

L16 – Week 8

Introduction to Statistical Learning
Theory: VC dimension and
Learnability

CS 295 Optimization for Machine Learning

Ioannis Panageas

A motivating example

Recap:

- We saw that the hypothesis classes of finite cardinality are PAC learnable using **Chernoff Bounds and Union Bound**. What if the class is not finite?

A motivating example

Recap:

- We saw that the hypothesis classes of finite cardinality are PAC learnable using **Chernoff Bounds and Union Bound**. What if the class is not finite?

Lemma (Threshold functions). Consider the Hypothesis class of threshold functions on the real line, that is

$$\mathcal{H} = \{h_a : a \in \mathbb{R}\}$$

where $h_a(x) = \mathbf{1}_{x < a}$. \mathcal{H} is PAC learnable using ERM algorithm (even if the class is infinite).

A motivating example

Recap:

- We saw that the hypothesis classes of finite cardinality are PAC learnable using **Chernoff Bounds and Union Bound**. What if the class is not finite?

Lemma (Threshold functions). Consider the Hypothesis class of threshold functions on the real line, that is

$$\mathcal{H} = \{h_a : a \in \mathbb{R}\}$$

where $h_a(x) = \mathbf{1}_{x < a}$. \mathcal{H} is PAC learnable using ERM algorithm (even if the class is infinite).

Remarks:

- Therefore it is not necessary that the hypothesis class is of finite cardinality.
- We will show the lemma above, i.e., (ϵ, δ) -learnable using $\frac{\log^2 \frac{2}{\delta}}{\epsilon}$ **samples**.

A motivating example

Proof. Let D be the marginal distribution over the domain and fix ϵ, δ . We need to show that taking S samples IID of size $\frac{\log(2/\delta)}{\epsilon}$ suffices so that with probability $1 - \delta$ the generalization error is at most ϵ .

A motivating example

Proof. Let D be the marginal distribution over the domain and fix ϵ, δ . We need to show that taking S samples IID of size $\frac{\log(2/\delta)}{\epsilon}$ suffices so that with **probability $1 - \delta$ the generalization error is at most ϵ .**

Let a^* be a number such that h_{a^*} has error zero (perfect fit).

Moreover, consider $a_0 < a^* < a_1$ such that

$$\Pr_{x \sim D} [x \in (a_0, a^*)] = \Pr_{x \sim D} [x \in (a^*, a_1)] = \epsilon.$$

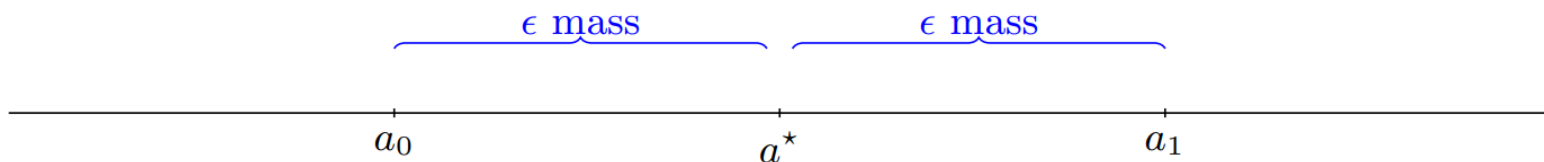
A motivating example

Proof. Let D be the marginal distribution over the domain and fix ϵ, δ . We need to show that taking S samples IID of size $\frac{\log(2/\delta)}{\epsilon}$ suffices so that with **probability $1 - \delta$ the generalization error is at most ϵ** .

Let a^* be a number such that h_{a^*} has error zero (perfect fit).

Moreover, consider $a_0 < a^* < a_1$ such that

$$\Pr_{x \sim D} [x \in (a_0, a^*)] = \Pr_{x \sim D} [x \in (a^*, a_1)] = \epsilon.$$



Observe that we might have to choose $a_0 = -\infty$ or $a_1 = +\infty$.

A motivating example

Proof cont. Let S be a set of IID samples and assume that the ERM algorithm returns a function h_S with threshold b_S .

If b_0 is the maximum x with label 1 and b_1 the minimum x with label 0 it holds that

$$b_S \in (b_0, b_1].$$

A motivating example

Proof cont. Let S be a set of IID samples and assume that the ERM algorithm returns a function h_S with threshold b_S .

If b_0 is the maximum x with label 1 and b_1 the minimum x with label 0 it holds that

$$b_S \in (b_0, b_1].$$

The error of h_S is at most ϵ if and only if $(b_0, b_1] \subseteq (a_0, a_1)$

A motivating example

Proof cont. Let S be a set of IID samples and assume that the ERM algorithm returns a function h_S with threshold b_S .

If b_0 is the maximum x with label 1 and b_1 the minimum x with label 0 it holds that

$$b_S \in (b_0, b_1].$$

The error of h_S is at most ϵ if and only if $(b_0, b_1] \subseteq (a_0, a_1)$

Let's bound the probability of this event!

A motivating example

Proof cont. Let S be a set of IID samples and assume that the ERM algorithm returns a function h_S with threshold b_S .

If b_0 is the maximum x with label 1 and b_1 the minimum x with label 0 it holds that

$$b_S \in (b_0, b_1].$$

The error of h_S is at most ϵ if and only if $(b_0, b_1] \subseteq (a_0, a_1)$

Let's bound the probability of this event!

By union bound we have

$$\Pr_{S \sim D^m} [(b_0 < a_0) \cup (b_1 > a_1)] \leq \Pr_{S \sim D^m} [(b_0 < a_0)] + \Pr_{S \sim D^m} [(b_1 > a_1)].$$

A motivating example

Proof cont.

$$\Pr_{S \sim D^m} [(b_0 < a_0)] \leq \Pr_S [\forall x \in S, x \notin (a_0, a^*)] = (1 - \epsilon)^m \leq e^{-\epsilon m}$$

A motivating example

Proof cont.

$$\Pr_{S \sim D^m} [(b_0 < a_0)] \leq \Pr_S [\forall x \in S, x \notin (a_0, a^*)] = (1 - \epsilon)^m \leq e^{-\epsilon m}$$

Using the same argument, we conclude that the error probability is $2e^{-\epsilon m} = \delta$. Solving for m we get

$$m = \frac{\log(2/\delta)}{\epsilon}.$$

A motivating example

Proof cont.

$$\Pr_{S \sim D^m} [(b_0 < a_0)] \leq \Pr_S [\forall x \in S, x \notin (a_0, a^*)] = (1 - \epsilon)^m \leq e^{-\epsilon m}$$

Using the same argument, we conclude that the error probability is $2e^{-\epsilon m} = \delta$. Solving for m we get

$$m = \frac{\log(2/\delta)}{\epsilon}.$$

All hypothesis classes are learnable then? Not really

VC dimension

Definition (Restriction). Let \mathcal{H} be a class of functions from \mathcal{X} to $\{0,1\}$ and let $C = \{c_1, \dots, c_m\}$. The restriction of \mathcal{H} to C is the set of functions from C to $\{0,1\}$ that can be derived from \mathcal{H} . That is

$$\mathcal{H}_C = \{h(c_1), \dots, h(c_m) : h \in \mathcal{H}\},$$

where we represent each function from C to $\{0,1\}$ as a vector in $\{0,1\}^{|C|}$.

VC dimension

Definition (Restriction). Let \mathcal{H} be a class of functions from \mathcal{X} to $\{0,1\}$ and let $C = \{c_1, \dots, c_m\}$. The restriction of \mathcal{H} to C is the set of functions from C to $\{0,1\}$ that can be derived from \mathcal{H} . That is

$$\mathcal{H}_C = \{h(c_1), \dots, h(c_m) : h \in \mathcal{H}\},$$

where we represent each function from C to $\{0,1\}$ as a vector in $\{0,1\}^{|C|}$.

Definition (Shattering). A hypothesis class \mathcal{H} shatters a finite set $C \subset \mathcal{X}$ if the restriction of \mathcal{H} to C is the set of all functions from C to $\{0,1\}$. That is $|\mathcal{H}_C| = 2^{|C|}$.

VC dimension

Definition (Restriction). Let \mathcal{H} be a class of functions from \mathcal{X} to $\{0,1\}$ and let $C = \{c_1, \dots, c_m\}$. The restriction of \mathcal{H} to C is the set of functions from C to $\{0,1\}$ that can be derived from \mathcal{H} . That is

$$\mathcal{H}_C = \{h(c_1), \dots, h(c_m) : h \in \mathcal{H}\},$$

where we represent each function from C to $\{0,1\}$ as a vector in $\{0,1\}^{|C|}$.

Definition (Shattering). A hypothesis class \mathcal{H} shatters a finite set $C \subset \mathcal{X}$ if the restriction of \mathcal{H} to C is the set of all functions from C to $\{0,1\}$. That is $|\mathcal{H}_C| = 2^{|C|}$.

Definition (VC dimension). The VC-dimension hypothesis class \mathcal{H} , denoted $VCdim(\mathcal{H})$, is the maximal size of a set C that can be shattered by \mathcal{H} . If \mathcal{H} can shatter sets of arbitrarily large size we say that \mathcal{H} has infinite VC-dimension.

Examples

- The class of threshold functions on real line has VC dimension 1. Why?
- The class of interval functions on real line has VC dimension 2. Why?
- The class of aligned rectangle functions on the plane has VC dimension 4. Why?

Examples

- The class of threshold functions on real line has **VC dimension 1. Why?**
- The class of interval functions on real line has **VC dimension 2. Why?**
- The class of aligned rectangle functions on the plane has **VC dimension 4. Why?**

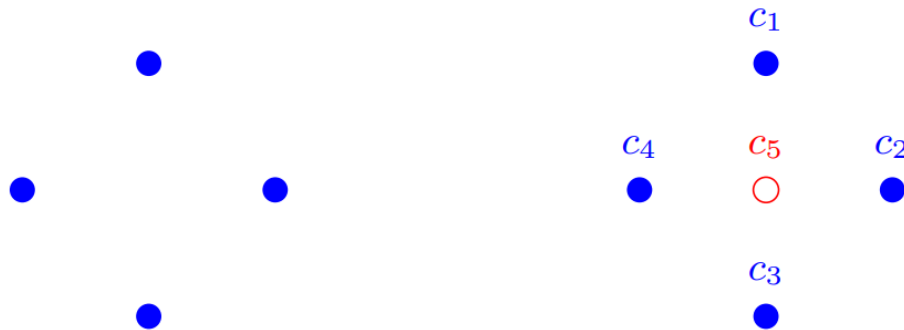


Figure 6.1 Left: 4 points that are shattered by axis aligned rectangles. Right: Any axis aligned rectangle cannot label c_5 by 0 and the rest of the points by 1.

Examples

- The class of threshold functions on real line has **VC dimension 1. Why?**
- The class of interval functions on real line has **VC dimension 2. Why?**
- The class of aligned rectangle functions on the plane has **VC dimension 4. Why?**



Figure 6.1 Left: 4 points that are shattered by axis aligned rectangles. Right: Any axis aligned rectangle cannot label c_5 by 0 and the rest of the points by 1.

- Any finite class H has **VC dimension at most $\log |H|$. Why?**

VC dimension of halfspaces

Theorem (Halfspaces). *The VC dimension of the class \mathcal{H} of homogenous halfspaces in \mathbb{R}^d is d . Note that $\mathcal{H} = \{h_w(x) : h_w(x) := \text{sign}(w^\top x)\}$.*

VC dimension of halfspaces

Theorem (Halfspaces). *The VC dimension of the class \mathcal{H} of homogenous halfspaces in \mathbb{R}^d is d . Note that $\mathcal{H} = \{h_w(x) : h_w(x) := \text{sign}(w^\top x)\}$.*

Proof. We first need to show that VC dimension is **at least d** by appropriately choosing a set C .

VC dimension of halfspaces

Theorem (Halfspaces). *The VC dimension of the class \mathcal{H} of homogenous halfspaces in \mathbb{R}^d is d . Note that $\mathcal{H} = \{h_w(x) : h_w(x) := \text{sign}(w^\top x)\}$.*

Proof. We first need to show that VC dimension is **at least d** by appropriately choosing a set C .

Consider the set of vectors e_1, \dots, e_d , where for every i the vector e_i is the all zeros vector except 1 in the i -th coordinate.

VC dimension of halfspaces

Theorem (Halfspaces). *The VC dimension of the class \mathcal{H} of homogenous halfspaces in \mathbb{R}^d is d . Note that $\mathcal{H} = \{h_w(x) : h_w(x) := \text{sign}(w^\top x)\}$.*

Proof. We first need to show that VC dimension is **at least d** by appropriately choosing a set C .

Consider the set of vectors e_1, \dots, e_d , where for every i the vector e_i is the all zeros vector except 1 in the i -th coordinate.

This set is shattered by the class of homogenous halfspaces because for every binary vector y_1, \dots, y_d , and for $w = (y_1, \dots, y_d)$, we get that $h_w(e_i) = y_i$.

VC dimension of halfspaces

Theorem (Halfspaces). *The VC dimension of the class \mathcal{H} of homogenous halfspaces in \mathbb{R}^d is d . Note that $\mathcal{H} = \{h_w(x) : h_w(x) := \text{sign}(w^\top x)\}$.*

Proof. We first need to show that VC dimension is **at least d** by appropriately choosing a set C .

Consider the set of vectors e_1, \dots, e_d , where for every i the vector e_i is the all zeros vector except 1 in the i -th coordinate.

This set is shattered by the class of homogenous halfspaces because for every binary vector y_1, \dots, y_d , and for $w = (y_1, \dots, y_d)$, we get that $h_w(e_i) = y_i$.

We need now to show that VC dimension is less than $d + 1$. Let x_1, \dots, x_{d+1} be a set of $d + 1$ vectors in \mathbb{R}^d .

VC dimension of halfspaces

Proof cont. Then, there must exist real numbers a_1, \dots, a_{d+1} , not all of them are zero, such that

$$\sum a_i x_i = 0 \text{ linearly dependent.}$$

VC dimension of halfspaces

Proof cont. Then, there must exist real numbers a_1, \dots, a_{d+1} , not all of them are zero, such that

$$\sum a_i x_i = 0 \text{ linearly dependent.}$$

Let $I = \{i : a_i > 0\}$ and $J = \{j : a_j < 0\}$.

If both I, J are non-empty then

$$\sum_{i \in I} a_i x_i = \sum_{j \in J} |a_j| x_j.$$

VC dimension of halfspaces

Proof cont. Then, there must exist real numbers a_1, \dots, a_{d+1} , not all of them are zero, such that

$$\sum a_i x_i = 0 \text{ linearly dependent.}$$

Let $I = \{i : a_i > 0\}$ and $J = \{j : a_j < 0\}$.

If both I, J are non-empty then

$$\sum_{i \in I} a_i x_i = \sum_{j \in J} |a_j| x_j.$$

If x_1, \dots, x_{d+1} are shattered then there exists a w such that $w^\top x_i > 0$ for $i \in I$ and $w^\top x_j < 0$ for $j \in J$.

VC dimension of halfspaces

If x_1, \dots, x_{d+1} are shattered then there exists a w such that $w^\top x_i > 0$ for $i \in I$ and $w^\top x_j < 0$ for $j \in J$.

If the above is true, we get that

$$0 < \sum_{i \in I} a_i w^\top x_i = w^\top \sum_{i \in I} a_i x_i$$

VC dimension of halfspaces

If x_1, \dots, x_{d+1} are shattered then there exists a w such that $w^\top x_i > 0$ for $i \in I$ and $w^\top x_j < 0$ for $j \in J$.

If the above is true, we get that

$$\begin{aligned} 0 < \sum_{i \in I} a_i w^\top x_i &= w^\top \sum_{i \in I} a_i x_i \\ &= w^\top \sum_{j \in J} |a_j| x_j \end{aligned}$$

VC dimension of halfspaces

If x_1, \dots, x_{d+1} are shattered then there exists a w such that $w^\top x_i > 0$ for $i \in I$ and $w^\top x_j < 0$ for $j \in J$.

If the above is true, we get that

$$\begin{aligned} 0 &< \sum_{i \in I} a_i w^\top x_i = w^\top \sum_{i \in I} a_i x_i \\ &= w^\top \sum_{j \in J} |a_j| x_j \\ &= \sum_{j \in J} |a_j| w^\top x_j < 0. \end{aligned}$$

VC dimension of halfspaces

If x_1, \dots, x_{d+1} are shattered then there exists a w such that $w^\top x_i > 0$ for $i \in I$ and $w^\top x_j < 0$ for $j \in J$.

If the above is true, we get that

$$\begin{aligned} 0 &< \sum_{i \in I} a_i w^\top x_i = w^\top \sum_{i \in I} a_i x_i \\ &= w^\top \sum_{j \in J} |a_j| x_j \\ &= \sum_{j \in J} |a_j| w^\top x_j < 0. \end{aligned}$$

Contradiction!

Example of infinite VC

Theorem (**sin has infinite VC**). *Consider the real line and let*

$$\mathcal{H} = \{x \rightarrow \lceil \sin(\theta x) \rceil : \theta \in \mathbb{R}\}.$$

The VC dimension of the hypothesis class above is infinite.

Proof. We need to show that for every d one can find d points that are shattered by \mathcal{H} .

Example of infinite VC

Theorem (**sin has infinite VC**). Consider the real line and let

$$\mathcal{H} = \{x \rightarrow \lceil \sin(\theta x) \rceil : \theta \in \mathbb{R}\}.$$

The VC dimension of the hypothesis class above is infinite.

Proof. We need to show that for every d one can find d points that are shattered by \mathcal{H} .

Consider $x \in (0, 1)$ and let $0.x_1x_2x_3\dots$, be the binary expansion of x . Then for any natural number m , $\lceil \sin(2^m \pi x) \rceil = 1 - x_m$, provided that there exists a $k \geq m$ such that $x_k = 1$.

Example of infinite VC

Theorem (**sin has infinite VC**). Consider the real line and let

$$\mathcal{H} = \{x \rightarrow \lceil \sin(\theta x) \rceil : \theta \in \mathbb{R}\}.$$

The VC dimension of the hypothesis class above is infinite.

Proof. We need to show that for every d one can find d points that are shattered by \mathcal{H} .

Consider $x \in (0, 1)$ and let $0.x_1x_2x_3\dots$, be the binary expansion of x . Then for any natural number m , $\lceil \sin(2^m \pi x) \rceil = 1 - x_m$, provided that there exists a $k \geq m$ such that $x_k = 1$.

Fix d and consider $C = \{1/2, 1/4, \dots, 1/2^d\}$ and moreover choose any binary vector of labels (y_1, \dots, y_d) . Set $x = 0.y_1\dots y_d1$ and use the above.

Why do we care about VC?

Theorem (Fundamental Theorem of Learnability). *The following are equivalent:*

- \mathcal{H} is PAC learnable.
- Any ERM rule is a successful PAC learner for \mathcal{H} .
- \mathcal{H} has finite VC dimension.

Remarks:

- The number of samples needed is $O\left(\frac{d \log \frac{1}{\epsilon} + \log \frac{1}{\delta}}{\epsilon}\right)$ where d is the VC dimension of the hypothesis class.

Conclusion

- Introduction to Statistical Learning.
 - VC dimension.
 - Examples.
 - Fundamental theorem of Learnability
- Last lecture we will be talking about **Stochastic Games and Multi-agent RL**.