CSE 513T Theory of Artificial Intelligence & Machine Learning March 6, 2018

Lecture 15

Lecturer: Brendan Juba *Scribe: Christopher McKiernan*

Contents

0.1 Final Project 1

1 Agnostic Learning **1**
1.1 Introduction 1
1.2 Definition 2

2 Empirical Risk Minimization **3**
2.1 Introduction 3
2.2 Definition 3
2.3 Summary 4

3 Exception Tolearance and Common Sense **5**
3.1 Introduction 5

0.1 Final Project

Start thinking about your final project topics.

- Groups of ≤ 3
- Output is a 5 page report
- 10 min presentation with 5 minutes of QA

Between homework 8 and 9 you are to submit a less than 1 page project proposal that includes the names of your team members. Before writing your project proposal run your idea by the professor to make sure it is of valid scope for the course.

1 Agnostic Learning

1.1 Introduction

PAC Learning we assumed the labels were produced by some simple known class $\mathscr{C}$ (e.g. decision trees or halfspaces).

Real natural concepts generally arent defined by such simple representations.
Take for example a bird...

What is a bird?

- Flies
  - Planes fly are they a bird?
  - Not all birds fly e.g. penguins and ostriches
- Vertebrate
- Has feathers
  - If you pluck all the feathers off a chicken, is it still a bird?

A more extreme example of this is Wittgenstein's example

What is a game?

How can I tell when people are playing a game?
Is there anything I can say about their observed behavior that can characterizes that they are playing a game?

- They are smiling and laughing
  - What if they are playing a very serious game.
- Talking about a score
- There is a ball

These are all roughly indicative of people playing a game, but none of them is a perfect rule that defines a game. Can't describe perfectly with a decision tree.

Even though these are not perfect, we hope to learn rules like this inspite of counter examples necessarily occuring given enough examples. This is the objective of agnostic learning.

1.2 Definition

Definition 1 *The agnostic learning problem for a class of representations* $\mathscr{C} = \bigcup_{n\in\mathbb{N}} \mathscr{C}_n$ *(Where* $\mathscr{C}_n$ *is a class of representations over n attributes given as input* n, ϵ, δ *and given access to examples drawn from an arbritary distribution D over* $\mathscr{X}_n$ *and labeled by an arbritary function* $f : \mathscr{X}_n \to 0, 1$*, produce as output a representation of some efficiently evaluatable* $h \in \mathscr{H}$*, s.t. with probability* $1 - \delta$

$$\Pr_{x\in D}[h(x) \neq f(x)] \leq \min_{c\in\mathscr{C}_n} \Pr_{x\in D}[c(x) \neq f(x)] + \epsilon$$

More generally we may also consider $\alpha(n, \epsilon, \epsilon^*)$ *approxiamate agnostic learning, where it only requires that with probability* $1 - \delta$

$$\Pr_{x\in D}[h(x) \neq f(x)] \leq \alpha(n, \epsilon, \epsilon^*)$$

Where $\epsilon^* = \min_{c\in\mathscr{C}_n} \Pr_{x\in D}[c(x) \neq f(x)]$

Note: Usual PAC learning is the special case where $\epsilon^* = 0$(So even if $\epsilon* = 0$, α has better be $\geq \epsilon$)

To distinquish "usual" PAC learning from agnostic learning it is stimetimes called "relizable learning". While agnostic learning is somtimes called "non-relizable learning"

2 Empirical Risk Minimization

2.1 Introduction

Our usual apprach of finding a representation in some limited $\mathscr{H}$ that is consistent with a training set cannot work. For this to be possible at all $\mathscr{H}$ must have a finite VC-dimension, but then maybe $f \notin \mathscr{H}$, and then every $h \in \mathscr{H}$ will make errors if we draw enough data...

Natural generilzation find the $h \in \mathscr{H}$ with the lowest "error rate" on the training set.

2.2 Definition

Definition 2 *The empirical error of h on a list of labeled examples $(x^{(1)}, b^{(1)}, ..., (x^{(m)}, b^{(m)}))$ is*

$$\hat{err}(h) = \frac{1}{m}\#\{i : h(x^{(1)}) \neq b^{(i)}\}$$

Definition 3 *The empirical risk minimizer (EMR) of a class $\mathscr{H}$ and a list of labeled examples is any $h \in \mathscr{H}$ s.t.*

$$\hat{err}(h^*) = \min_{h \in \mathscr{H}} \hat{err}(h)$$

Key idea: *If we have enough data, empirical error $\approx$ true error, then choosing ERM suffices.*

Definition 4 *For representation class $\mathscr{H}$, distribution D, target property f, a list of labeled examples is said to be ϵ-represenative if,*

$$\text{for } err(h) \Pr_{x \in D}[h(x) \neq f(x)]$$
$$\text{for every } h \in \mathscr{H} \mid \hat{err}(h) - err(h) \mid \leq \epsilon$$

E.g. HW5#2 if H has VC-dimension d, a set of $m = O(\frac{1}{\epsilon^2}(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta}))$ examples is ϵ-representative with probability $1 - \delta$

Lemma 5 *For a $\frac{\epsilon}{2}$-representative set of examples for a class $\mathscr{H}$*
Any ERM $h^* \in \mathscr{H}$ satisfies

$$err(h^*) \leq \min_{h \in \mathscr{H}} err(h) + \epsilon$$

So such an h^* is satisfactory as a solution to agnostic learning [1] as long as $\mathscr{H} \geq \mathscr{C}$ since then

$$\min_{h \in \mathscr{H}_n} err(h) \leq \min_{c \in \mathscr{C}_n} err(c) = \epsilon^*$$

Proof *We have by defnition h^* satisfies*

$$\hat{err}(h^*) \leq \min_{h \in \mathscr{H}} \hat{err}(h) \leq \hat{err}(h^{**})$$

Where e.g. h^{**} achieves $\min_{h \in \mathscr{H}} err(h)$. Since the sample is $\frac{\epsilon}{2}$-represenative,

$$err(h^*) \leq \hat{err}(h^*) + \frac{\epsilon}{2}$$
$$err(h^{**}) \leq err(h) + \frac{\epsilon}{2} \text{ so ...}$$
$$err(h^*) \leq \hat{err}(h^*) + \frac{\epsilon}{2} \leq \min_{h \in \mathscr{H}} \hat{err}(h) + \frac{\epsilon}{2} \leq \min_{h \in \mathscr{H}} \{err(h) + \frac{\epsilon}{2}\} + \frac{\epsilon}{2} \leq \min_{h \in \mathscr{H}} err(h) + \epsilon$$

Unforunately finding an ERM is almost always NP-Complete.
An example of an ERM that is not NP-Complete follows:

Corollary 6 *The agnostic learning task for conjunctions of $\leq$ k literals (k=O(1)) can be solved in time*

$O(\frac{n^k}{\epsilon^2}(k \log n \log \frac{1}{\epsilon} + \log \frac{1}{\delta}))$ *using ...*

$O(\frac{1}{\epsilon^2}(k \log n \log \frac{1}{\epsilon} + \log \frac{1}{\delta}))$ *examples*

Why?
Enumerating the conjunction of k literals $O(n^k)$.
Choose one with smallest error.
VC dimension max # of shattered points d $\rightarrow$ must have $\geq 2^d$ labeling functions.
VC dimension for this class must satisfy $2d < O(n^k) \rightarrow d \leq k \log n + O(1)$
$\rightarrow$ only need quoted # of examples.

Doesn't necessarily mean agnostic leaning is NP-hard, maybe by using an even more expressive class of representations we can match the error rate of the ERM.

Still don't know efficient algorithms for most representations..
All we know algorithms for approx agnostic learning of conjunctions e.g. with $\alpha = n^{\frac{1}{3}}\epsilon^*$
HW 8 we'll do $O(n\epsilon^*)$

2.3 Summary

- Computational problem finding a low-error hypothesis -hard
- Statisically easy to get enough data to obtain good estimates for all representation in a class $\mathscr{H}$, can estimate "on demand" - reasoning with examples

3 Exception Tolearance and Common Sense

3.1 Introduction

We saw natural concepts aren't characterized by simple rules, must tolerate "counter examples" to rules. Hallmark of "common sense" reasoning" - "leaping to conclussions". rough rules for natural concepts... For example

Suppose I place a book in my bag. In the absence of further evidence to the contrary. I know my book is in the bag. Even though I can't see it. (no direct knowledge of it).
But I might later discover that the book is no longer there.

- Someone might steal it
- Hole in the bag
- Bag wasn't zipped up

It's still useful to assume that unless I know otherwise the book will still be there, known as "common sense law of inertia".

Also note additional knowledge may force me to retract conclussions, so knowledge is "nonmomotonic".

In contrast to classical knowledge classically adding representations to KB $\rightarrow$
fewer possible consistent assignment $\rightarrow$ more representations must be drawn as a conclussion. (valid wrt KB) so common sense reasoning is not governed by classical semantics.

But reasoning with examples will be able to exhibit this effect only because we can tolerate exceptions when drawining conclussions. i.e. relies on ability to certify
$(1-\epsilon)$ - validity when $\epsilon >> 0$, notciably large.

We'll view this operation of incorporating aditional evidence as filtering the set of examples or equivalently, reasoning with examples drawn from a conditional distribution.