Valiant's Theory

The PAC Model

Theory of Machine Learning – Spring 22

January 13, 2022

Last class

Logistics

- **Course webpage:** Canvas, can find lecture schedule, slides, scribe template, . . .
- TA: Chris Harker
- Scribe for today?

ML until the 1980s

- Many informal notions of learning: Rosenblatt and the “perceptron” algorithm, neural networks, . . .
- Limitations of perceptrons
- No formal theory to reason about, no clear definitions

Question: Can we formally define “learning”?

Theory of the Learnable

- Leslie Valiant 1983 – Theory of the Learnable (CACM)
- Drawing the boundaries of learnability – how to define it? what is possible?
- Really a theory of *supervised* learning
- I.e., deals with classification or *prediction* problems

- given some description of a "scenario". what to do next.?
- given an input, make prediction – label.

Theory of the Learnable

image: set of all pixel values

- **Input:** "features" of input
- **Hypothesis/model:** function from input to prediction/label

  hypothesis $h : I \to L$
  (all inputs) (all labels)
- **Definition of a learning algorithm.** an algorithm that can find a *good* hypothesis without explicitly being told what it is!

  - what all does a learning alg. need?

Most natural way. Give examples of inputs and ~~outputs~~ their labels

Good hypothesis:

"Low error": agreement with "true" label.

Qn: should it agree on all inputs?

Input: $x \rightsquigarrow$ collection of feature values.

$$\left(\quad , \quad , \; \dots \right) \in \mathbb{R}^m.$$

pixel 1-value

pixel 2 value

m pixel.

Ans: No, but we must have agreement on all "inputs of interest".

Good hypothesis?

(given examples)

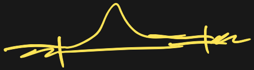
- Must do well on given inputs (hopefully perfectly)
- Must also do well on "unseen" inputs (generalization)
- **How to formalize this?**

Valiant's key assumption. Assume an "input distribution" (unknown to the learner)

We care about error "wrt" this distribution.

probability distribution on the spac of all inputs.

Good hypothesis



assuming there is a *true* label for each input

$\exists$ some (unknown to learner) distn $\mathcal{D}$ on the space of all inputs

Risk minimization is the goal:

Given a hypothesis h, a true label function ℓ, the risk of h w.r.t. a dist $\mathcal{D}$ $:= R_{\mathcal{D}}(h) = \Pr_{x \sim \mathcal{D}}[h(x) \neq \ell(x)]$.

Definition of learnability

$\to$ We say that a hypothesis ℓ is learnable, if $\forall \mathcal{D}$ for any $\varepsilon > 0$, there exists an $m(\varepsilon)$ (training size) such that given m iid examples $x_1, \dots, x_m$ from $\mathcal{D}$ and $\ell(x_1), \dots, \ell(x_m)$, we can produce a h s.t. $R_{\mathcal{D}}(h) \leq \varepsilon$, with prob $> 90\%$.

"We can produce"?

$\mathcal{D}_1, \mathcal{D}_2$

$h_1 \qquad h_2$

- $\exists$ an efficient algorithm A. (poly in m — # training examples)

  that takes $(x_1, \ell(x_1)), (x_2, \ell(x_2)) \dots (x_m, \ell(x_m))$

  outputs h.

★ Inherently a probabilistic statement.

★ Training samples come from same dist as test

Complexity of ground truth label

Importance of *hypothesis class*

($\mathbb{R}^m$) (labels, ± 1).

Label function $\ell : \mathcal{I} \to \mathcal{L}$

How "rich" can this fn be?

- Sample complexity ~~&~~ training time depend on "how complex" ℓ is.

± 1

$\text{sign}(x_1^2 + 3x_2 + x_4)$

- Assume: label function ℓ is in a certain ~~bigger~~ "hypothesis class" $\mathcal{H}$ (which is known to algo A).

$\mathcal{H}$: set of all polynomials of degree d in input features $x_1, \dots, x_m$.

Learnability with finite hypothesis classes

Theorem: (informal): Any finite hypothesis class $\mathcal{H}$ is learnable with $\sim \log|\mathcal{H}|$ training examples.

$\forall \varepsilon$, you can produce h such that

$$R_{\mathcal{D}}(h) \leq \varepsilon \quad \text{w.p.} \geq 90\%$$

using $\frac{\log|\mathcal{H}|}{\varepsilon^2}$ samples.