Линейные классификаторы

и немножко теории…

Perceptron

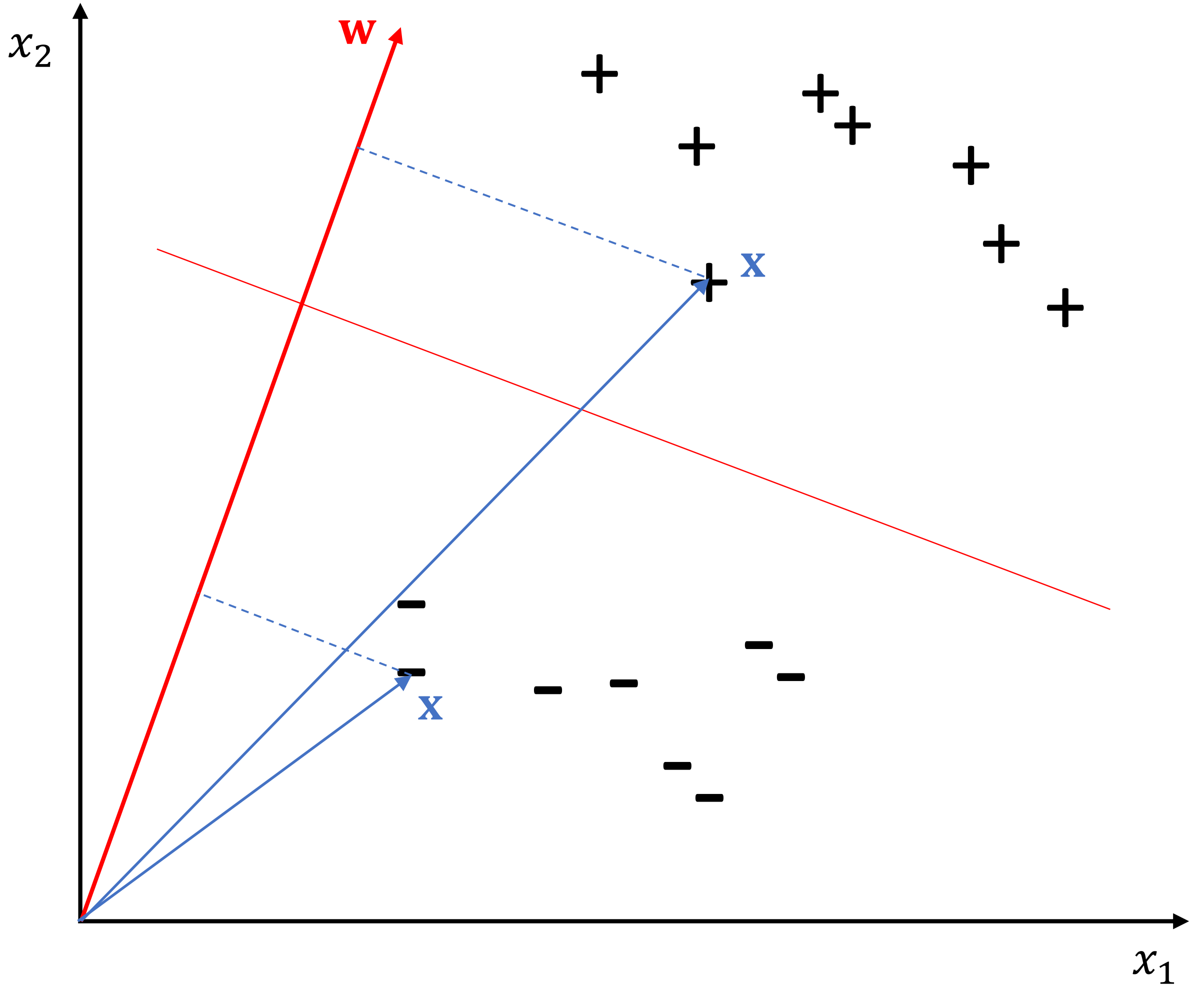
$$y \in \{-1,1\}$$

$$h(\mathbf{x}) = \text{sign}\left(\left(\sum_{i=1}^{d} w_i x_i\right) - \mathit{threshold}\right)$$

$$h(\mathbf{x}) = \text{sign}\left(\left(\sum_{i=1}^{d} w_i x_i\right) - w_0\right)$$

$$h(\mathbf{x}) = \text{sign}\left(\sum_{i=0}^{d} w_i x_i\right)$$

$$h(\mathbf{x}) = \text{sign}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}\right)$$



Обучение перцептрона

$$h(\mathbf{x}) = \mathrm{sign}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}\right)$$

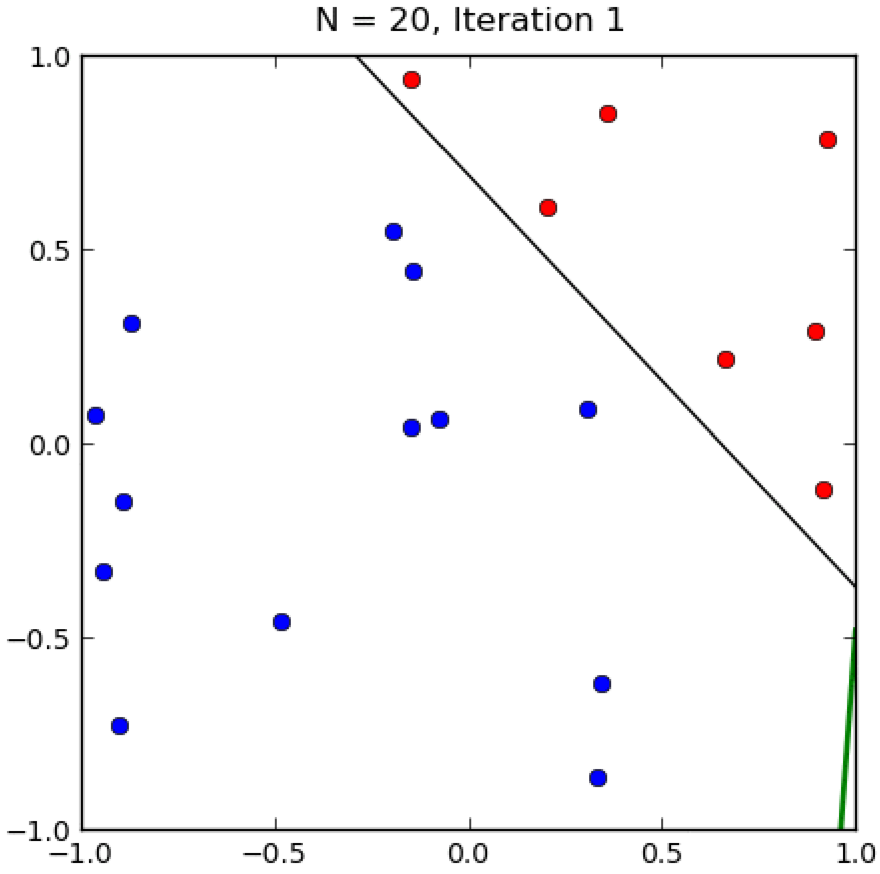
Data: $(\mathbf{x_1}, \boldsymbol{y_1}), (\mathbf{x_2}, \boldsymbol{y_2}), \dots, (\mathbf{x_N}, \boldsymbol{y_N})$

Алгоритм:

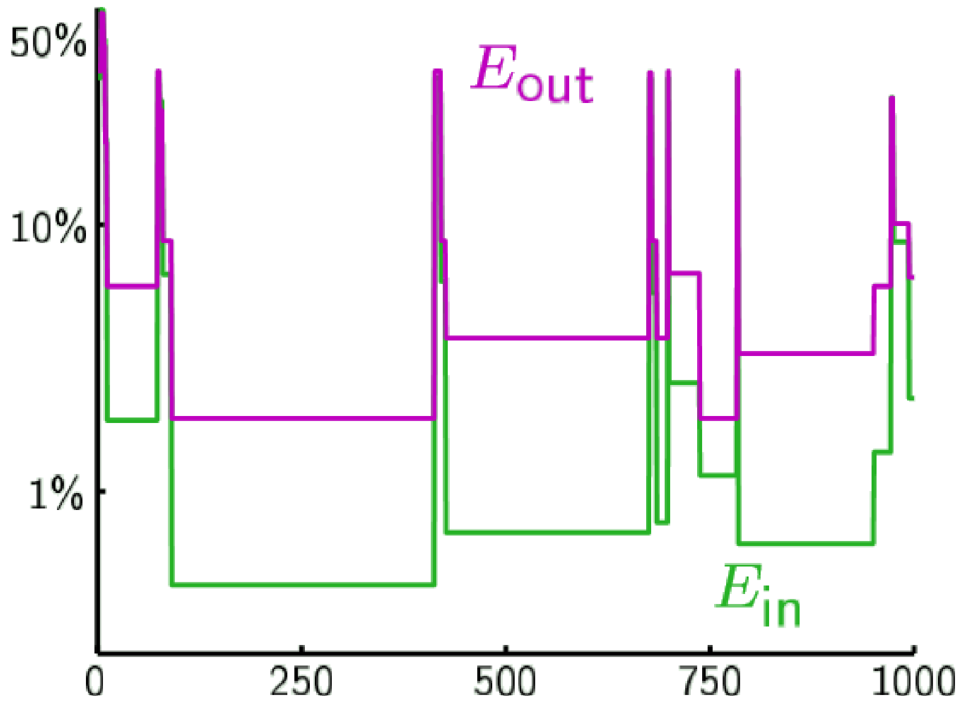
Начнем со случайного вектора w

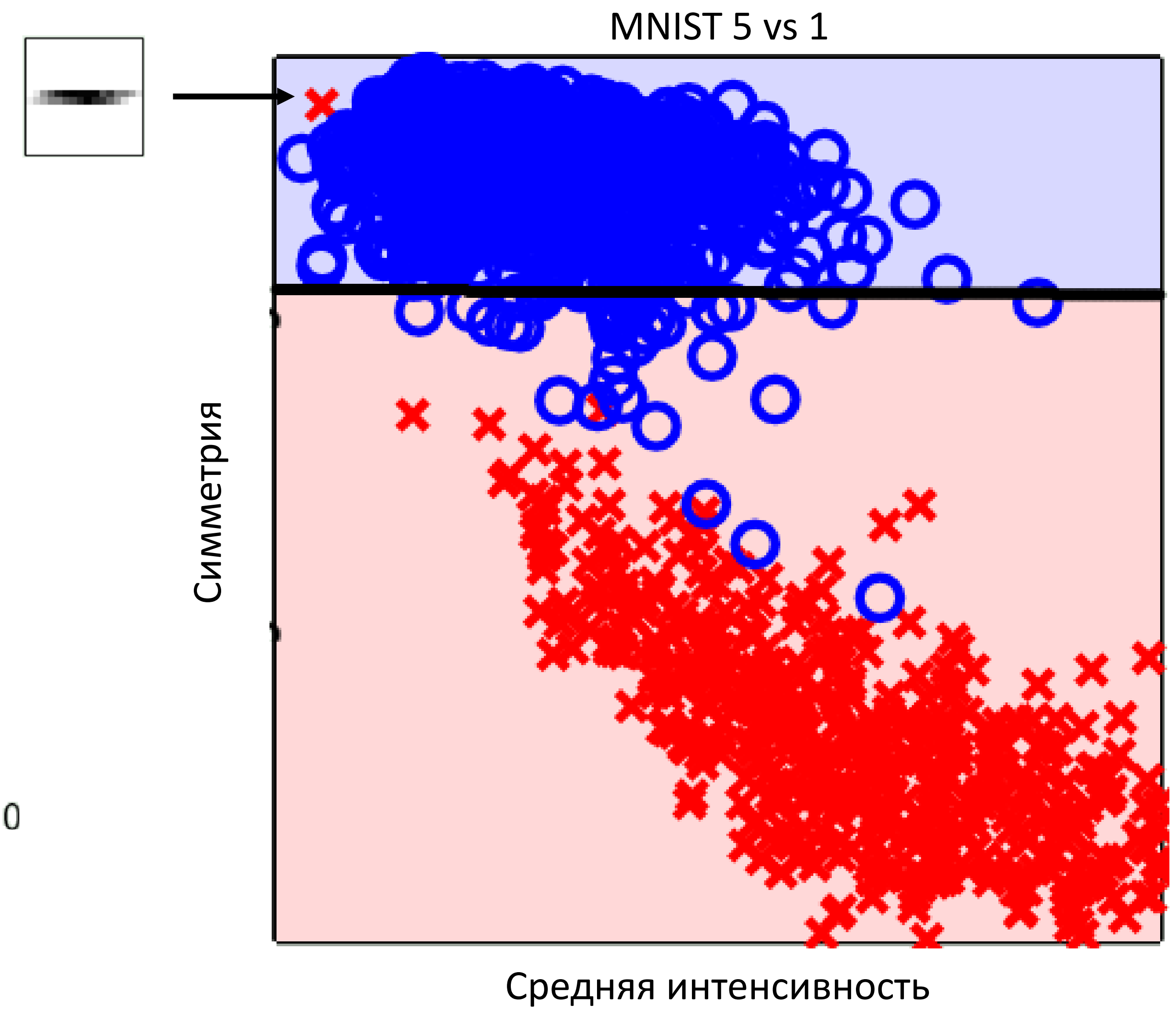
Найдем такой $\mathbf{x_i}$, что $h(\mathbf{x_i}) \neq \mathrm{y_i}$

$$\mathbf{w} \leftarrow \mathbf{w} + y_i \mathbf{x_i}$$



Перцептрон с карманом





Подсчет ошибки

$$E_{in}(h) = \frac{1}{N}\sum_{i=1}^{N} e\big(h(\mathrm{x}_i), f(\mathrm{x}_i)\big)$$

in sample (в выборке)

$$E_{out}(h) = \mathrm{E}_{\mathrm{x}}[e\big(h(\mathrm{x}), f(\mathrm{x})\big)]$$

out of sample (вне выборки)

Неравенство Хёфдинга

$$P[\underbrace{|E_{in}(h) - E_{out}(h)|}_{\substack{\text{Ошибка обобщения} \\ \text{Generalization error}}} > \epsilon] \leq 2e^{-\epsilon^2 N}$$

Иногда ошибкой обобщения называют E_{out}

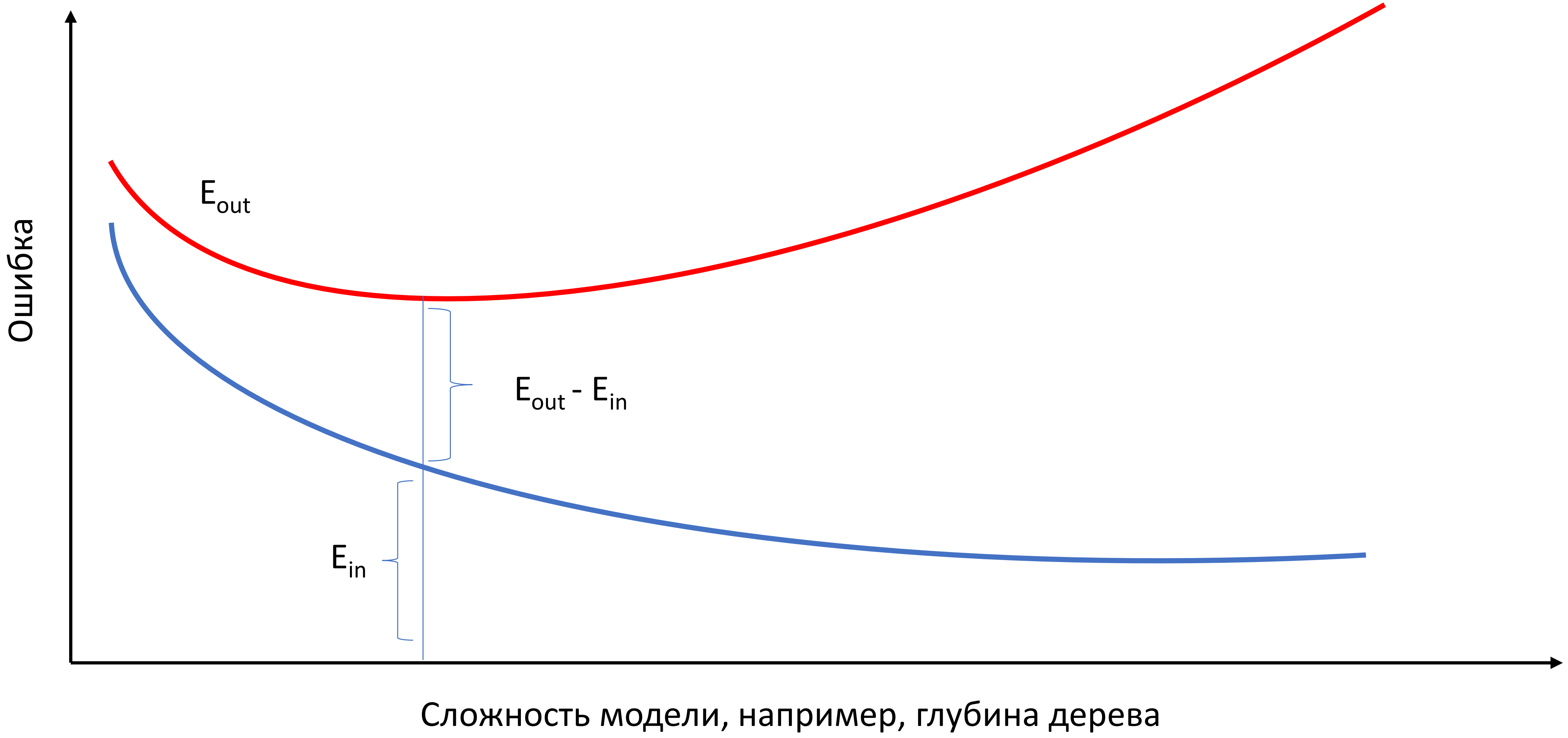
Неравенство Хёфдинга

$$P[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2e^{-\epsilon^2 N}$$

$$P[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq \mathrm{M}2e^{-\epsilon^2 N}$$

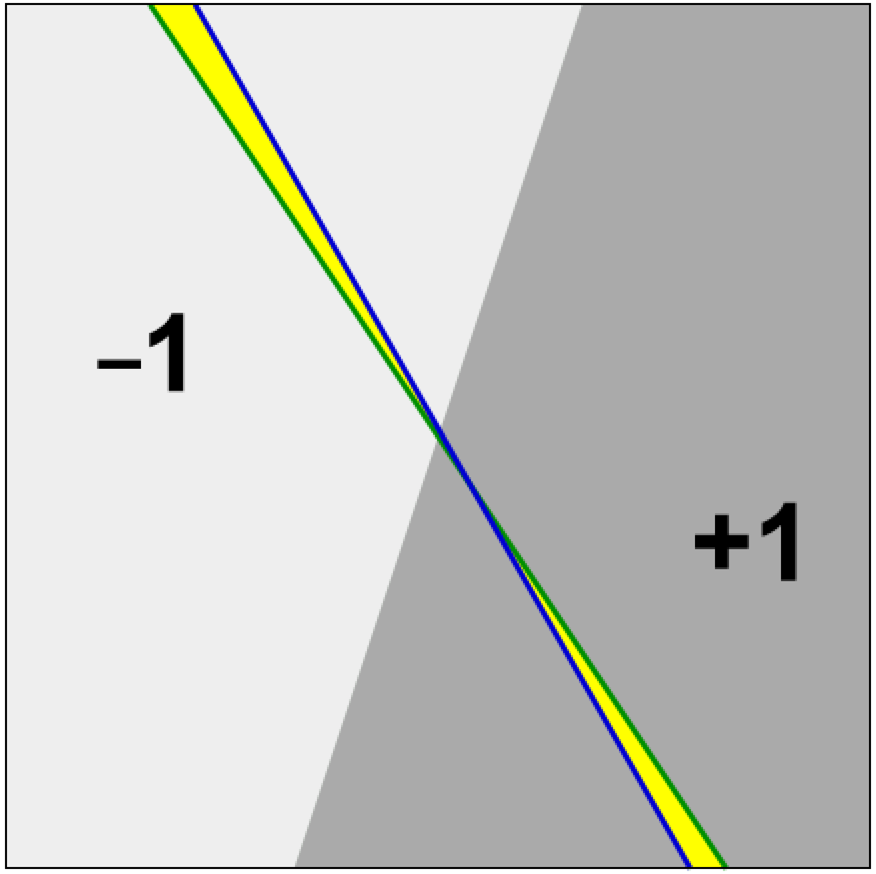
M – количество гипотез

Сложность модели



Близкие гипотезы

$$|E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| \approx |E_{\text{in}}(h_2) - E_{\text{out}}(h_2)|$$



От гипотез к дихотомиям

- Гипотеза: $h\colon \mathrm{X} \to \{-1, +1\}$
- Дихотомия: $h\colon \{\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_N}\} \to \{-1, +1\}$
- Максимальное количество дихотомий: 2^N

Функция роста (growth function)

$$m_H(N) = \max_{x_1, \dots, x_N} |H(\mathbf{x_1}, \dots, \mathbf{x_N})|$$

$$m_H(N) \le 2^N$$

Неравенство Вапника-Червоненкиса

$$P[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq \mathrm{M}2e^{-\epsilon^2 N}$$

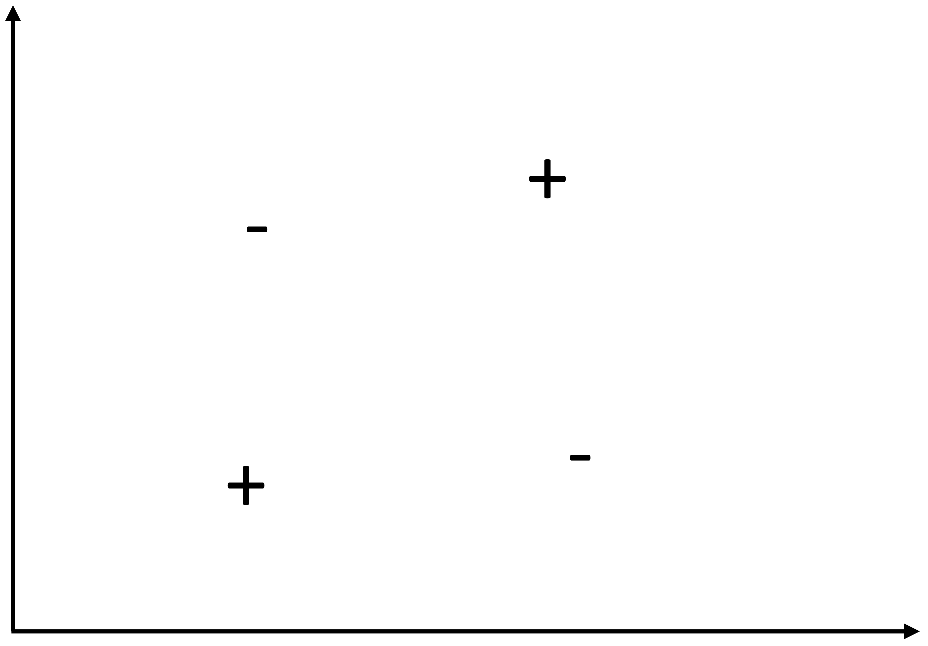
$$\Downarrow \quad \Downarrow \quad \Downarrow$$

$$P[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq m_H(2N)4e^{-\epsilon^{\frac{1}{8}}N}$$

Точка “поломки” (Breakpoint)

Определение: если никакой набор данных размера ***k*** нельзя распределить на все возможные случаи набором гипотез ***H***, то ***k*** – точка поломки для ***H***.

$$\min(k\colon\ m_H(k) < 2^k)$$



Для 2D перцептрона, ***k = 4.***

Доказательство полиномиальности функции роста в присутствии точки поломки

- $B(N,k) = m_H(N)$ с точкой поломки k
- $B(N,k) = \alpha + 2\beta$
- $\alpha + \beta \le B(N-1,k)$
- $\beta \le B(N-1,k-1)$
- $B(N,k) \le B(N-1,k) + B(N-1,k-1)$
- Докажем, что $B(N,k) \le \sum_{i=0}^{k-1} C_N^i$

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\ldots$	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$	
	$+1$	$+1$	$\ldots$	$+1$	$+1$	1 вариант для x_N
	-1	$+1$	$\ldots$	$+1$	-1	
α	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
	$+1$	-1	$\ldots$	-1	-1	
	-1	$+1$	$\ldots$	-1	$+1$	
	$+1$	-1	$\ldots$	$+1$	$+1$	2 варианта для x_N
	-1	-1	$\ldots$	$+1$	$+1$	
β	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
	$+1$	-1	$\ldots$	$+1$	$+1$	
	-1	-1	$\ldots$	-1	$+1$	
	$+1$	-1	$\ldots$	$+1$	-1	
	-1	-1	$\ldots$	$+1$	-1	
β	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
	$+1$	-1	$\ldots$	$+1$	-1	
	-1	-1	$\ldots$	-1	-1	

Индукция

$$B(N,k) \le \sum_{i=0}^{k-1} C_N^i$$

$$B(N,1) = 1, \qquad B(1, k > 1) = 2$$

$$B(N,k) \le B(N-1,k) + B(N-1,k-1) \le \sum_{i=0}^{k-1} C_{N-1}^i + \sum_{i=0}^{k-2} C_{N-1}^i$$

$$= 1 + \sum_{i=1}^{k-1} C_N^i + \sum_{i=1}^{k-1} C_{N-1}^{i-1} = 1 + \sum_{i=1}^{k-1} (C_{N-1}^i + C_{N-1}^{i-1}) = 1 + \sum_{i=1}^{k-1} C_N^i = \sum_{i=0}^{k-1} C_N^i$$

Размерность Вапника-Червоненкиса

$d_{VC}(H)$ для набора гипотез H, это наибольшее значение N, для которого $m_H(N) = 2^N$.

$$d_{VC}(H) = \mathrm{k} - 1, \text{где } \mathrm{k} \ - \text{точка поломки}$$

Функция роста и VC-размерность

$$m_H(N) \le \sum_{i=0}^{k-1} C_N^i$$

$$m_H(N) \le \sum_{i=0}^{d_{VC}} C_N^i \le N^{d_{VC}} + 1$$

VC-размерность для перцептрона

Для размерности d, $d_{VC} = d + 1$

$$\mathbf{X} = \begin{bmatrix} -\mathbf{x}_1^{\mathbf{T}}- \\ -\mathbf{x}_2^{\mathbf{T}}- \\ -\mathbf{x}_3^{\mathbf{T}}- \\ \vdots \\ -\mathbf{x}_4^{\mathbf{T}}- \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 \\ & & & \vdots & & \\ 1 & 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

$$\operatorname{sign}(\mathbf{Xw}) = \mathbf{y}$$

$$\mathbf{Xw} = \mathbf{y}$$

$$\mathbf{w} = \mathbf{X}^{-1}\mathbf{y}$$

$\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3} \dots, \mathbf{x_{d+1}}, \mathbf{x_{d+2}}$

$$\mathbf{x}_j = \sum_{i \neq j} \mathbf{x}_i a_i \qquad \mathrm{w}^{\mathrm{T}}\mathrm{x_j} = \sum_{i \neq j} \mathrm{w}^{\mathrm{T}}\mathrm{x_i} a_i \qquad y_i = \operatorname{sign}(a_i) \qquad y_j = -1$$

Сколько нужно данных?

$$P[|E_{in}(h) - E_{out}(h)| > \epsilon] \le m_H(2N)4e^{-\epsilon^{\frac{1}{8}N}}$$

$$\approx N^{d_{VC}}\ e^{-N}$$

$$N \ge 10\ d_{VC}$$

Валидация

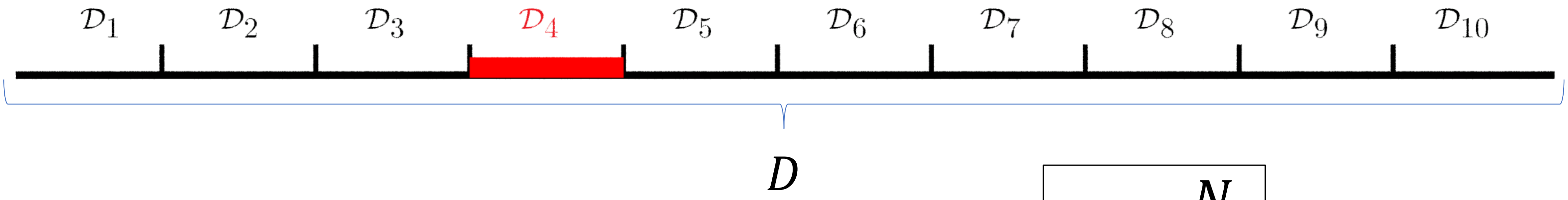
$$(\mathbf{x_1}, \boldsymbol{y_1}), (\mathbf{x_2}, \boldsymbol{y_2}), \dots, (\mathbf{x_K}, \boldsymbol{y_K}) \in \boldsymbol{D_{val}}$$

$$E_{val}(h) = \frac{1}{K}\sum_{i=1}^{K} e\big(h(\mathrm{x}_i), f(\mathrm{x}_i)\big)$$

$$P[|E_{val}(h) - E_{out}(h)| > \epsilon] \le 2e^{-\epsilon^2 N}$$

$$\boxed{K = \frac{N}{5}}$$

Кросс-валидация



$$\boxed{K = \frac{N}{10}}$$

Train-Val-Test

- Обучаем алгоритм на **train**
- Оптимизируем алгоритм (гиперпараметры) на **val (cross-val)**
- Проверяем алгоритм на **test**