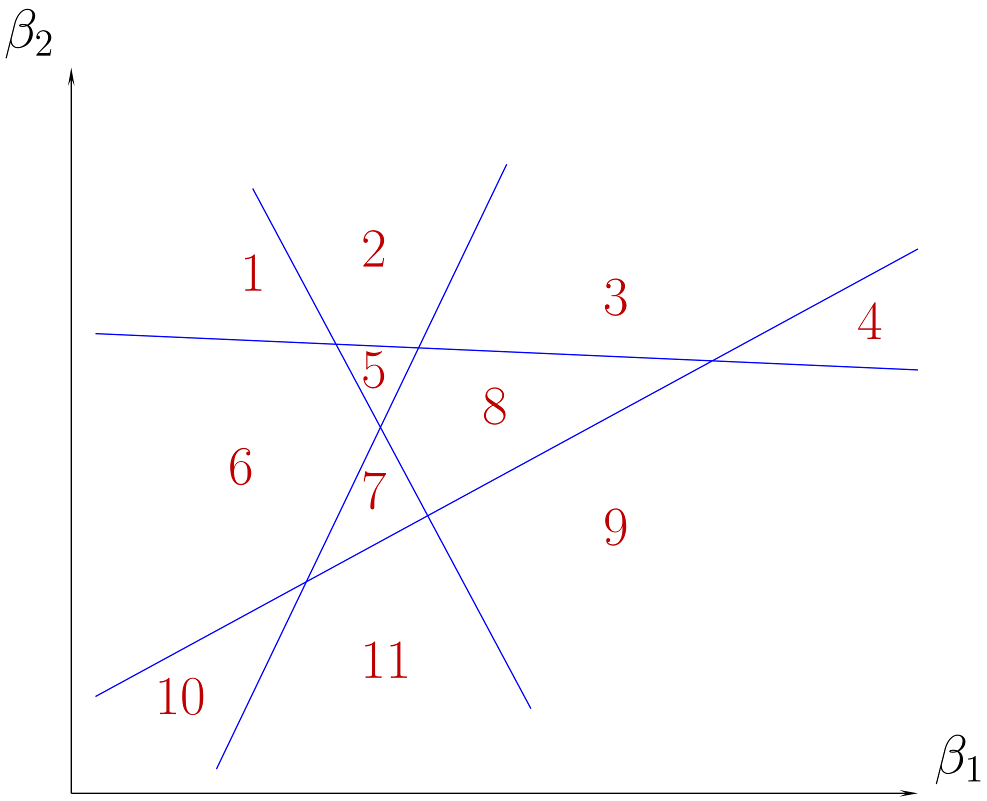
Будем считать только первые, а затем результат умножим на 2.

Для них можно считать, что $\beta_0 = -1$:

$$\begin{cases} -1 + \langle \beta, x \rangle \leq 0, \text{ если } f(x) = 0 \\ -1 + \langle \beta, x \rangle > 0, \text{ если } f(x) = 1 \end{cases}$$



β_2
β_1



Найдем максимальное количество $\Gamma(d, N)$ областей, на которые можно разбить $\mathbf{R}^d$ N гиперплоскостями. Справедливо рекуррентное соотношение

$$\Gamma(d, N) = \Gamma(d,\ N-1) + \Gamma(d-1,\ N-1),$$

откуда

$$\Gamma(d, N) = \binom{N}{0} + \binom{N}{1} + \cdots + \binom{N}{d}.$$

Теорема 17.11 (Лемма Зауэра; Sauer, 1972; Shelah, 1972) *Пусть класс* $\mathscr{F}$ *имеет конечную размерность Вапника–Червоненкиса* $\mathrm{VC}(\mathscr{F}) = h$*, тогда*

$$\Delta(\mathscr{F},\, N) \leq \binom{N}{0} + \binom{N}{1} + \cdots + \binom{N}{h} < \left(\frac{eN}{h}\right)^h$$