Optimization Methods in Machine Learning
Lecure 4: Vapnik-Chervonenkis (VC) dimension

Katya Scheinberg

Lehigh University

Spring 2016

Overview

1. Motivation
2. Growth Function
3. Vapnik-Chervonenkis (VC) dimension

The material for this lecture is taken from a short course taught at UT Austin by N. Srebro and K. Scheinberg in 2011.

Motivation

- Recall that the bound on the difference in the expected (R) and empirical ($\hat{R}$) errors: with probability at least $1-\delta$, for all $h \in \mathcal{H}$,

$$\left| R(h) - \hat{R}(h) \right| \leq \sqrt{\frac{\log |\mathcal{H}| + \log \frac{2}{\delta}}{2m}}$$

- Recall that the cardinality of a hypothesis class is difficult to measure, many classes may be infinite in cardinality because of their parameters being continuous.
- We need a more consistent way of measuring complexity of a class of hypothesis.

Growth Function

- We define the number of different behaviors that a hypothesis class has on a specific set of points as the **growth function**.

Definition

We define the growth function $\Pi(\mathcal{H}, S)$ for a hypothesis class $\mathcal{H}$ and a set of points $S = \{x_1, \ldots, x_m\}$. We look at all the possible labelings we can have $y_1, \ldots, y_m$ such that there exists some classifier in the class that actually gives these labelings.
The **growth function**
$\Pi(\mathcal{H}, S) =$ Number of different behaviors(predictions) the class of hypothesis H can generate on a sample S.

Example (4.1)

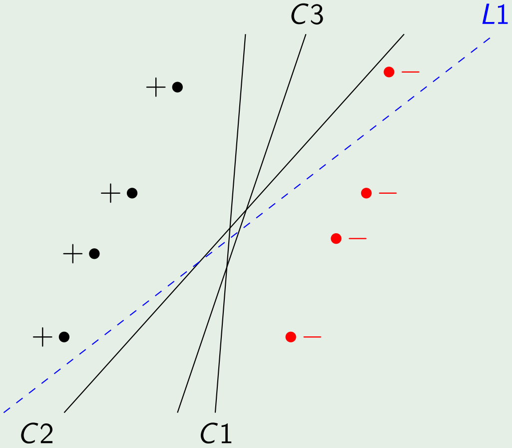


Figure: Linear classifiers with same (C1,C2,C3) and different (L1) behavior on training sample.

Why do we care about the growth function

Lemma

With probability at least $1-\delta$, for all hypotheses $h \in \mathcal{H}$

$$\left|R(h) - \hat{R}(h)\right| \leq \sqrt{\frac{\log \Pi(\mathcal{H}, 2m) + \log \frac{4}{\delta}}{m}}.$$

Growth Function, Simple Bounds

Here, instead of providing labelings to specific data sample, we are looking for the maximum number of growth function for any m data points on given space.
How big can the growth function be for a given hypothesis class $\mathcal{H}$?

- $$\Pi\left(\mathcal{H}, m\right) \leq 2^m$$
- $$\Pi\left(\mathcal{H}, m\right) \leq \left|\mathcal{H}\right|$$

Vapnik-Chervonenkis (VC) dimension

- The VC-dimension of a hypothesis class the maximal number of points for which you can get all possible behaviors.

Definition

VC-dimension($\mathcal{H}$) $\triangleq$maximal number of points m such that $\Pi(\mathcal{H}, m) = 2^m$.

- In terms of the growth function, we can just write the VC-dimension as the maximal m such that $\Pi(\mathcal{H}, m) = 2^m$.
- Being able to achieve any labeling of a given set of points is also known as *shattering* the points.
- We say that if we have m points and we can get all possible behaviors on our hypothesis class for these points then these points are shattered.

The Shatter Lemma

The following lemma connects m, VC-dimension and the growth function.

Lemma

$$\Pi\left(\mathcal{H}, m\right) \leq \sum_{i=0}^{VC\text{-}Dim(\mathcal{H})} \left(\begin{array}{c} m \\ i \end{array} \right) \leq \left(\frac{e \cdot m}{D}\right)^{D} \overset{(D \geq 3)}{\leq} m^{D}$$

where D is VC-dimension$(\mathcal{H})$.

The Deviation bounds using VC-dimension:

Lemma

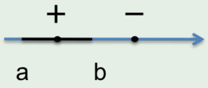
With probability at least $1-\delta$, for all hypotheses $h \in \mathcal{H}$

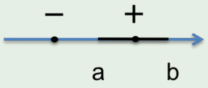
$$\left|R\left(h\right) - \hat{R}\left(h\right)\right| \leq \sqrt{\frac{D\log\left(2m\right) + \log\frac{4}{\delta}}{m}}.$$

Vapnik-Chervonenkis (VC) dimension

Example (4.2)

Let us consider a hypothesis $h = \{x \in [a, b]\}$, which labels x positive if x is in the interval [a,b], negative otherwise.
For the case of two points, we have two labelings:





However, for the case of three points, there exists some labelings that can not be generated, one case is presented below:

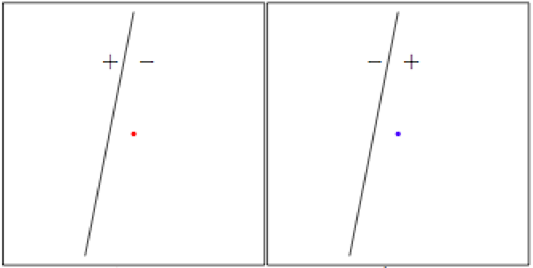
So the VC dimension for this hypothesis is 2, note that it is the same as the number of parameters of the hypothesis.

Vapnik-Chervonenkis (VC) dimension

Let us consider a particular example of linear separators in $\mathbb{R}^2$.

Example (4.3)

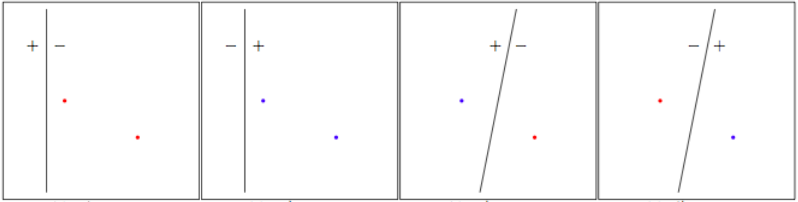
If we have only a single point, we can either label it positive or negative, hence we can label it in all 2^1 different ways. For each such labeling we can find a linear predictor that is consistent with it.



(a) 1^{st} separator (b) 2^{nd} separator

Example (4.4)

For the case of 2 points, there are $4 = 2^2$ kinds of labeling. For each such labeling we can find a linear predictor that is consistent with it.

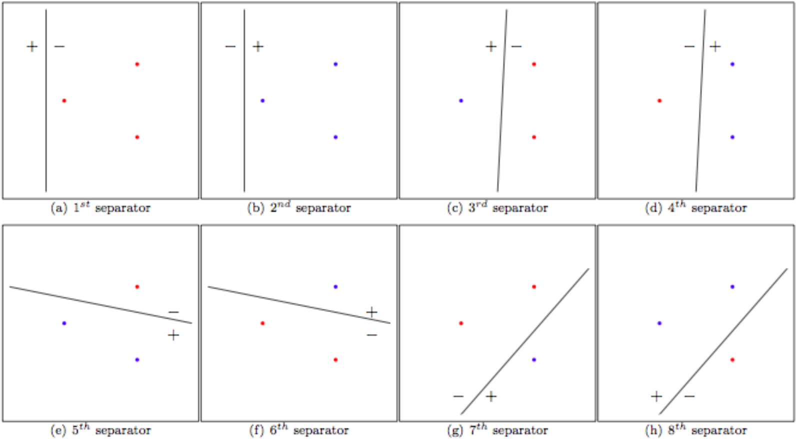


(a) 1^{st} separator (b) 2^{nd} separator (c) 3^{rd} separator (d) 4^{th} separator

Vapnik-Chervonenkis (VC) dimension

Example (4.5)

For the case of 3 points, there are $8 = 2^3$ kinds of labeling. For each such labeling we can find a linear predictor that is consistent with it.



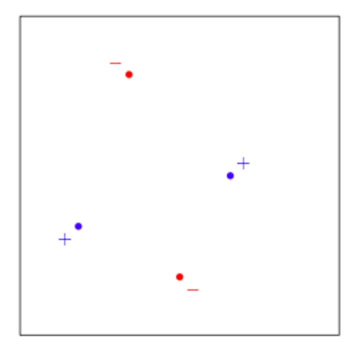
(a) 1^{st} separator (b) 2^{nd} separator (c) 3^{rd} separator (d) 4^{th} separator

(e) 5^{th} separator (f) 6^{th} separator (g) 7^{th} separator (h) 8^{th} separator

Vapnik-Chervonenkis (VC) dimension

Example (4.6)

However, things are a little different with the case of 4 points. For the case of 4 points, there are $2^4 - 2 = 14$ kinds of labeling. As the usual 2^m number of labelings, this time there are two labeling that is not achievable by linear classifiers. Below presents one of them:



Vapnik-Chervonenkis (VC) dimension

From the previous examples, we see that:

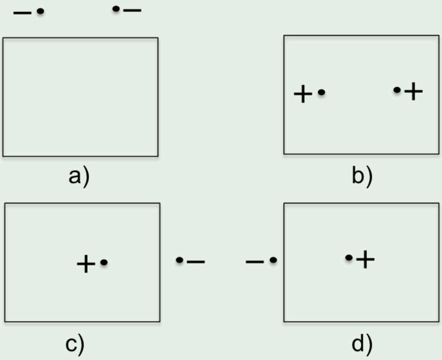
- For $m = 1$, $\Pi(\mathcal{H}, m) = 2^1 = 2 = 2^m$;
- For $m = 2$, $\Pi(\mathcal{H}, m) = 2^2 = 4 = 2^m$;
- For $m = 3$, $\Pi(\mathcal{H}, m) = 2^3 = 8 = 2^m$;
- For $m = 4$, $\Pi(\mathcal{H}, m) = 14 = 2^4 - 2 \neq 2^m$.

Therefore, VC-dimension $(\mathcal{H}) = 3$.

Vapnik-Chervonenkis (VC) dimension

Example (4.7)

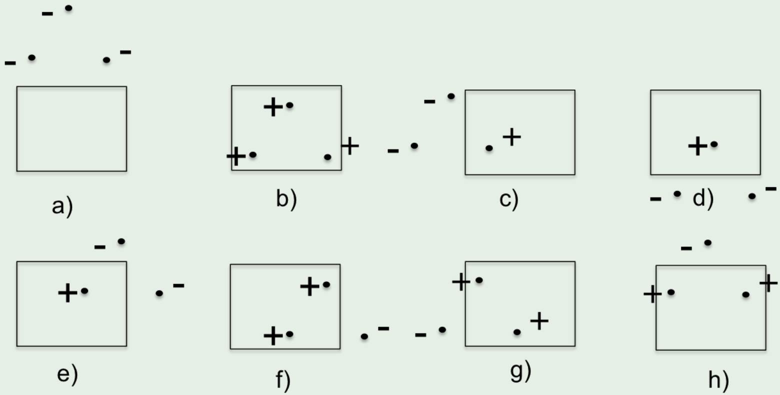
Now we look at another example, where the hypothesis labels the point inside the rectangle decided by the two points (a,b), (c,d) positive, and otherwise negative. The case for two points:



Vapnik-Chervonenkis (VC) dimension

Example (4.8)

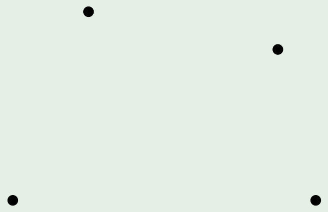
The case for three points:



Vapnik-Chervonenkis (VC) dimension

Example (4.9)

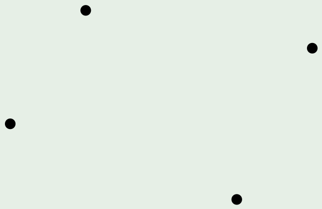
The case for four points is a little different; It is not possible to produce all the labeling for certain situations, one of them is presented below:



Vapnik-Chervonenkis (VC) dimension

Example (4.10)

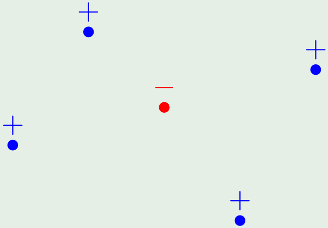
On the other hand this configuration can be shattered:



Vapnik-Chervonenkis (VC) dimension

Example (4.11)

But not this one:



Vapnik-Chervonenkis (VC) dimension

From the previous examples, we see that:

- For $m = 1$, $\Pi(\mathcal{H}, m) = 2^1 = 2 = 2^m$;
- For $m = 2$, $\Pi(\mathcal{H}, m) = 2^2 = 4 = 2^m$;
- For $m = 3$, $\Pi(\mathcal{H}, m) = 2^3 = 8 = 2^m$;
- For $m = 4$, $\Pi(\mathcal{H}, m) = 2^4 = 16 = 2^m$.
- For $m = 5$, $\Pi(\mathcal{H}, m) < 2^5 = 2^m$.

Therefore, VC-dimension $(\mathcal{H}) = 4$.