MATH466/MATH766
Math of machine learning
04/02 Lecture 20 Vapnik-Chervonenkis dimension

References:

- ch7.9 of The Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani and Jerome Friedman
- https://svivek.com/teaching/lectures/slides/colt/vc-dimensions.pdf

Todays contents:

- VC dimension

Important concepts:

- cardinality
- shattering
- VC dimension

Recommend reading:

-

Goal: quatify the "capacity" of a function space

Intuition: think about binary classification

given n points in $\mathbb{R}^2$ $X = \{x_1, \dots, x_n\}$

dots scattered on a piece of paper

binary classification $X = X_1 \cup X_2$, $X_1 \cap X_2 = \phi$

draw boundaries to separate the dots into 2 groups

model: the way you can separate the dots

boundaries that you are allowed to draw

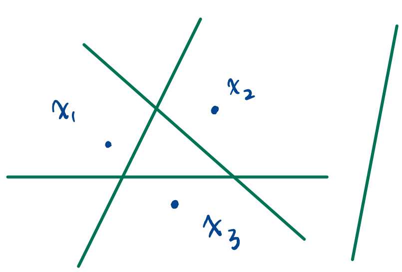
all possibilities of the classification: 2^X, power set of X

number of all possible cases $2^{\text{card}(X)}$ (or $2^{|X|}$)

Q: How many dots you can perfectly separate

in every possible way with you model

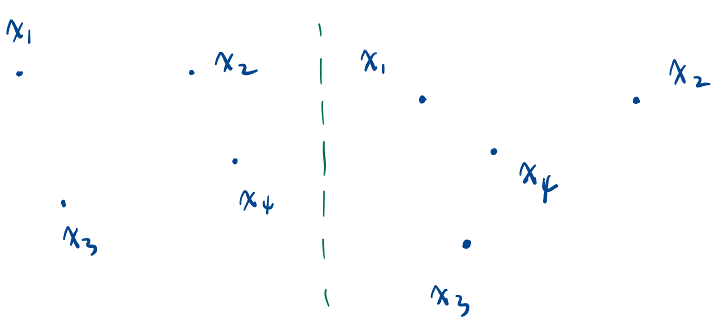
e.g. binary classification in $\mathbb{R}^2$ with linear model



$\exists$ 3 points

that can be separated

in any desired 2 groups



$\forall$ 4 points

$\exists$ some 2 groups

cannot be separated by

straight lines.

Formal definition :

Shattering : Suppose X is a finite set with n elements

i.e. $\mathrm{card}(X) = n$

$\mathcal{A}$ is a class of sets.

The class $\mathcal{A}$ shatters the set X

if for each subset $X_i \in 2^X$

there exists some element $A \in \mathcal{A}$

such that $X_i = X \cap A$

i.e. $\mathcal{A}$ shatters X if $\{ X \cap A : A \in \mathcal{A} \} = 2^X$

e.g. $\mathcal{A} = \{ A \subseteq \mathbb{R}^2 : A = \{ x \in \mathbb{R}^2 : a^T x + b \geq 0 \} \text{ for some } a, b \}$

if $X = \{ (0,0), (0,1), (1,0) \}$

then $\mathcal{A}$ shatters X

if $X = \{ x_1, x_2, x_3, x_4 \}$

then $\mathcal{A}$ does not shatter X

Shatter coefficient : Let $\mathcal{A}$ be a collection of subsets of $\mathbb{R}^d$

the n-th shutter coefficient of $\mathcal{A}$ is

$$S_{\mathcal{A}}(n) := \max_{X = \{x_1, \dots, x_n\} \in \mathbb{R}^d} \mathrm{card}\left(\{ X \cap A : A \in \mathcal{A} \} \right)$$

VC dimension : The VC dimension of a collection of sets $\mathcal{A}$

is the largest integer n such that $S_{\mathcal{A}}(n) = 2^n$

the VC dimension of $\mathcal{A}$ is the largest cardinality

of sets shattered by $\mathcal{A}$.

Examples:

1. If $\mathcal{A}$ is finite, then $V_{\mathcal{A}} \le \log_2 \text{card}(\mathcal{A})$

2. $\mathcal{A} = \{ A : A = \{ x \in \mathbb{R}^2 : a^T x + b \ge 0 \} \}$. $V_{\mathcal{A}} = 3$

① Find a set of 3 points that is shattered by $\mathcal{A}$.

$X = \{ \underline{0}, e_1, e_2 \} = \{ (0,0), (0,1), (1,0) \}$

For any $X_i \subseteq X$, find $A \in \mathcal{A}$ s.t. $X \cap A = X_i$

(i) $0 \notin X_i$, ϕ. $A_0 = \{ x : -\frac{1}{2} \ge 0 \} = \phi$

$\{e_1\}$. $A_1 = \{ x : e_1^T x - \frac{1}{2} \ge 0 \}$

similar for $\{e_2\}$

$\{e_1, e_2\}$ $A_2 = \{ x : e_1^T x + e_2^T x - \frac{1}{2} \ge 0 \}$

(ii) $0 \in X_i$, $\{0\} = X \setminus \{e_1, e_2\}$

Pick $A = A_2^c = \{ x : -e_1^T x - e_2^T x + \frac{1}{2} \ge 0 \}$

$\{0, e_2\} = X \setminus \{e_1\}$

Pick $A = A_1^c$ similar for $\{0, e_1\}$

$\{0, e_1, e_2\} = X \setminus \phi$

Pick $A = A_0^c = \mathbb{R}^2$

② Prove that for any set containing more than 3 points it cannot be shattered by $\mathcal{A}$

$X = \{ x_1, x_2, \cdots x_k \}$ $(k > 3)$

key observation: (i) if a linear separates X into X_1, X_2 it also separates $\text{conv}(X_1)$ and $\text{conv}(X_2)$

(ii) if $\text{card}(X) > 3$, we can always divide X into X_1, X_2 $\left(\begin{array}{l} X_1 \cup X_2 = X \\ X_1 \cap X_2 = \phi \end{array} \right)$

such that $\text{conv}(X_1) \cap \text{conv}(X_2) \ne \phi$

To show (ii), consider $\sum_{i=1}^{k} \lambda_i x_i = 0$, $\sum_{i=1}^{k} \lambda_i = 0$

k unknowns with 3 equation, homogenous system

$\Rightarrow$ there exists at least one non-zero solution λ^*.

$I_1 = \{ i : \lambda_i^* \geq 0 \}$, $I_2 = \{ i : \lambda_i^* < 0 \}$,

$X_1 = \{ x_i : i \in I_1 \}$, $X_2 = \{ x_i : i \in I_2 \}$, $X_1 \cap X_2 = \phi$

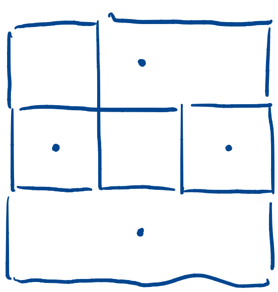
$$\mathrm{conv}(X_1) \cap \mathrm{conv}(X_2) = \frac{1}{\sum_{i \in I_1} \lambda_i^*} \sum_{i \in I_1} \lambda_i^* x_i = \frac{1}{\sum_{i \in I_2} \lambda_i^*} \sum_{i \in I_2} \lambda_i^* x_i$$

HW: generalize to linear models in high-dimension

3. $\mathcal{A}$ is the collection of axis-aligned rectangles in $\mathbb{R}^2$

$V_{\mathcal{A}} = 4$

① $X = \{ (1, 0), (0, 1), (-1, 0), (0, -1) \}$

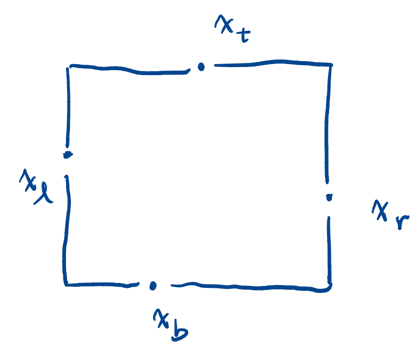


② $X = \{ x_1, \ldots, x_k \}$ $(k > 4)$

consider the "extremal points"

$X_e := \{ x_l, x_r, x_t, x_b \}$

any rectangle that contains X_e must contain X.



4. Think about an example where $\mathrm{card}(\mathcal{A}) = \infty$ but $V_{\mathcal{A}} = 1$

VC dimension of a collection of binary classifier

$\mathcal{H} := \{ h : X \to \{0,1\} \}$

① $\mathcal{A} := \{ A_h := \{ x : h(x) = 1 \} \} \subseteq 2^X$

② $\tilde{\mathcal{A}} := \{ \tilde{A}_h : h \in \mathcal{H} \} \subseteq 2^{X \times \{0,1\}}$

$\tilde{A}_h := \{ (x,y) \in X \times \{0,1\} : h(x) \neq y \}$

$V_{\mathcal{H}} := V_{\mathcal{A}} = V_{\tilde{\mathcal{A}}}$