Computational Learning Theory: Shattering and VC Dimensions
This lecture: Computational Learning Theory

• The Theory of Generalization

• Probably Approximately Correct (PAC) learning

• Positive and negative learnability results

• Agnostic Learning

• Shattering and the VC dimension
This lecture: Computational Learning Theory

• The Theory of Generalization

• Probably Approximately Correct (PAC) learning

• Positive and negative learnability results

• Agnostic Learning

• Shattering and the VC dimension
Infinite Hypothesis Space

- The previous analysis was restricted to finite hypothesis spaces
- Some infinite hypothesis spaces are more expressive than others
  - E.g., Rectangles, vs. 17-sided convex polygons vs. general convex polygons
  - Linear threshold function vs. a combination of LTUs

Vapnik-Chervonenkis (VC) dimension provides a measure of the expressive capacity of a set of functions, analogous to $|H|$.
Infinite Hypothesis Space

- The previous analysis was restricted to finite hypothesis spaces
- Some infinite hypothesis spaces are more expressive than others
  - E.g., Rectangles, vs. 17- sides convex polygons vs. general convex polygons
  - Linear threshold function vs. a combination of LTUs
- Need a measure of the expressiveness of an infinite hypothesis space other than its size
Infinite Hypothesis Space

• The previous analysis was restricted to finite hypothesis spaces
• Some infinite hypothesis spaces are more expressive than others
  – E.g., Rectangles, vs. 17- sides convex polygons vs. general convex polygons
  – Linear threshold function vs. a combination of LTUs
• Need a measure of the expressiveness of an infinite hypothesis space other than its size
• The Vapnik-Chervonenkis dimension (VC dimension) provides such a measure
  – “What is the expressive capacity of a set of functions?”
Infinite Hypothesis Space

• The previous analysis was restricted to finite hypothesis spaces
• Some infinite hypothesis spaces are more expressive than others
  – E.g., Rectangles, vs. 17-sides convex polygons vs. general convex polygons
  – Linear threshold function vs. a combination of LTUs
• Need a measure of the expressiveness of an infinite hypothesis space other than its size
• The Vapnik-Chervonenkis dimension (VC dimension) provides such a measure
  – “What is the expressive capacity of a set of functions?”
• Analogous to $|H|$, there are bounds for sample complexity using $VC(H)$
Learning Rectangles

Assume the target concept is an axis parallel rectangle
Learning Rectangles

Assume the target concept is an axis parallel rectangle

Points inside are positive

Points outside are negative
Learning Rectangles

Assume the target concept is an axis parallel rectangle
Learning Rectangles

Assume the target concept is an axis parallel rectangle
Learning Rectangles

Assume the target concept is an axis parallel rectangle
Learning Rectangles

Assume the target concept is an axis parallel rectangle
Learning Rectangles

Assume the target concept is an axis parallel rectangle
Learning Rectangles

Assume the target concept is an axis parallel rectangle
Learning Rectangles

Assume the target concept is an axis parallel rectangle

Will we be able to learn the target rectangle?

Can we come close?
Let’s think about expressivity of functions

Suppose we have two points.

Can linear classifiers correctly classify any labeling of these points?
Let’s think about expressivity of functions

There are four ways to label two points
Let’s think about expressivity of functions

There are four ways to label two points

And it is possible to draw a line that separates positive and negative points in all four cases.
Let’s think about expressivity of functions

There are four ways to label two points

And it is possible to draw a line that separates positive and negative points in all four cases

We say that linear functions are expressive enough to *shatter two* points
Let’s think about expressivity of functions

There are four ways to label two points

And it is possible to draw a line that separates positive and negative points in all four cases

We say that linear functions are expressive enough to *shatter* two points

What about fourteen points?
Shattering
Shattering
Shattering
Shattering

What about this labeling?
Shattering

This particular labeling of the points cannot be separated by any line
This particular labeling of the points cannot be separated by any line
Shattering

This particular labeling of the points cannot be separated by any line.
Shattering

This particular labeling of the points cannot be separated by any line
Shattering

Linear functions are **not** expressive enough to shatter fourteen points

Because there is *at least one labeling* that can not be separated by them
Shattering

Linear functions are not expressive enough to shatter fourteen points.

Because there is at least one labeling that can not be separated by them.

Of course, a more complex function could separate them.
Shattering

**Definition:** A set $S$ of examples is shattered by a set of functions $H$ if for every partition of the examples in $S$ into positive and negative examples there is a function in $H$ that gives exactly these labels to the examples.

**Intuition:** A rich set of functions shatters large sets of points.
Shattering

**Definition:** A set $S$ of examples is **shattered** by a set of functions $H$ if for every partition of the examples in $S$ into positive and negative examples there is a function in $H$ that gives exactly these labels to the examples.

**Intuition:** A rich set of functions shatters large sets of points.

**Example 1:** Hypothesis class of left bounded intervals on the real axis: $[0,a)$ for some real number $a > 0$
Left bounded intervals

Example 1: Hypothesis class of left bounded intervals on the real axis: \([0,a)\) for some real number \(a>0\)
Left bounded intervals

**Example 1**: Hypothesis class of left bounded intervals on the real axis: $[0,a)$ for some real number $a > 0$.

![Diagram of a real axis with a point at 0 and an open interval $[0,a)$ highlighted.](image)
Left bounded intervals

**Example 1:** Hypothesis class of left bounded intervals on the real axis: $[0,a)$ for some real number $a>0$

If we have a set $S$ with only this one point

If the point is labeled $+$, we can find an $a$ that is to the right of that point

This hypothesis correctly labels the point as positive
Left bounded intervals

Example 1: Hypothesis class of left bounded intervals on the real axis: \([0, a)\) for some real number \(a > 0\)

If the point is labeled \(-\), we can find an \(a\) that is to the right of that point.

This hypothesis correctly labels the point as negative.

If we have a set \(S\) with only this one point.
Left bounded intervals

Example 1: Hypothesis class of left bounded intervals on the real axis: [0,a) for some real number a>0

If we have a set S with only this one point, we can find an $a$ that is to the right of that point. If the point is labeled $-$, we can find an $a$ that is to the right of that point. This hypothesis correctly labels the point as negative.

Any set of one point can be shattered by the hypothesis class of left bounded intervals.
Left bounded intervals

**Example 1**: Hypothesis class of left bounded intervals on the real axis: $[0,a)$ for some real number $a > 0$

Let us consider a set with two points

If we have a set $S$ with these two points
Left bounded intervals

**Example 1**: Hypothesis class of left bounded intervals on the real axis: \([0,a)\) for some real number \(a > 0\)

Let us consider a set with two points

If we have a set \(S\) with these two points

We can label the points such that no hypothesis in our class can match the labels
Left bounded intervals

Example 1: Hypothesis class of left bounded intervals on the real axis: \([0,a)\) for some real number \(a > 0\)

Let us consider a set with two points

If we have a set \(S\) with these two points

We can label the points such that no hypothesis in our class can match the labels
**Left bounded intervals**

**Example 1:** Hypothesis class of left bounded intervals on the real axis: \([0,a)\) for some real number \(a > 0\)

Let us consider a set with two points. We can label the points such that no hypothesis in our class can match the labels.

Let's consider the points 0 and \(a\) on the real axis. One point is labeled with a negative sign, and the other with a positive sign. The left bounded interval \([0,a)\) incorrectly labels this point as negative.

We can label the points such that no hypothesis in our class can match the labels.
Left bounded intervals

**Example 1:** Hypothesis class of left bounded intervals on the real axis: \([0,a)\) for some real number \(a>0\)

Let us consider a set with two points

We can label the points such that no hypothesis in our class can match the labels

Incorrectly labels this point as positive
Left bounded intervals

Example 1: Hypothesis class of left bounded intervals on the real axis: \([0,a)\) for some real number \(a>0\)

Let us consider a set with two points

We can label the points such that no hypothesis in our class can match the labels
Shattering

**Definition**: A set $S$ of examples is **shattered** by a set of functions $H$ if for every partition of the examples in $S$ into positive and negative examples there is a function in $H$ that gives exactly these labels to the examples.

**Intuition**: A rich set of functions shatters large sets of points.

**Example 1**: Hypothesis class of left bounded intervals on the real axis: $[0,a)$ for some real number $a>0$. 
Shattering

**Definition:** A *set S of examples* is **shattered** by a *set of functions* H if for *every* partition of the examples in S into positive and negative examples there is a function in H that gives exactly these labels to the examples

**Intuition:** A rich set of functions shatters large sets of points

**Example 1:** Hypothesis class of left bounded intervals on the real axis: \([0,a)\) for some real number \(a>0\)

Sets with **one** point **can** be shattered

That is: Given one point, for any labeling of the points, we can find a concept in this class that is consistent with it
Shattering

**Definition:** A set $S$ of examples is **shattered** by a set of functions $H$ if for every partition of the examples in $S$ into positive and negative examples there is a function in $H$ that gives exactly these labels to the examples.

**Intuition:** A rich set of functions shatters large sets of points.

**Example 1:** Hypothesis class of left bounded intervals on the real axis: $[0,a)$ for some real number $a>0$.

- Sets with **one** point can be shattered.
- Sets with **two** points cannot be shattered.

That is: Given one point, for any labeling of the points, we can find a concept in this class that is consistent with it. That is: given two points, you can label them in such a way that no concept in this class will be consistent with their labeling.
Shattering

**Definition**: A set $S$ of examples is **shattered** by a set of functions $H$ if for every partition of the examples in $S$ into positive and negative examples there is a function in $H$ that gives exactly these labels to the examples.

**Example 2**: Hypothesis class is the set of intervals on the real axis: $[a,b]$, for some real numbers $b>a$.
**Definition:** A set $S$ of examples is shattered by a set of functions $H$ if for every partition of the examples in $S$ into positive and negative examples there is a function in $H$ that gives exactly these labels to the examples.

**Example 2:** Hypothesis class is the set of intervals on the real axis: $[a, b]$, for some real numbers $b > a$.

Points in this region will be labeled as positive.

Points outside the shaded region will be labeled as negative.
Real intervals

Example 2: Hypothesis class is the set of intervals on the real axis: $[a, b]$, for some real numbers $b > a$
Shattering

**Definition**: A set $S$ of examples is **shattered** by a set of functions $H$ if for *every* partition of the examples in $S$ into positive and negative examples there is a function in $H$ that gives exactly these labels to the examples.

**Example 2**: Hypothesis class is the set of **intervals on the real axis**: $[a, b], \text{for some real numbers } b>a$.
Shattering

**Definition:** A set $S$ of examples is **shattered** by a set of functions $H$ if for every partition of the examples in $S$ into positive and negative examples there is a function in $H$ that gives exactly these labels to the examples.

**Example 2:** Hypothesis class is the set of intervals on the real axis: $[a,b]$, for some real numbers $b>a$.

All sets of one or two points can be shattered.
But sets of **three** points **cannot** be shattered.

Proof? Enumerate all possible three points.
**Definition:** A set $S$ of examples is shattered by a set of functions $H$ if for every partition of the examples in $S$ into positive and negative examples there is a function in $H$ that gives exactly these labels to the examples.

**Example 3:** Half spaces in a plane
Shattering

**Definition:** A set $S$ of examples is **shattered** by a set of functions $H$ if for every partition of the examples in $S$ into positive and negative examples there is a function in $H$ that gives exactly these labels to the examples.

**Example 3:** Half spaces in a plane

Can one point be shattered?

Two points?

Three points? Can any three points be shattered?
Half spaces on a plane: 3 points
**Shattering**

**Definition:** A set $S$ of examples is **shattered** by a set of functions $H$ if for every partition of the examples in $S$ into positive and negative examples there is a function in $H$ that gives exactly these labels to the examples.

**Example 3:** Half spaces in a plane

Can four points be shattered?

Suppose three of them lie on the same line, label the outside points + and the inner one –

Otherwise, make a convex hull. Label points outside + and the inner one –

**Four** points **cannot** be shattered!
Half spaces on a plane: 4 points
Shattering: The adversarial game

You

An adversary
Shattering: The adversarial game

**You**

**An adversary**

**You:** Hypothesis class $H$ can shatter

*these* $d$ points
Shattering: The adversarial game

You: Hypothesis class $H$ can shatter these $d$ points

Adversary: That’s what you think! Here is a labeling that will defeat you.
You: Hypothesis class $H$ can shatter these $d$ points

Adversary: That’s what you think! Here is a labeling that will defeat you.

You: Aha! There is a function $h \in H$ that correctly predicts your evil labeling
Shattering: The adversarial game

**You**

**You**: Hypothesis class $H$ can shatter *these* $d$ points

**You**: Aha! There is a function $h \in H$ that correctly predicts your evil labeling

**An adversary**

**Adversary**: That’s what you think! Here is a labeling that will defeat you.

**Adversary**: Argh! You win this round. But I’ll be back.....
Some functions can shatter infinite points!

If arbitrarily large finite subsets of the instance space $X$ can be shattered by a hypothesis space $H$.

**Intuition**: A rich set of functions shatters large sets of points
Some functions can shatter infinite points!

If arbitrarily large finite subsets of the instance space $X$ can be shattered by a hypothesis space $H$.

An unbiased hypothesis space $H$ shatters the entire instance space $X$, i.e., it can induce every possible partition on the set of all possible instances.

The larger the subset $X$ that can be shattered, the more expressive a hypothesis space is, i.e., the less biased it is.

**Intuition:** A rich set of functions shatters large sets of points.
A set $S$ of examples is **shattered** by a set of functions $H$ if for every partition of the examples in $S$ into positive and negative examples there is a function in $H$ that gives exactly these labels to the examples.
A set $S$ of examples is **shattered** by a set of functions $H$ if for every partition of the examples in $S$ into positive and negative examples there is a function in $H$ that gives exactly these labels to the examples.

**Definition:** The VC dimension of hypothesis space $H$ over instance space $X$ is the size of the largest **finite** subset of $X$ that is shattered by $H$. 
Vapnik-Chervonenkis Dimension

A set $S$ of examples is **shattered** by a set of functions $H$ if for every partition of the examples in $S$ into positive and negative examples there is a function in $H$ that gives exactly these labels to the examples.

**Definition**: The VC dimension of hypothesis space $H$ over instance space $X$ is the size of the largest **finite** subset of $X$ that is shattered by $H$.

- If there **exists** any subset of size $d$ that can be shattered, $VC(H) \geq d$.
  - Even one subset will do.
- If no subset of size $d$ can be shattered, then $VC(H) < d$. 
What we have managed to prove

<table>
<thead>
<tr>
<th>Concept class</th>
<th>VC Dimension</th>
<th>Why?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Half intervals</td>
<td>1</td>
<td>There is a dataset of size 1 that can be shattered</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No dataset of size 2 can be shattered</td>
</tr>
<tr>
<td>Intervals</td>
<td>2</td>
<td>There is a dataset of size 2 that can be shattered</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No dataset of size 3 can be shattered</td>
</tr>
<tr>
<td>Half-spaces in the plane</td>
<td>3</td>
<td>There is a dataset of size 3 that can be shattered</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No dataset of size 4 can be shattered</td>
</tr>
</tbody>
</table>
## More VC dimensions

<table>
<thead>
<tr>
<th>Concept class</th>
<th>VC Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear threshold unit in (d) dimensions</td>
<td>(d + 1)</td>
</tr>
<tr>
<td>Neural networks</td>
<td>Number of parameters</td>
</tr>
<tr>
<td>1 nearest neighbors</td>
<td>infinite</td>
</tr>
</tbody>
</table>

**Intuition**: A rich set of functions shatters large sets of points
### More VC dimensions

<table>
<thead>
<tr>
<th>Concept class</th>
<th>VC Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear threshold unit in d dimensions</td>
<td>$d + 1$</td>
</tr>
<tr>
<td>Neural networks</td>
<td>Number of parameters</td>
</tr>
<tr>
<td>1 nearest neighbors</td>
<td>infinite</td>
</tr>
</tbody>
</table>

What is the number of parameters needed to specify a linear threshold unit in $d$ dimensions?

**Intuition:** A rich set of functions shatters large sets of points.
## More VC dimensions

<table>
<thead>
<tr>
<th>Concept class</th>
<th>VC Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear threshold unit in d dimensions</td>
<td>$d + 1$</td>
</tr>
<tr>
<td>Neural networks</td>
<td>Number of parameters</td>
</tr>
<tr>
<td>1 nearest neighbors</td>
<td>infinite</td>
</tr>
</tbody>
</table>

What is the number of parameters needed to specify a linear threshold unit in d dimensions? $d + 1$

**Intuition:** A rich set of functions shatters large sets of points.
### More VC dimensions

<table>
<thead>
<tr>
<th>Concept class</th>
<th>VC Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear threshold unit in d dimensions</td>
<td>$d + 1$</td>
</tr>
<tr>
<td>Neural networks</td>
<td>Number of parameters</td>
</tr>
<tr>
<td>1 nearest neighbors</td>
<td>infinite</td>
</tr>
</tbody>
</table>

**What is the number of parameters needed to specify a linear threshold unit in d dimensions?** $d + 1$

**Local minima in learning means neural networks may not find the best parameters.**

**Intuition:** A rich set of functions shatters large sets of points
# More VC dimensions

<table>
<thead>
<tr>
<th>Concept class</th>
<th>VC Dimension</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear threshold unit in d dimensions</td>
<td>$d + 1$</td>
<td>What is the number of parameters needed to specify a linear threshold unit in d dimensions? $d + 1$</td>
</tr>
<tr>
<td>Neural networks</td>
<td>Number of parameters</td>
<td>Local minima in learning means neural networks may not find the best parameters</td>
</tr>
<tr>
<td>1 nearest neighbors</td>
<td>infinite</td>
<td><strong>Exercise</strong>: Try to prove this after we see nearest neighbors</td>
</tr>
</tbody>
</table>

**Intuition**: A rich set of functions shatters large sets of points
Why VC dimension?

• Remember sample complexity
  – Occam’s razor
  – Agnostic learning

• Sample complexity in both cases depends on the log of the size of the hypothesis space

• For infinite hypothesis spaces, its VC dimension behaves like $\log(|H|)$
VC dimension and Occam’s razor for consistent learners

• Using VC(H) as a measure of expressiveness, we have an Occam theorem for infinite hypothesis spaces

• Given a sample D with m examples, find some \( h \in H \) is consistent with all m examples. If

\[
m > \frac{1}{\varepsilon} \left( 8 \text{VC}(H) \log \frac{13}{\varepsilon} + 4 \log \frac{2}{\delta} \right)
\]

Then with probability at least \( 1 - \delta \), the hypothesis \( h \) has error less than \( \varepsilon \).

That is, if \( m \) is polynomial we have a PAC learning algorithm; To be efficient, we need to produce the hypothesis \( h \) efficiently
VC dimension and Agnostic Learning

Similar statement for the agnostic setting as well

If we have $m$ examples, then with probability $1 - \delta$, the true error of a hypothesis $h$ with training error $\text{err}_S(h)$ is bounded by

$$\text{err}_D(h) \leq \text{err}_S(h) + \sqrt{\frac{\text{VC}(H) \left( \ln \frac{2m}{\text{VC}(H)} + 1 \right) + \ln \frac{4}{\delta}}{m}}$$

(Phew!)
Exercises

What is the VC dimension axis parallel rectangles (which we saw at the beginning of this lecture)?

Your homework asks you to compute the VC dimension of different classes of functions
PAC learning: What you need to know

• What is PAC learning?
  – Remember: We care about generalization error, not training error

• Finite hypothesis spaces
  – Connection between size of hypothesis space and sample complexity
  – Derive and understand the sample complexity bounds
  – Count number of hypotheses in a hypothesis class

• Infinite hypotheses classes
  – What is shattering and VC dimension?
  – How to find VC dimension of simple concept classes?
  – Higher VC dimensions ⇒ more sample complexity
Computational Learning Theory

• Probably Approximately Correct (PAC) learning
  – A general definition that assumes fixed, but perhaps unknown distribution

• Occam’s razor for consistent learners in finite hypothesis spaces
  – Positive and negative learnability results in this setting

• Agnostic Learning and the associated Occam razor

• Shattering and the VC dimension

• Many extensions to the theory exist
  – Noisy data, known data distributions, probabilistic models
  – One important extension: **PAC-Bayes theory** that makes assumptions about the prior distribution over hypothesis spaces
COLT still doesn’t explain why learning works in all cases

Why computational learning theory

• Answers questions such as
  – What is learnability? How good is my class of functions?
  – Is a concept learnable? How many examples do I need?

• Mistake bounds imply PAC-learnability

• Raises interesting theoretical questions
  – If a concept class is weakly learnable (i.e. there is a learning algorithm that can produce a classifier that does slightly better than chance), does this mean that the concept class is strongly learnable?
  
  – We have seen bounds of the form
    \[ \text{true error} < \text{training error} + (\text{a term with } \pm \text{ and VC dimension}) \]
    Can we use this to define a learning algorithm?
Why computational learning theory

• Answers questions such as
  – What is learnability? How good is my class of functions?
  – Is a concept learnable? How many examples do I need?

• Mistake bounds imply PAC-learnability

• Raises interesting theoretical questions
  – If a concept class is weakly learnable (i.e. there is a learning algorithm that can produce a classifier that does slightly better than chance), does this mean that the concept class is strongly learnable?

  Boosting
  – We have seen bounds of the form
    \[
    \text{true error} < \text{training error} + (a \text{ term with } \epsilon, \delta \text{ and VC dimension})
    \]
    Can we use this to define a learning algorithm?

    Structural Risk Minimization principle
    Support Vector Machine