EECS 349: Machine Learning

Bryan Pardo

Topic: Concept Learning
Concept Learning

- Much of learning involves acquiring general concepts from specific training examples

- *Concept*: subset of objects from some space

- Concept learning: Defining a function that specifies which elements are in the concept set.
A Concept

• Let there be a set of unique instances $X$.
  \[ X = \{\text{duck, rabbit, chicken, pig, planet, bag-of-money, cow, dog}\} \]

• Let there be a set of unique labels $L$
  \[ L = \{0, 1\} \]

• A concept $C$ is...

  A subset of $X$
  \[ \text{e.g. } C = \text{mammals} = \{\text{rabbit, pig, cow, dog}\} \]

  A function that returns \textbf{1} (or “true”) only for elements in the concept
  \[ \text{e.g. } C(\text{rabbit}) = 1, \ C(\text{duck}) = 0 \]
Some more definitions

Call the set of unique examples \( X \) the *instance space*.

The set of unique labels \( L \) is the *label space*.

A hypothesis \( h(x) \) is, like a concept \( c(x) \), a function whose domain is \( X \) and whose range is \( L \).

The set of unique concepts is the *concept space* \( C \).

The set of unique hypotheses a learner will consider is called the *hypothesis space* \( H \).

It is usually not true that \( H=C \) (we’ll see why in a moment)
GIVEN:
- A space of instances $X$
- Target concept function $c$:
  E.g., Mammal: $X \rightarrow \{0,1\}$
- Hypothesis space $H$
- Training data $D$
  positive and negative examples: $<x_1, c(x_1)>, ..., <x_n, c(x_n)>$

FIND:
- A hypothesis $h$ in $H$ such that $h(x) = c(x)$ for all $x$ in $D$.  

Concept Learning Task
**Telling hypotheses/concepts apart**

- **Definition:** Two functions $f_1$ and $f_2$ are *distinguishable*, given the data $D$, if they differ in their labeling of at least one of the examples in $D$.

- **Definition:** A set of hypotheses is distinguishable, given $D$, iff ALL pairs of hypotheses in the set are distinguishable given $D$. Call $H_D$ a largest set of distinguishable hypotheses, given $D$. 
Inductive Learning Hypothesis

- Any hypothesis found to approximate the target function well over the training examples, will also approximate the target function well over the unobserved examples.

- This might not be true. When it it isn’t the hypothesis does not generalize well.

- In fact, the target concept may not even be in the hypothesis space.

- ...but maybe we can find a hypothesis that is good enough for our purposes.
Version Spaces

- Hypothesis $h$ is **consistent** with a set of training examples $D$ of the target concept $c$ iff $h(x) = c(x)$ for each training example $<x, c(x)>$ in $D$.

  $$\text{Consistent}(h, D) \equiv (\forall <x, c(x)> \in D) \ h(x) = c(x)$$

- A **version space**: all the hypotheses that are consistent with the training examples.

  $$VS_{H, D} \equiv \{ h \in H \mid \text{Consistent}(h, D) \}$$
A visualization

$h(x)$ from a linear regressor

$c(x)$, the actual concept

● = Training data
● = Validation data
As images, the two rabbits are unique instances
Encoding Matters: Feature vector

- This feature encoding makes the rabbits identical to each other...

<table>
<thead>
<tr>
<th>Number of Feet</th>
<th>Fur</th>
<th>Size</th>
<th>Has wings</th>
<th>Warm Blood</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>No</td>
<td>S</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>S</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>S</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>M</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>0</td>
<td>No</td>
<td>XXL</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>0</td>
<td>No</td>
<td>M</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>L</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>S</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>S</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

...and to the dog

Moral: pick the right encoding!
How many unique instances?

<table>
<thead>
<tr>
<th>Number of Feet</th>
<th>Fur</th>
<th>Size</th>
<th>Has wings</th>
<th>Warm Blood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integers 0 to 99</td>
<td>Yes, No</td>
<td>S, M, L, XL, XXL</td>
<td>Yes, No</td>
<td>Yes, No</td>
</tr>
</tbody>
</table>

100 * 2 * 5 * 2 * 2 = 4000 instances
How many unique concepts?

<table>
<thead>
<tr>
<th>Number of Feet</th>
<th>Fur</th>
<th>Size</th>
<th>Has wings</th>
<th>Warm Blood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integers 0 to 99</td>
<td>Yes,No</td>
<td>S,M,L,XL,XXL</td>
<td>Yes,No</td>
<td>Yes,No</td>
</tr>
</tbody>
</table>

$$100 \times 2 \times 5 \times 2 \times 2 = 4000 \text{ instances}$$

$$2^{4000} \text{ concepts}$$
How big is the version space for this data?

<table>
<thead>
<tr>
<th>Number of Feet</th>
<th>Fur</th>
<th>Size</th>
<th>Has wings</th>
<th>Warm Blood</th>
<th>C(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>No</td>
<td>S</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>S</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>S</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>M</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>No</td>
<td>XXL</td>
<td>No</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>No</td>
<td>M</td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>L</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>S</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>S</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
</tbody>
</table>

TRICK QUESTION! Until you know what set of hypothesis a learner can consider, you can’t say how big the version space is.
Example: MC2 Hypothesis Space

- MC2 (Mitchell, Chapter 2) hypothesis space
  Hypothesis $h$ is a conjunction of constraints on attributes

- Each constraint can be:
  A specific value: e.g. $Number\ of\ Feet = 4$
  A don’t care value: e.g. $Fur = ?$
  No value allowed: e.g. $Size = \emptyset$

- Instances $x$ that satisfy $h$ have $h(x) = 1$, else $h(x) = 0$

- Example hypotheses:

<table>
<thead>
<tr>
<th>Number of Feet</th>
<th>Fur</th>
<th>Size</th>
<th>Has wings</th>
<th>Warm Blood</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>?</td>
<td>Yes</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>$\emptyset$</td>
<td>S</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
How many unique hypotheses?

- Given this encoding of hypotheses how many hypotheses are possible?

<table>
<thead>
<tr>
<th>Number of Feet</th>
<th>Fur</th>
<th>Size</th>
<th>Has wings</th>
<th>Warm Blood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integers 0 to 99</td>
<td>Yes, No</td>
<td>S, M, L, XL, XXL</td>
<td>Yes, No</td>
<td>Yes, No</td>
</tr>
</tbody>
</table>

$$(102) * 4 * 7 * 4 * 4 = 45,696$$ hypotheses

Compare that to $2^{4000}$ concepts
Questions

• Does the MC2 hypothesis space contain the concept “has either 2 feet or 4 feet”?  

• NO! It cannot represent concepts that accept subsets of values in an attribute

• Why would we use such a limited hypothesis space?
Now how big is the version space?

<table>
<thead>
<tr>
<th>Number of Feet</th>
<th>Fur</th>
<th>Size</th>
<th>Has wings</th>
<th>Warm Blood</th>
<th>C(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>No</td>
<td>S</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>S</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>S</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>M</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>No</td>
<td>XXL</td>
<td>No</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>No</td>
<td>M</td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>L</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>S</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>S</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
</tr>
</tbody>
</table>
Why not just use the concept space?

How big is the number $2^{4000}$?

- Many approaches (e.g. Linear regression) limit the range of possible hypotheses.

- If you know something about the structure of a problem, you can limit the set of hypotheses you consider to be some tractable subset.

- Of course, if you’re wrong about the structure....
Deductive reasoning

• Tries to show a conclusion MUST follow from a set of premises (axioms)
• What we typically think of as “Logic”
  (1\textsuperscript{st} order, 2\textsuperscript{nd} order, etc.)
• Covered in EECS 348.
• Example
  All men are mortal.
  Socrates is a man.
  Therefore, Socrates is mortal.
Inductive reasoning

• The premises of an inductive argument indicate support (often probabilistic support) but do not ensure the conclusions are true.

• Example
  – 93% of students are right-handed.
  – Will is a student.
  – Therefore, Will is right-handed.
Inductive Bias

• NOT the same as bias in a statistical estimator

• DEFINITION: The set of axioms that would need to be added to the knowledge of the system so that a deductive reasoner would make the same inference as the inductive reasoner.

  – Example: Will does whatever the majority does.
Unbiased Learner

- Idea: Choose $H$ that expresses every teachable concept, that means $H$ is the set of all possible subsets of $X$

- $|X|=96$, therefore $|H|=2^{96} \sim 10^{28}$ concepts

- $H$ surely contains the target concept
- But there are too many concepts to pick them randomly to try
- Why not try them in some helpful order?
Unbiased Learner

Assume positive examples \((x_1, x_2, x_3)\) and negative examples \((x_4, x_5)\)

How would we classify some new instance \(x_6\)?

For any instance not in the training examples
    half of the version space says +
    the other half says –

* To learn the target concept one would have to present *every* single instance in \(X\) as a training example (Rote learning)
What kinds of biases are there?

- **Choice of data set**
  - e.g. Training an image classifier on photos from a foodie website means it won’t work well on car photos

- **Data representation**
  - How you code & represent the data has huge impact

- **Hypothesis space**
  - e.g. Linear regression only does straight lines and can’t fit a curve

- **Order in which we select hypotheses to test**
  - If your hypothesis space has $10^{10}$ hypotheses, you can’t try them all

- **Choice of performance measure**
  - Mean squared error? Maximum Margin? It makes a big difference
Summary

• Concept learning can be thought of as search through a space of hypotheses to find one (or more) that match the data.

• An unbiased learner cannot make inductive leaps to classify unseen examples.

• Inductive learning algorithms can classify unseen examples only because of inductive bias.

• There are biases in the learning algorithm, data representation, hypothesis space.