Lecture 4: The Vapnik-Chervonenkis Dimension

CSE 427: Machine Learning

Md Zahidul Hasan

Lecturer, Computer Science
BRAC University

Spring 2023

Infinite Hypothesis Classes Can Be Learnable Too

In the last lecture, we saw that the class of axis-aligned rectangles can be learned even though the class is not finite. We present another example of threshold functions that shows that infinite hypothesis classes can be learned sometimes.

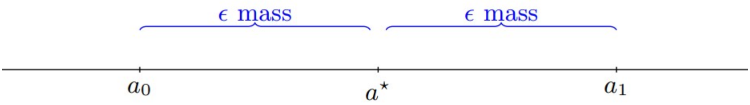
Example

Let $\mathcal{H}$ be the set of all threshold functions over the real line. A threshold function looks like this: $h_a(x) = \mathbb{I}_{x \leq a}$. So, $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$. Clearly, $\mathcal{H}$ is infinite. But the following lemma will prove that $\mathcal{H}$ is PAC-learnable.

Proof: Let's say, the concept we are trying to learn has threshold $a*$. We are given a sample of m points. Let's say, b_0 is the rightmost point with label 1, and b_1 is the leftmost point with label 0. Let's choose two points a_0 and a_1 on respectively the left and right side of $a*$ so that the probability of finding a point from $a*$ to these points according to the distribution is ϵ.

Threshold Functions are PAC-learnable

$$\underset{x \sim \mathcal{D}_x}{\mathbb{P}}\left[x \in (a_0, a^\star)\right] = \underset{x \sim \mathcal{D}_x}{\mathbb{P}}\left[x \in (a^\star, a_1)\right] = \epsilon.$$



So our ERM algorithm will return a hypothesis h_S that has its threshold between b_0 and b_1. Let's say C is the event that $R_D(h_S) \leq \epsilon$, A be the event that $b_0 \geq a_0$ and B be the event that $b_1 \leq a_1$. If, A and B happen together, C definitely happens.

$A \wedge B \implies C$

$\implies Pr(A \wedge B) \leq Pr(C)$.

$\implies Pr(\overline{A \wedge B}) \geq Pr(\overline{C})$

$\implies Pr(\overline{A} \vee \overline{B}) \geq Pr(\overline{C})$.

$\implies Pr(R_D(h_S) > \epsilon) \leq Pr((b_0 < a_0) \vee (b_1 > a_1))$.

Threshold Functions contd.

Using the union bound we get,
$Pr(R_D(h_S) > \epsilon) \leq Pr(b_0 < a_0) + Pr(b_1 > a_1)$.
$b_0 < a_0$ happens when none of the m points fall in the range $(a_0, a*)$ and $b_1 > a_1$ happens when none of the m points fall in the range $(a*, a_1)$.
So, $Pr[b_0 < a_0] = (1-\epsilon)^m$. Similarly, $Pr[b_1 > a_1] = (1-\epsilon)^m$.
$Pr(R_D(h_S) > \epsilon) \leq 2(1-\epsilon)^m$
Since $1 + x \leq e^x$ for all $x \in \mathbb{R}$, we have that,
$Pr(R_D(h_S) > \epsilon) \leq 2e^{-\epsilon m} = \delta$.
Solving for m gives $m = \frac{1}{\epsilon} \log \frac{2}{\delta}$ which concludes the proof. $\square$
We will gradually learn that the size of a hypothesis class may not be a good characterization of its learnability. Rather the Vapnik-Chervonenkis dimension is much more insightful. In the later slides, we will discuss this concept in depth.

Restrictions and Dichotomies

Restriction of $\mathcal{H}$ to S

Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0, 1\}$. And let $S = (x_1, x_2, \cdots, x_m)$ be a sample of size m derived from $\mathcal{X}$. The restriction of $\mathcal{H}$ to S is defined to be a set of unique vectors as follows:

$$\mathcal{H}_S = \{((h(x_1), h(x_2), \cdots, h(x_m)) : h \in \mathcal{H}\}$$

Example

Let's say, $\mathcal{H}$ is a set of threshold hypotheses.
$\mathcal{H} = (h_1 = \mathbb{I}_{x \le 4}, h_2 = \mathbb{I}_{x \le 11}, h_3 = \mathbb{I}_{x \le 11.5})$. Let's say the sample is $S = (x_1 = 2, x_2 = 3, x_3 = 5, x_4 = 12)$. Then $\mathcal{H}_S =$
$\{((h_1(x_1), h_1(x_2), h_1(x_3), h_1(x_4)), ((h_2(x_1), h_2(x_2), h_2(x_3), h_2(x_4))$
$((h_3(x_1), h_3(x_2), h_3(x_3), h_3(x_4))\}$
$= \{(1, 1, 0, 0), (1, 1, 1, 0), (1, 1, 1, 0)\}$. After removing duplicates, we get, $\mathcal{H}_S = \{(1, 1, 0, 0), (1, 1, 1, 0)\}$. Each of these unique vectors is called a Dichotomy.

Shattering

Shattering

When the restriction of a hypothesis class $\mathcal{H}$ is a set of all possible vectors of length m, then $\mathcal{H}$ is said to shatter the training set S. In the case of binary classification, $\mathcal{H}$ shatters S if and only if $|\mathcal{H}_S| = 2^{|S|}$.

Example

Let's say $\mathcal{H}$ is the set of all interval functions. In other words, a hypothesis in $\mathcal{H}$ may look like $h_{a,b}(x) = \mathbb{I}_{a \leq x \leq b}$ where $a, b \in \mathbb{R}$. Let's take two sets, $S_1 = (2, 3)$ and $S_2 = (3, 4, 7)$. We will see that $\mathcal{H}$ shatters S_1 but cannot shatter S_2. If we pick $h_{4,5}(x)$, then S_1 gives the vector $(0, 0)$. If we pick $h_{0,5}(x)$, then S_1 gives the vector $(1, 1)$. If we pick $h_{1,2.5}(x)$, then S_1 gives the vector $(1, 0)$. If we pick $h_{2.88,11}(x)$, then S_1 gives the vector $(1, 0)$. So, all possible vectors of length 2 have been achieved. But no matter what hypothesis we choose, S_2 can never give $(1, 0, 1)$.

Shattering = No Restriction

When a training set is shattered by $\mathcal{H}$, the training set is not restricted at all. The restriction of $\mathcal{H}$ to S is maximum when $|\mathcal{H}_S| = 1$. When the training set isn't restricted, an adversary can choose a distribution where the true error is minimized by a function handpicked by the adversary. Imagine, a doctor is trying to guess the lifestyle of a patient. If they don't have any context about the patient and still say "Maybe you aren't physically active", there is a good chance that it's a wrong remark. But if you tell the doctor that the patient has diabetes and if they can say "Maybe you eat a lot of carbohydrates", then there's a good chance that what he is saying is true.

No-Free-Lunch Revisited

If $\mathcal{H}$ is a hypothesis class from $\mathcal{X}$ to $\{0, 1\}$. Assume there exists a set C of size $2m$ that's shattered by $\mathcal{H}$. Then, for any learning algorithm $\mathcal{A}$, there exists a distribution that allows a training sample S of size m to be chosen so that, $Pr[R_D(h_S) \geq \frac{1}{8}] \geq \frac{1}{7}$.

VC Dimension

VC-dimension

The VC-dimension of a hypothesis class $\mathcal{H}$ with respect to a domain $\mathcal{X}$ is the size of the largest set in $\mathcal{X}$ that can be shattered by $\mathcal{H}$. It's written as $VCdim(\mathcal{H})$.

Theorem

If $VCdim(\mathcal{H}) = \infty$, then $\mathcal{H}$ is not PAC-learnable.

Proof: Since the VC-dimension is infinite, for any m, there exists a shattered set of size $n \geq 2m$. So, by the no-free-lunch theorem, $\mathcal{H}$ can't be learned with any training set of size m. So, $\mathcal{H}$ is not PAC-learnable. $\square$

Example

Threshold Functions: Let's take $S = \{1\}$. If we take $\mathbb{I}_{x \leq 2}$ and $\mathbb{I}_{x \leq 0}$, we can get both 1 and 0. But if we take, $S = \{1, 2\}$, then no threshold function can give $(0, 1)$. So, $VCdim(\mathcal{H}) = 1$.

VC-dimension Examples

Example

Interval Functions: Let's say $\mathcal{H} = \{h_{a,b}(x) = \mathbb{I}_{a \leq x \leq b} : a, b \in \mathbb{R}\}$. Let's say $S = \{p, p+2\}$. $h_{p-2,p-1}(x), h_{p-2,p+1}(x)$, $h_{p+1,p+3}(x)$, $h_{p-2,p+3}(x)$ can give all $2^2 = 4$ dichotomies. But if $S = (a, b, c)$ with $a \leq b \leq c$, then no choice of hypothesis can give $(1, 0, 1)$. Therefore, $VCdim(\mathcal{H}) = 2$.

Example

Axis-aligned Rectangles: $\mathcal{H} = \{h_{a,b,c,d}(x, y) = \mathbb{I}_{a \leq x \leq b \wedge c \leq y \leq d}\}$. It's pretty easy to prove that $VCdim(\mathcal{H}) \geq 4$. We will prove that $VCdim(\mathcal{H}) = 4$. Take any 5 points. There is a highest point a, lowest point b, rightmost point c, and a leftmost point d. There's one point that's still not named. That one cannot be separated from the rest with an axis-aligned rectangle.

VC-dimension of Hyperplanes

Let's say we are trying to classify d dimensional points using hyperplanes. In n-dimension, affine hyperplanes are defined by the following equation:

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b$$

First, let's try to do it for 2-dimensional points. We have to separate them by a line. It's pretty easy to check that we can shatter any sets of points of size less than 4.

(a) (b)

But if the set has 4 points. We can choose 3 points and create a convex hull. The 4th point can fall inside the convex hull as in figure (b) or it can fall outside as in figure (a). In both cases, we cannot realize the illustrated dichotomies. So, $VCdim(\mathcal{H}) = 3$.

VC-dimension of Hyperplanes contd.

In $\mathbb{R}^d$, we can show that we can shatter a set of $d+1$ points. Let's pick the origin as $x_0 = \{0, 0, \cdots, 0\}$. And the unit vectors of the d axes as x_is. So, $x_i = \{0, \cdots, 1, \cdots, 0\}$ will have all components 0 except that the i'th component will be 1. The following hypothesis can yield any dichotomy that's required:

$$h(x) = sign(y_1 x^{(1)} + y_2 x^{(2)} + \cdots + y_d x^{(d)} + \frac{y_0}{2})$$
$$\text{So, } h(x_0) = sign(y_1 x_0^{(1)} + y_2 x_0^{(2)} + \cdots + y_d x_0^{(d)} + \frac{y_0}{2})$$
$$h(x_0) = sign(y_1 0 + y_2 0 + \cdots + y_d 0 + \frac{y_0}{2}) = sign(\frac{y_0}{2}) = y_0$$
$$h(x_i) = sign(y_1 0 + \cdots + y_i 1 + \cdots + y_d 0 + \frac{y_0}{2}) = sign(y_i + \frac{y_0}{2}) = y_i$$

So, $VCdim(\mathcal{H}) \geq d+1$. Now, we will prove that no $d+2$ points in $\mathbb{R}^d$ can be shattered with a $d-$dimensional hyperplane.

Radon's Theorem

Any set of $d+2$ points in $\mathbb{R}^d$ can be partitioned into two subsets X_1 and X_2 so that their convex hulls intersect.

VC-dimension of Hyperplanes contd.

Let's say the $d+2$ points are $\{x_1, x_2, \cdots, x_{d+2}\}$. Consider the two equations:

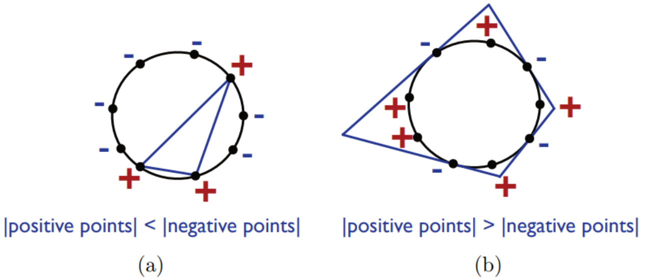
$$\sum_{i=1}^{d+2} a_i = 0, \qquad \sum_{i=1}^{d+2} \alpha_i x_i = 0.$$

There is 1 equation on the left and d on the right and $d+2$ unknowns. So, there are many non-zero solutions to the equations. Let's pick one such solution $\{\beta_1, \beta_2, \cdots, \beta_{d+2}\}$. Let's define $X_1 = \{x_i : 1 \leq i \leq n \wedge \beta_i > 0\}$ and $X_2 = \{x_i : 1 \leq i \leq n \wedge \beta_i < 0\}$. So, $\beta = \sum_{x_j \in X_1} \beta_j = -\sum_{x_j \in X_2} \beta_j$. Then we can rewrite the first equation as:

$$z = \sum_{x_i \in X_1} \frac{\beta_i}{\beta} x_i = \sum_{x_i \in X_2} \frac{-\beta_i}{\beta} x_i$$

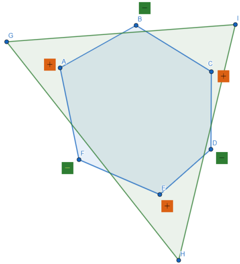
Since, $0 \leq \frac{\beta_i}{\beta} \leq 1$, and $\sum_{x_i \in X_i} \frac{\beta_i}{\beta} = \sum_{x_i \in X_i} \frac{-\beta_i}{\beta} = 1$, we can say that the point z is contained by the convex hulls of both X_1 and X_2. So, the hulls overlap. Therefore, no hyperplane can give us the dichotomy where all points in X_1 have the label $+1$ and all points in X_2 have the label -1.

VC-dimension of Convex Polygons



We will first show that convex $d-$gons can shatter a set of $2d+1$ points. For that, take all the $2d+1$ points on a circle. If the number of negative points is greater than that of the positive points, then choose the positive points and create their convex hulls. Since the number of positive points is less than or equal to d, so a we will only need a polygon of at most d sides. But if the number of positive points is greater, then take tangents to the circles at the negative points. That will do.

VC-dimension of Convex Polygons contd.



Now we will show that no set of $2d+2$ points can be shattered by any n sided convex polygons where $n \leq d$. There can be two cases. **Case 1:** All the points are vertices of their convex hull. In that case, if we color the vertices alternatively like this $\{+1, -1, +1, -1, \cdots\}$, then due to the convexity of the hull, in order to separate every two adjacent vertices, we will need a convex polygon of at least $d+1$ sides. **Case 2:** At least one point p isn't a vertex of the convex hull of the points. In that case, we can't achieve the labeling where p is negative and the rest are positive.

Growth Function

Definition

Growth Function: The growth function of a hypothesis $\mathcal{H}$ with respect to a domain $\mathcal{X}$ and a sample size m is:

$$\mathcal{G}_{\mathcal{H}}(m) = \max_{\mathcal{S} \subseteq \mathcal{X}: |S| = m} |\mathcal{H}_S|$$

The growth function is the maximum number of ways into which m points can be classified by the function class: $\mathcal{H}$. If $VCdim(\mathcal{H}) = d$, then for all $m \leq d$, $\mathcal{G}_{\mathcal{H}}(m) = 2^m$.

Example

Let's go back to the example of threshold functions. Let's say $S = (a, b)$ is a sample of size 2. Without loss of generality, we can assume that $a < b$. We can see that we can never come up with the dichotomy $(0, 1)$. So, the dichotomies we can achieve are $\{(1,1), (1,0), (0,0)\}$. Hence, $\mathcal{G}_{\mathcal{H}}(2) = 3$.

Sauer-Shelah-Perles

Let $\mathcal{H}$ be a hypothesis set with $VCdim(H) = d$. Then for all $m \in \mathbb{N}$, the following holds:

$$\mathcal{G}_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$$

Proof: Since, $\mathcal{G}_{\mathcal{H}}(m) = \max_{\mathcal{S} \subseteq \mathcal{X}:|S|=m} |\mathcal{H}_S|$, we will try to find an upper bound for $|\mathcal{H}_S|$. In fact, it can be showed that,

$$\forall h \in \mathcal{H}, \qquad |\mathcal{H}_S| \leq |A \subseteq S : \mathcal{H} \;\; shatters \;\; A|$$

We will prove this claim using mathematical induction. The case is trivial when $m = 1$. Let's assume that the claim is true for $m > 1$. Let's say, $S^{(m)} = \{s_1, s_1, \cdots, s_m\}$ and $S^{(m-1)} = \{s_1, s_2, \cdots, s_{m-1}\}$. Let's define two sets of dichotomies,
$Y_0 = \{(y_0, y_1, \cdots, y_{m-1}) : (y_0, y_1, \cdots, y_{m-1}, 0) \in \mathcal{H}_{S^{(m)}} \vee (y_0, y_1, \cdots, y_{m-1}, 1) \in \mathcal{H}_{S^{(m)}}\}$

Sauer's Lemma contd.

$Y_1 = \{(y_0, y_1, \cdots, y_{m-1}) : (y_0, y_1, \cdots, y_{m-1}, 0) \in \mathcal{H}_{S^{(m)}} \wedge (y_0, y_1, \cdots, y_{m-1}, 1) \in \mathcal{H}_{S^{(m)}}\}$ It's easy to notice that $|\mathcal{H}_{S^{(m)}}| = |Y_0| + |Y_1|$. Because if $(y_0, y_1, \cdots, y_{m-1}, 0)$ and $(y_0, y_1, \cdots, y_{m-1}, 1)$ both are in $\mathcal{H}_{S^{(m)}}$, then $(y_0, y_1, \cdots, y_{m-1})$ is included once in Y_0 and once in Y_1. All others are included once in Y_0. Also notice that, Y_0 is $\mathcal{H}_{S^{(m-1)}}$. So, using the inductive hypothesis,

$$\begin{aligned} |Y_0| = |\mathcal{H}_{S^{(m-1)}}| &\leq |A \subseteq S^{(m-1)} : \mathcal{H} \ \ shatters \ \ A| \\ &\leq |A \subseteq S^{(m)} : s_m \notin S^{(m)} \wedge \mathcal{H} \ \ shatters \ \ A| \end{aligned}$$

Let's define a new hypothesis set $H' \subseteq \mathcal{H}$ as follows:

$$\mathcal{H}' = \{h \in \mathcal{H} : \exists h' \in \mathcal{H} \ \ s.t. \ \ (h(s_1) = h'(s_1)) \wedge (h(s_2) = h'(s_2)) \wedge (h(s_3) = h'(s_3)) \wedge \cdots (h(s_m) \neq h'(s_m))\}$$

Sauer's Lemma contd.

So, it's evident that if $\mathcal{H}'$ shatters a set $A \subseteq S^{(m-1)}$, it also shatters $A \cup \{s_m\}$ and $Y_1 = \mathcal{H}'_{S^{m-1}}$. So, from induction:

$$\begin{aligned}|Y_1| = |\mathcal{H}'_{S^{m-1}}| &\leq |\{A \subseteq S^{(m-1)} : \mathcal{H}' \ \ shatters \ \ A\}| \\ &= |\{A \subseteq S^{(m-1)} : \mathcal{H}' \ \ shatters \ \ A \cup \{s_m\}\}| \\ &= |\{A \subseteq S^{(m)} : s_m \in A \wedge \mathcal{H}' \ \ shatters \ \ A\}| \\ &\leq |\{A \subseteq S^{(m)} : s_m \in A \wedge \mathcal{H} \ \ shatters \ \ A\}|\end{aligned}$$

Finally,

$$|\mathcal{H}_{S^{(m)}}| = |Y_0| + |Y_1| \leq |\{A \subseteq S^{(m)} : s_m \notin A \wedge \mathcal{H} \ \ shatters \ \ A\}| + |\{A \subseteq S^{(m)} : s_m \in A \wedge \mathcal{H} \ \ shatters \ \ A\}| = |\{A \subseteq S^{(m)} : \mathcal{H} \ \ shatters \ \ A\}|$$

But $|\{A \subseteq S^{(m)} : \mathcal{H} \ \ shatters \ \ A\}|$ is the number of subsets of $S^{(m)}$ that can be shattered using $\mathcal{H}$. Since $VCdim(\mathcal{H}) = d$, no subset with greater size can be shattered. Assuming the worst case, we can say that, $|\{A \subseteq S^{(m)} : \mathcal{H} \ \ shatters \ \ A\}| \leq \sum_{i=0}^{d} \binom{m}{i}$ which concludes our proof.

Corollary from Sauer's Lemma

Corollary

Let $\mathcal{H}$ be a hypothesis with $VCdim(\mathcal{H}) = d$, then for all $m \geq d$,

$$\mathcal{G}_{\mathcal{H}}(m) \leq (\frac{em}{d})^d = \mathcal{O}(m^d).$$

Proof: From the following proof, we can see the relationship between the VC-dimension and the growth function.

$$\begin{gathered}
\mathcal{G}_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \\
\leq \sum_{i=0}^{d} \binom{m}{i} (\frac{m}{d})^{d-i} \\
\leq \sum_{i=0}^{m} \binom{m}{i} (\frac{m}{d})^{d-i} \\
= (\frac{m}{d})^{d-m} \sum_{i=0}^{m} \binom{m}{i} (\frac{m}{d})^{m-i} \\
= (\frac{m}{d})^{d-m} (1 + \frac{m}{d})^{m} \\
= (\frac{m}{d})^{d} (1 + \frac{d}{m})^{m} \\
\leq (\frac{m}{d})^{d} e^{d}.
\end{gathered}$$

Generalization Bounds

Generalization Bound Using VC-dimension

Let H be a binary classification hypothesis class with $VCdim(\mathcal{H}) = d$. Then for any $1 > \delta > 0$ and for any sample S of size m, the following inequalities hold for all $h \in \mathcal{H}$ with probability $1 - \delta$:

$$R(h) \leq R_S(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$
$$R(h) \leq R_S(h) + \sqrt{\frac{8d \log \frac{2em}{d} + 8 \log \frac{4}{\delta}}{m}}$$

Generalization Bound Using Growth Function

Let H be a binary classification hypothesis class with growth function $\mathcal{G}_{\mathcal{H}}$. Then for any $1 > \delta > 0$ and for any sample S of size m, the following holds for all $h \in \mathcal{H}$ with probability $1 - \delta$:

$$R(h) \leq R_S(h) + \frac{4 + \sqrt{\log\left(\mathcal{G}_{\mathcal{H}}(2m)\right)}}{\delta\sqrt{2m}}$$