Machine Learning

Lecture 13: Computational Learning Theory

Overview

- Are there general laws that govern learning?
  - ***Sample Complexity:*** How many training examples are needed to learn a successful hypothesis?
  - ***Computational Complexity:*** How much computational effort is needed to learn a successful hypothesis?
  - ***Mistake Bound:*** How many training examples will the learner misclassify before converging to a successful hypothesis?

Some terms

X is the set of all possible instances

C is the set of all possible concepts c

where $c : X \rightarrow \{0,1\}$

H is the set of hypotheses considered

by a learner, $H \subseteq C$

L is the learner

D is a probability distribution over X

that generates observed instances

Definition

- The ***true error*** of hypothesis *h,* with respect to the target concept *c* and observation distribution *D* is the probability that *h* will misclassify an instance drawn according to *D*

$$error_D \equiv \underset{x \in D}{P}[c(x) \neq h(x)]$$

- In a perfect world, we'd like the true error to be 0

The world isn't perfect

- We typically can't provide every instance for training.
- Since we can't , there is always a chance the examples provided the learner will be misleading
  - "No Free Lunch" theorem
- So we'll go for a weaker thing:

  PROBABLY APPROXIMATELY CORRECT learning

Definition: PAC - learnable

A concept class **C** is “PAC learnable” by a hypothesis class **H** iff there exists a learning algorithm **L** such that..

….given any target concept **c** in **C**,

any target distribution **D** over the possible examples **X**,

and any pair of real numbers 0< ε, δ <1

… that **L** takes as input a training set of **m** examples drawn according to **D,** where the size of **m** is bounded above by a polynomial in $1/\varepsilon$ and $1/\delta$

… and outputs an hypothesis **h** in **H** about which we can say, with confidence (probability over all possible choices of the training set) greater than $1 - \delta$

…. that the error of the hypothesis is less than ε.

$$error_D \equiv \underset{x \in D}{P}[c(x) \neq h(x)] \leq \varepsilon$$

For *Finite* **Hypothesis Spaces**

- A hypothesis is ***consistent*** with the training data if it returns the correct classification for every example presented it.
- A ***consistent learner*** returns only hypotheses that are consistent with the training data.
- Given a consistent learner, the number of examples sufficient to assure that any hypothesis will be probably (with probability ***(1- δ)***) approximately (within error ε) correct is…

$$m \geq \frac{1}{\varepsilon}\left(\ln|H| + \ln(1/\delta)\right)$$

Theorem:

If the hypothesis space H is finite, and D is a sequence of $m \geq 1$ independent random examples of some target concept c, then for any $0 \leq \epsilon \leq 1$, the probability that $VS_{H,D}$ contains a hypothesis with error greater than ϵ is less than

$$|H|e^{-\epsilon m}$$

Proof sketch:

Prob(1 hyp. w/ error $> \epsilon$ consistent w/ 1 ex.) $< 1 - \epsilon \leq e^{-\epsilon}$

Prob(1 hyp. w/ error $> \epsilon$ consistent with m exs.) $< e^{-\epsilon m}$

Prob(1 of $|H|$ hyps. consistent with m exs.) $< |H|e^{-\epsilon m}$

Interesting! This bounds the probability that any consistent learner will output a hypothesis h with $error(h) \geq \epsilon$

If we want this probability to be at most δ

$$|H|e^{-\epsilon m} \leq \delta$$

then

$$m \geq \frac{1}{\epsilon}(\ln|H| + \ln(1/\delta))$$

Problems with PAC

- The PAC Learning framework has 2 disadvantages:
  - It can lead to weak bounds
  - Sample Complexity bound cannot be established for infinite hypothesis spaces

- We introduce the VC dimension for dealing with these problems (particularly the second one)

The VC-Dimension

- **<u>Definition:</u>** A set of instances S is shattered by hypothesis space H <u>iff</u> for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy.

- **<u>Definition:</u>** The Vapnik-Chervonenkis dimension, VC(H), of hypothesis space H defined over instance space X is the size of the largest finite subset of X shattered by H. If arbitrarily large finite sets of X can be shattered by H, then VC(H)=∞

Sample Complexity with VC

- ***Bound*** on sample complexity, using the ***VC-Dimension (Blumer et al. 1989)***:

$$m \geq \frac{1}{\varepsilon}\left(4\log_2(2/\delta) + 8VC(H)\log_2(13/\varepsilon)\right)$$

Sample Complexity for Infinite Hypothesis Spaces II

Consider any concept class ***C*** such that ***VC(C)*** $\geq$ ***2,*** any learner ***L,*** and any ***0*** < ε < ***1/8,*** and ***0*** < δ < ***1/100.*** Then there exists a distribution ***D*** and target concept in ***C*** such that if ***L*** observes fewer examples than

$$\max[1/\varepsilon \log(1/\delta),(VC(C)-1)/(32\varepsilon)]$$

then with probability at least δ, ***L*** outputs a hypothesis ***h*** having $error_D(h) > \varepsilon$.

The *Mistake Bound* Model of Learning

- Different from the PAC framework
- Considers learners that
  - receive a sequence of training examples
  - Predict the target value for each example
- The question asked in this setting is: ***"How many mistakes will the learner make in its predictions before it learns the target concept?***

Optimal Mistake Bounds

- $M_A(C)$ is the maximum number of mistakes made by algorithm A over all possible learning sequences before learning the concept C
- Let ***C*** be an arbitrary nonempty concept class. The optimal mistake bound for ***C***, denoted ***Opt(C)***, is the minimum over all possible learning algorithms ***A*** of $M_A(C)$.
  $Opt(C) = min_{A \in Learning\_Algorithm} M_A(C)$

Optimal Mistake Bounds

- For any concept class ***C***, the optimal mistake bound is bound as follows:

$$VC(C) \leq Opt(C) \leq \log_2(|C|)$$

A Case Study: The Weighted-Majority Algorithm

a_i denotes the i[th] prediction algorithm in the pool A of algorithm. w_i denotes the weight associated with a_i.

- For all i initialize w_i <-- 1
- For each training example $\langle x, c(x) \rangle$
  - Initialize q_0 and q_1 to 0
  - For each prediction algorithm ai
    - If $a_i(x)=0$ then q_0 <-- q_0+w_i
    - If $a_i(x)=1$ then q_1 <-- q_1+w_i
  - If $q_1 > q_0$ then predict $c(x)=1$
  - If $q_0 > q_1$ then predict $c(x) =0$
  - If $q_0=q_1$ then predict 0 or 1 at random for $c(x)$
  - For each prediction algorithm a_i in A do
    - If $a_i(x) \neq c(x)$ then w_i <-- βw_i

Relative Mistake Bound for the Weighted-Majority Algorithm

- Let ***D*** be any sequence of training examples, let ***A*** be any set of n prediction algorithms, and let ***k*** be the minimum number of mistakes made by any algorithm in ***A*** for the training sequence ***D***. Then the number of mistakes over ***D*** made by the ***Weighted-Majority*** algorithm using $\beta = 1/2$ is at most $2.4(k + \log_2 n)$.
- This theorem can be generalized for any $0 \leq \beta \leq 1$ where the bound becomes

$$(k \log_2 1/\beta + \log_2 n)/\log_2(2/(1+\beta))$$