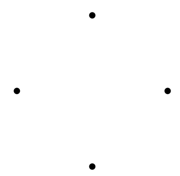
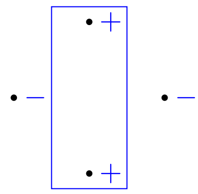
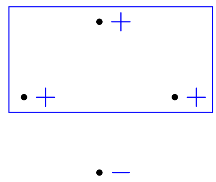
Notes 5: VC dimension

1. Vapnik–Chervonenkis dimension

Related to mistake lower bounds in Online Learning
Usually an integer, telling us how expressive a concept class $\mathcal{C}$ is
Given concept class $\mathcal{C}$ over instance space X, subset $S \subseteq X$ is **shattered by** $\mathcal{C}$ if all "dichotomies" of S can be induced by $\mathcal{C}$, i.e.:

$$\forall T \subseteq S, \exists c \in \mathcal{C} \text{ s.t. } c \cap S = T$$

$X =$ the plane $= \mathbb{R}^2$ $\quad \mathcal{C} =$ axis-aligned rectangles

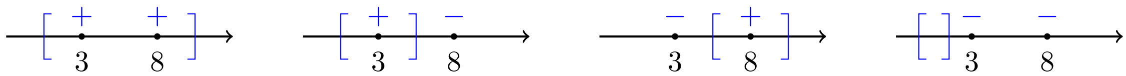
$S =$  has all dichotomies such as  and 

VCDim($\mathcal{C}$) is the size of the largest subset $S \subseteq X$ shattered by $\mathcal{C}$
VCDim($\mathcal{C}$) $= d$ if and only if

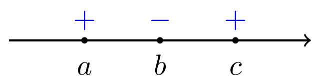
(1) some subset $S \subseteq X$ with $|S| = d$ is shattered by $\mathcal{C}$; and
(2) all subsets of size $d + 1$ is not shattered by $\mathcal{C}$

VCDim($\mathcal{C}$) can be ∞

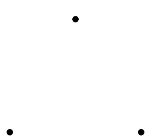
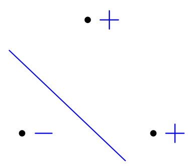
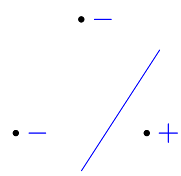
Example: Closed intervals of the real line
$X = \mathbb{R}$ $\quad \mathcal{C} =$ closed intervals $= \{[a, b] \mid a, b \in \mathbb{R}\}$ where $[a, b] = \{x \in \mathbb{R} \mid a \leqslant x \text{ and } x \leqslant b\}$
Every two points (e.g. 3 and 8) can be shattered $\quad \Longrightarrow \quad$ VCDim($\mathcal{C}$) $\geqslant 2$



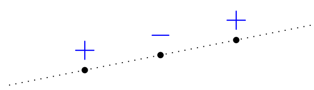
No three points ($a < b < c$) can be shattered $\quad \Longrightarrow \quad$ VCDim($\mathcal{C}$) $\leqslant 2$



Example: Halfspaces in the plane $\quad X = \mathbb{R}^2 \quad \mathcal{C} =$ LTF
Any three non-collinear points can be shattered $\quad \Longrightarrow \quad$ VCDim($\mathcal{C}$) $\geqslant 3$

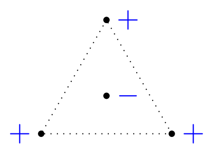
$S =$  has all dichotomies such as  and 

No four points can be shattered $\quad \Longrightarrow \quad$ VCDim($\mathcal{C}$) $\leqslant 3$
Case 1: contains three collinear points

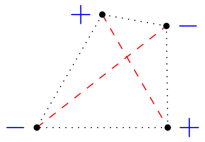


Case 2: No three points collinear

Case 2a: Some point inside the triangle formed by three other points



Case 2b: Four points form a convex quadrilateral $\quad \Longleftrightarrow \quad$ the two diagonals cross



endpoints of two diagonals get different labels

Example: $X = \{0,1\}^n$ $\qquad \mathcal{C} = \{\text{monotone conjunction}\}$ $\qquad$ e.g. $c(x) = x_2 \wedge x_5 \wedge x_7 \in \mathcal{C}$

$\mathrm{VCDim}(\mathcal{C}) \geqslant n$: $\qquad S = \{a_j = \text{vector with 0 at position } j \text{ and 1 everywhere else} \mid 1 \leqslant j \leqslant n\}$

$$\text{e.g. } n = 4 \qquad S = \begin{Bmatrix} 0111, \\ 1011, \\ 1101, \\ 1110 \end{Bmatrix}, \qquad T = \begin{Bmatrix} 0111, \\ \\ 1101, \\ 1110 \end{Bmatrix} \text{ induced by } c(x) = x_2$$

Every subset $T \subseteq S$ is induced by $c \in \mathcal{C}$ containing precisely variables x_j s.t. $a_j \notin T$

$\mathrm{VCDim}(\mathcal{C}) \leqslant n$: $\qquad$ because $|\mathcal{C}| = 2^n$ and

Observation: $\qquad \mathrm{VCDim}(\mathcal{C}) \geqslant d$ implies $|\mathcal{C}| \geqslant 2^d$

---

2. Online Mistake Lowerbounds from VC dimension

Claim 1. *Any deterministic algorithm for learning $\mathcal{C}$ makes $\geqslant \mathrm{VCDim}(\mathcal{C})$ mistakes on some sample sequence*

Proof. $S = \{x^1, \ldots, x^d\}$ be shattered set of size $d = \mathrm{VCDim}(\mathcal{C})$
Instance sequence is $x^1, \ldots, x^d$
On sample x^i, algorithm predicts $b_i \in \{0,1\}$
Can find $c \in \mathcal{C}$ s.t. $c(x^i) = \bar{b}_i$ for all $1 \leqslant i \leqslant n$ $\qquad$ (opposite of all predictions) $\square$

Claim 2. *Some fixed sample sequence causes every randomized algorithm for learning $\mathcal{C}$ to make $\geqslant \mathrm{VCDim}(\mathcal{C})/2$ mistakes in expectation*

Previous claim follows from the next claim (via Yao's minimax principle, not covered in this course)

Claim 3. *Some distribution of random sample sequences causes every deterministic algorithm for learning $\mathcal{C}$ to make $\geqslant \mathrm{VCDim}(\mathcal{C})/2$ mistakes in expectation*

Proof. $S = \{x^1, \ldots, x^d\}$ be shattered set of size $d = \mathrm{VCDim}(\mathcal{C})$
Sample sequence is $(x^1, y^1), \ldots, (x^d, y^d)$, where $y^1, \ldots, y^d$ are uniformly random bits
Any algorithm predicting d uniformly random bits makes $d/2$ mistakes in expectation
For every choice of random bits $y^1, \ldots, y^d$, some $c \in \mathcal{C}$ correctly labels all instances $x^1, \ldots, x^d$ $\square$