Linear classifiers

Perceptron

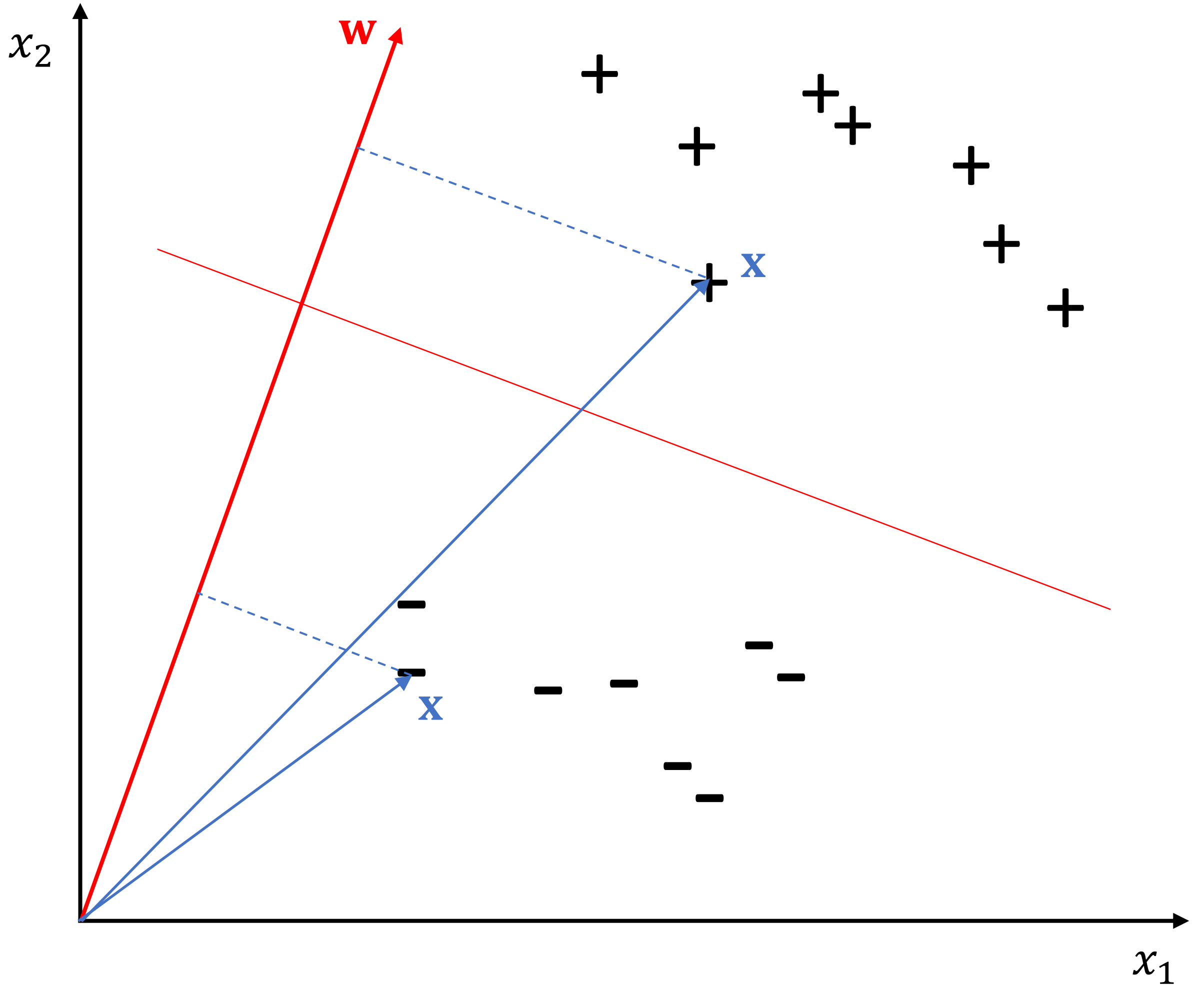
$$y \in \{-1,1\}$$

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

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

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

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



Perceptron training

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

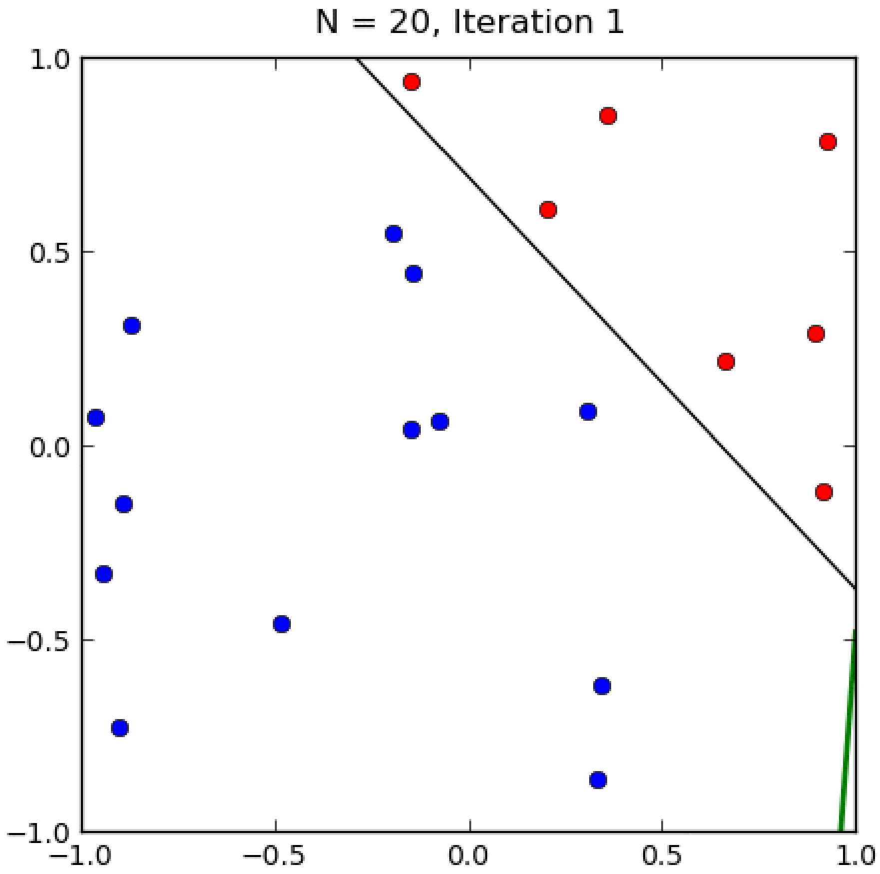
Data: $(\mathbf{x_1}, \boldsymbol{y_1}), (\mathbf{x_2}, \boldsymbol{y_2}), \dots, (\mathbf{x_N}, \boldsymbol{y_N})$

Algorithm:

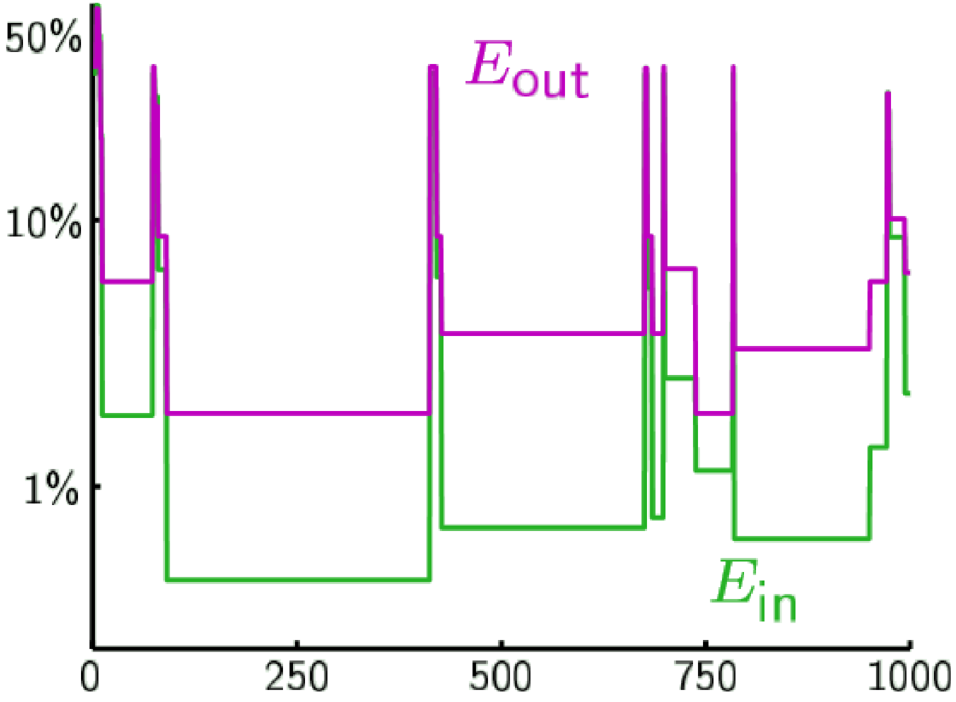
Start with the random **w**

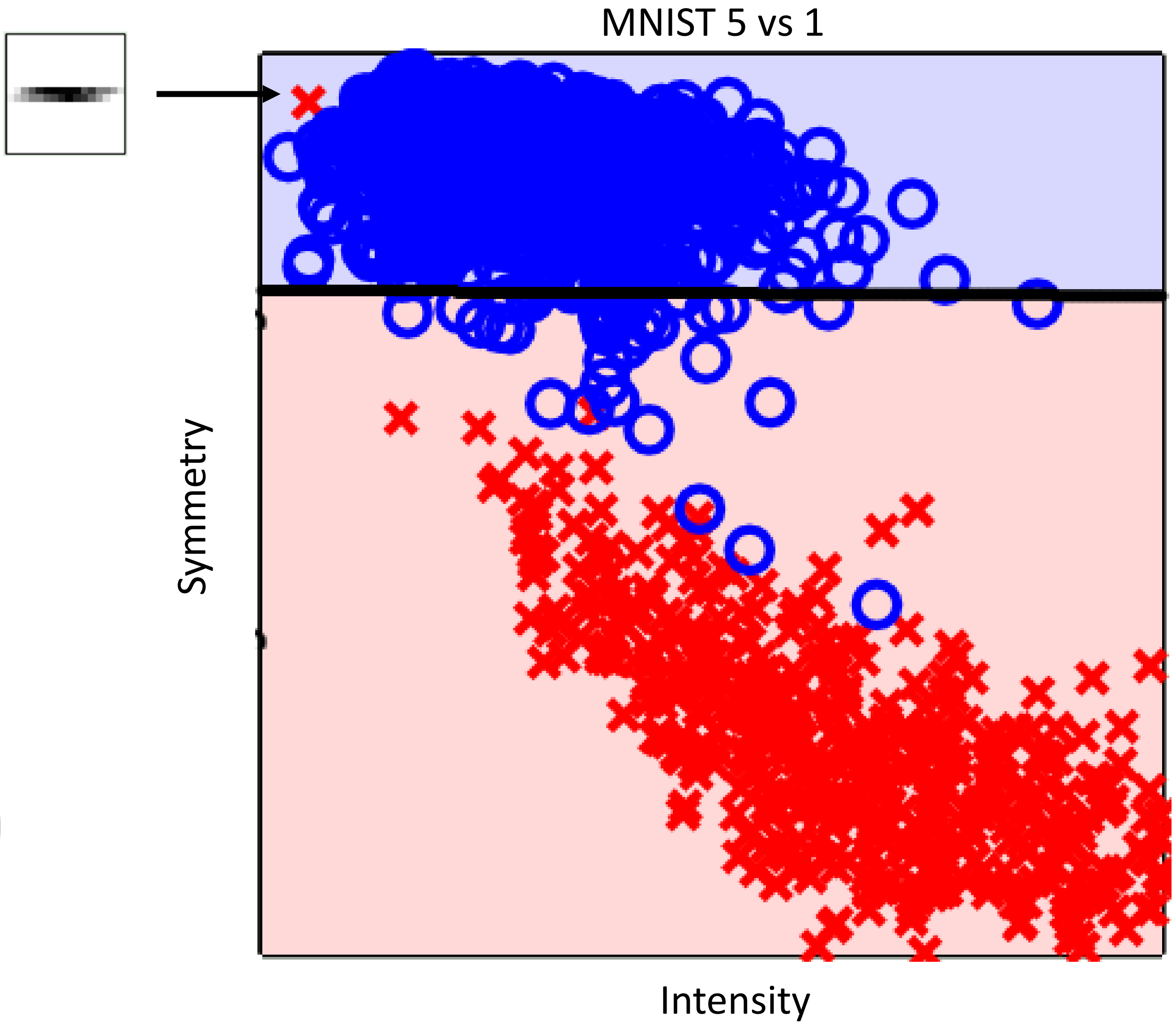
Find such $\mathbf{x_i}$, that $h(\mathbf{x_i}) \neq \mathrm{y_i}$

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

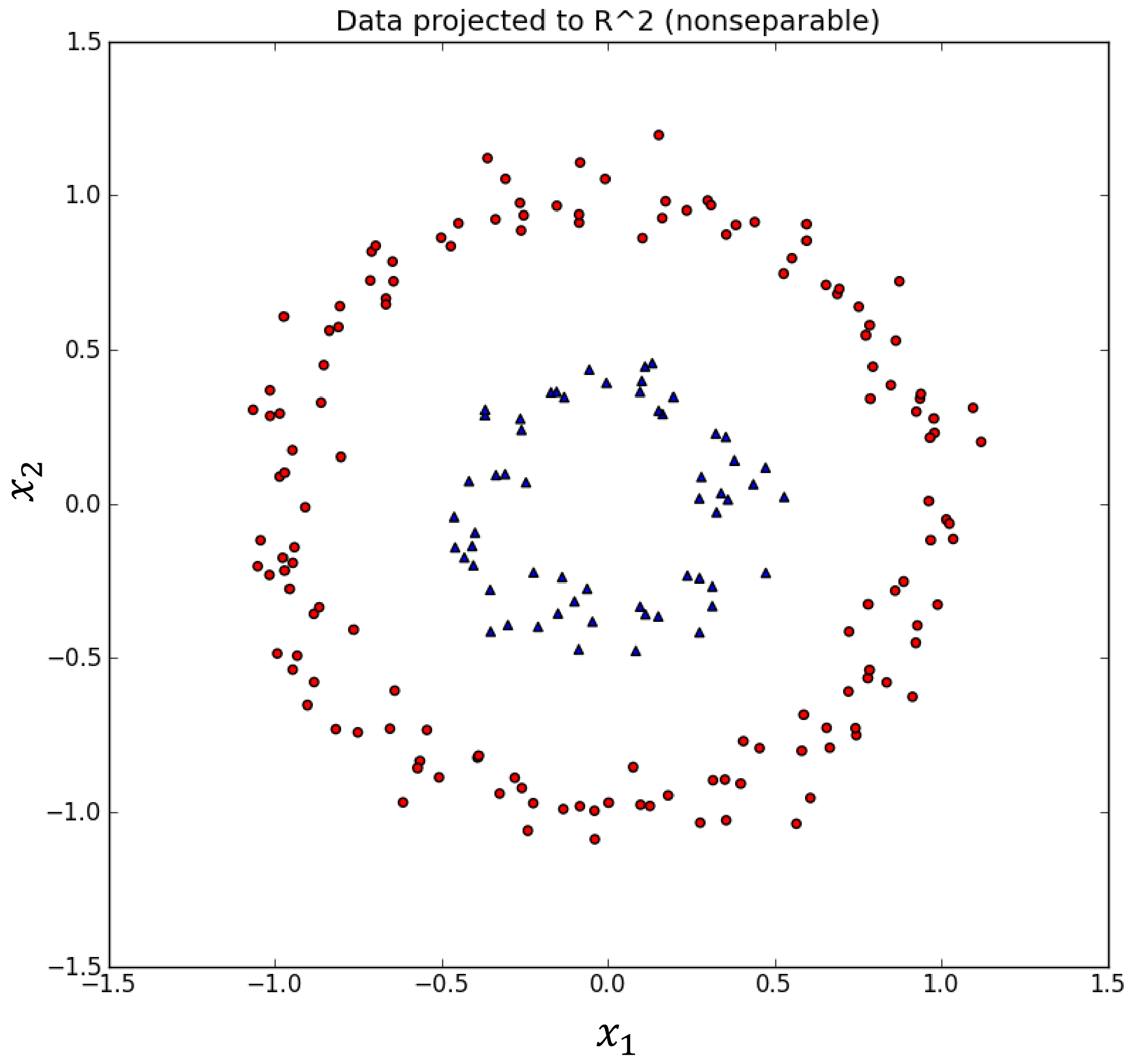


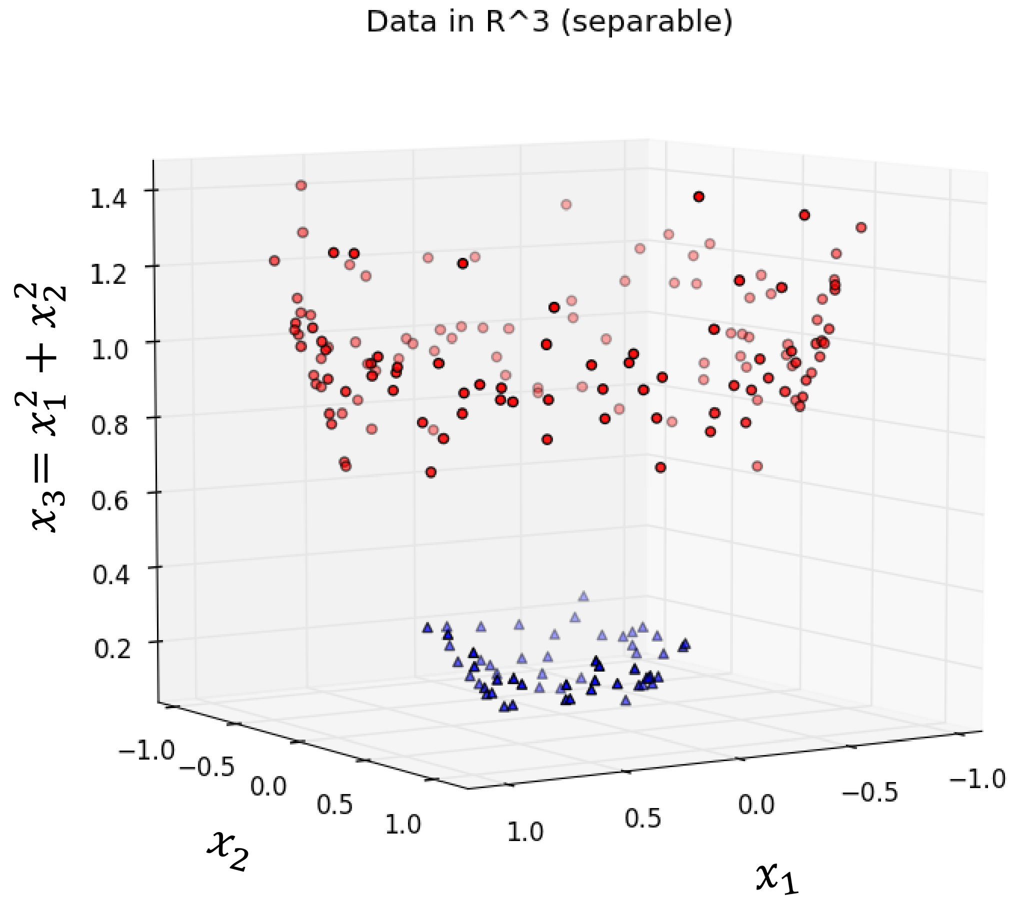
Pocket algorithm





Feature engineering





Theory of error

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

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

Hoeffding's inequality

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

Sometimes E_{out} is denoted as generalization error

Hoeffding's inequality

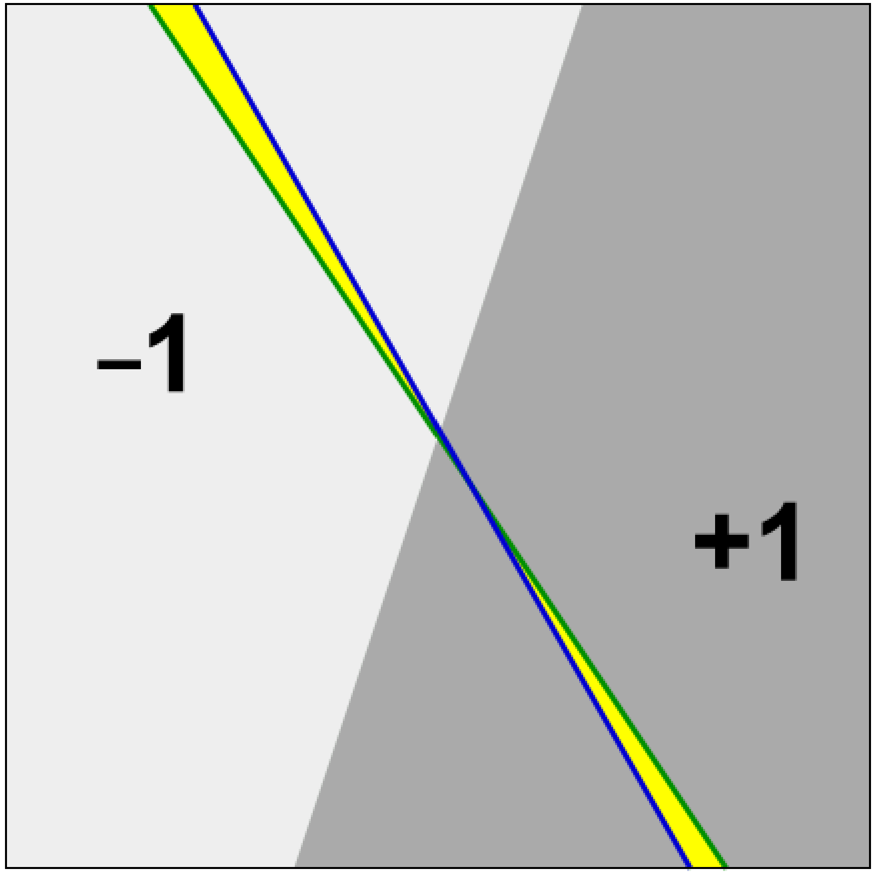
$$P[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2e^{-\epsilon^2 N}$$

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

M is the number of hypothesis

Close hypotheses

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



From hypotheses to dichotomies

- Hypothesis: $h\colon \mathrm{X} \rightarrow \{-1, +1\}$
- Dichotomy: $h\colon \{\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_N}\} \rightarrow \{-1, +1\}$
- Maximum number of dichotomies: 2^N

Growth function

$$m_H(N) = \max_{\mathrm{x}_1, \dots, \mathrm{x}_N} |H(\mathbf{x_1}, \dots, \mathbf{x_N})|$$

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

Vapnik–Chervonenkis Inequality

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

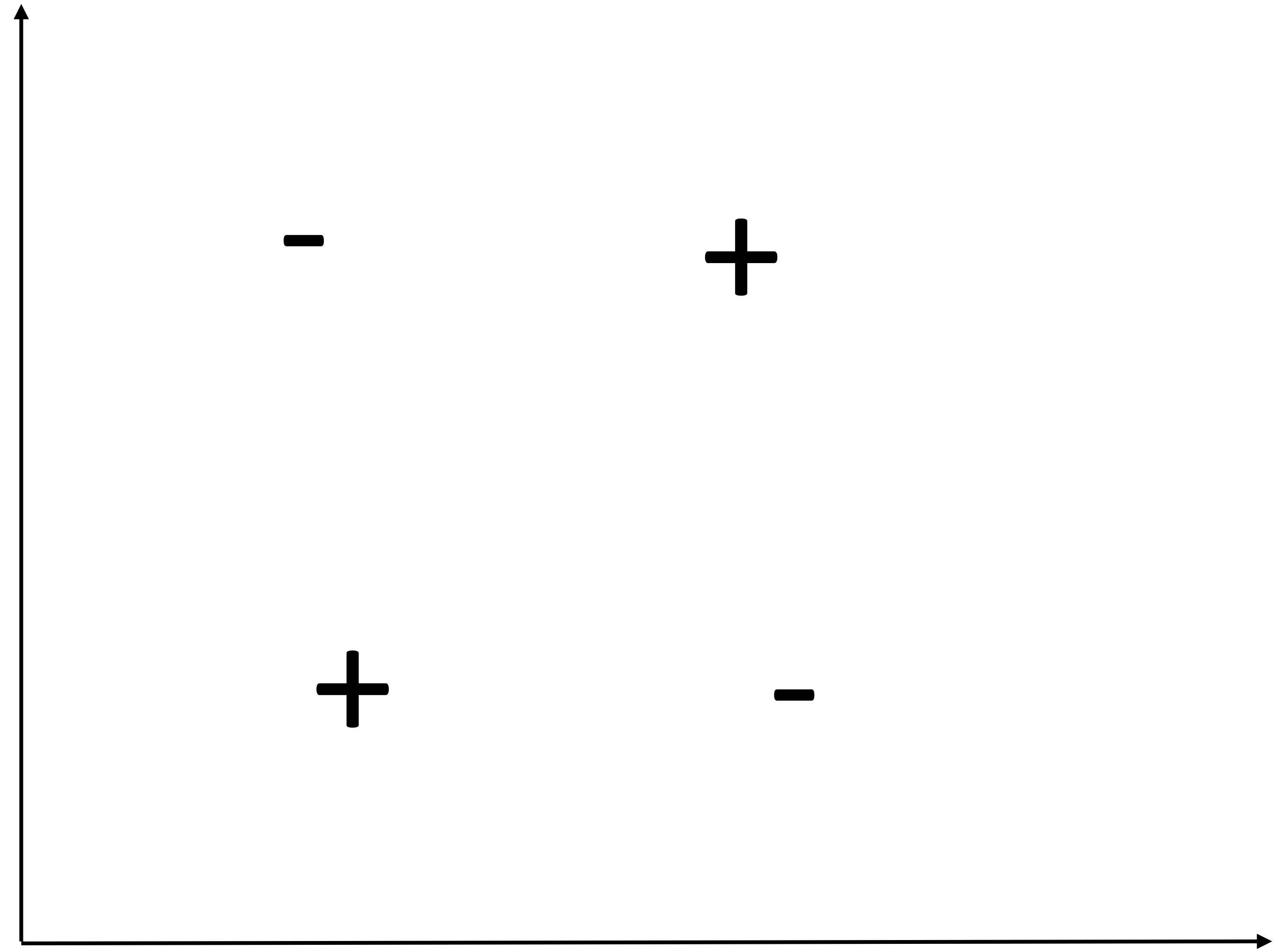
$$\Downarrow \quad \Downarrow \quad \Downarrow$$

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

Breakpoint

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

For 2D perceptron, ***k = 4.***



Proof of polynomiality of a growth function in the presence of a breakpoint

- $B(N, k) = m_H(N)$ with breakpoint k
- $B(N, k) = \alpha + 2\beta$
- $\alpha + \beta \leq B(N-1, k)$
- $\beta \leq B(N-1, k-1)$
- $B(N, k) \leq B(N-1, k) + B(N-1, k-1)$
- Let's prove that $B(N, k) \leq \sum_{i=0}^{k-1} C_N^i$

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\ldots$	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$	
	$+1$	$+1$	$\ldots$	$+1$	$+1$	
	-1	$+1$	$\ldots$	$+1$	-1	
α	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	1 class for x_N
	$+1$	-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	$+1$	2 classes for x_N
	$+1$	-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	

Proof

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

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

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

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

VC-dimension

$d_{VC}(H)$ for hypotheses type *H*, is the maximum number N, such that $m_H(N) = 2^N$.

$$d_{VC}(H) = \mathrm{k} - 1, \text{ where k } - \text{is the breakpoint.}$$

Growth function and VC-dimension

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

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

VC-dimension of perceptron

If d is the space dimensionality, $d_{VC} = d + 1$

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

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

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

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

$$\mathbf{X} = \begin{bmatrix} - \mathbf{x_1^T} - \\ - \mathbf{x_2^T} - \\ - \mathbf{x_3^T} - \\ \vdots \\ - \mathbf{x}_{\boldsymbol{k}}^{\mathbf{T}} - \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 \\ & & & \vdots & & \\ 1 & 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

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

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

Sufficient data

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

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

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

Validation

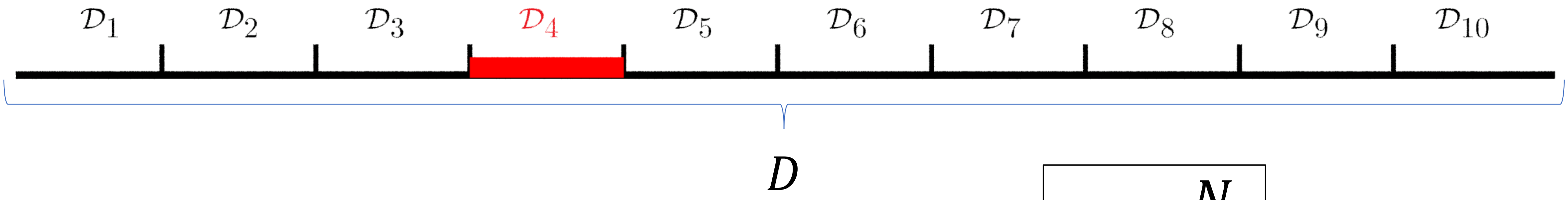
$$(\mathbf{x_1}, \boldsymbol{y_1}), (\mathbf{x_2}, \boldsymbol{y_2}), \ldots, (\mathbf{x_K}, \boldsymbol{y_K}) \in \boldsymbol{D_{val}}$$

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

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

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

Cross-validation



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

Train-Val-Test

- Train the algorithm on the **train** dataset.

- Optimize hyperparameters on **val (cross-val).**

- Check the final performance on **test.**