Machine Learning 10-701

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

March 15, 2011

Today:

- Computational Learning Theory
- Mistake bounds

Recommended reading:

- Mitchell: Ch. 7
- suggested exercises: 7.1, 7.2, 7.7

Computational Learning Theory

What general laws constrain inductive learning?

We seek theory to relate:

- Probability of successful learning
- Number of training examples
- Complexity of hypothesis space
- Accuracy to which target function is approximated
- Manner in which training examples presented

* see Annual Conference on Learning Theory (COLT)

Mistake Bounds

So far: how many examples needed to learn?

What about: how many mistakes before convergence?

Let's consider similar setting to PAC learning:

- Instances drawn at random from X according to distribution $\mathcal{D}$ $= P(x)$
- Learner must classify each instance before receiving correct classification from teacher
- Can we bound the number of mistakes learner makes before converging?

Mistake Bounds: Find-S

$x = \langle x_1, x_2 \dots x_n \rangle$ boolean $y \in \{0,1\}$

e.g. $h = (x_2 = 1) \wedge (x_7 = 0) \rightarrow y = 1$

$= l_2 \quad \neg l_7 \rightarrow y = 1$

Consider Find-S when H = conjunction of boolean literals

Find-S:

- Initialize h to the most specific hypothesis $l_1 \wedge \cancel{\neg l_1} \wedge l_2 \wedge \cancel{\neg l_2} \dots \cancel{l_n} \wedge \neg l_n$
- For each positive training instance x
  - Remove from h any literal that is not satisfied by x
- Output hypothesis h.

Start with $2n$ lits.

mistake 1: remove n
= first + example

Mistake 2: remove 1 or more

⋮

k: 1

How many mistakes before converging to correct h? $\leq n+1$

Mistake Bounds: Halving Algorithm

1. Initialize VS ← H
2. For each training example,
   - remove from VS every hypothesis that misclassifies this example

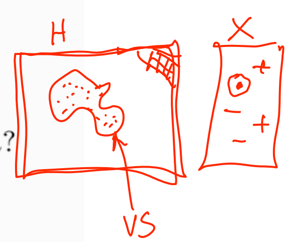
Consider the Halving Algorithm:

- Learn concept using version space CANDIDATE-ELIMINATION algorithm
- Classify new instances by majority vote of version space members

How many mistakes before converging to correct h?

- ... in worst case?
- ... in best case?

$\boxed{\frac{1}{2} \lfloor \log_2 |H| \rfloor}$



initial size of VS $= |H|$

after 1 mistake $\leq |H| \frac{1}{2}$

k mistakes $\leq |H| \left(\frac{1}{2}\right)^k \rightarrow k \leq \lfloor \log_2 |H| \rfloor$

Optimal Mistake Bounds

Let $M_A(C)$ be the max number of mistakes made by algorithm A to learn concepts in C. (maximum over all possible $c \in C$, and all possible training sequences)

$$M_A(C) \equiv \max_{c \in C} M_A(c)$$

Definition: Let C be an arbitrary non-empty concept class. The **optimal mistake bound** for C, denoted $Opt(C)$, is the minimum over all possible learning algorithms A of $M_A(C)$.

$$Opt(C) \equiv \min_{A \in learning\ algorithms} M_A(C)$$

$$\boxed{VC(C) \leq Opt(C) \leq M_{Halving}(C) \leq log_2(|C|).}$$

Weighted Majority Algorithm

a_i denotes the i^{th} prediction algorithm in the pool A of algorithms. w_i denotes the weight associated with a_i.

- For all i initialize $\boxed{w_i \leftarrow 1}$
- For each training example $\langle x, c(x) \rangle$
  - Initialize q_0 and q_1 to 0
  - For each prediction algorithm a_i
    - If $a_i(x) = 0$ then $q_0 \leftarrow q_0 + w_i$
      If $a_i(x) = 1$ then $q_1 \leftarrow q_1 + w_i$
  - If $q_1 > q_0$ then predict $c(x) = 1$
    If $q_0 > q_1$ then predict $c(x) = 0$
    If $q_1 = q_0$ 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 $w_i \leftarrow \beta w_i$

when β=0, equivalent to the Halving algorithm…

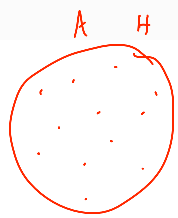
$\beta = 0.5$

Weighted Majority

Even algorithms that learn or change over time...

[Relative mistake bound for WEIGHTED-MAJORITY] 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 $\beta = \frac{1}{2}$ is at most

$$2.4(k + \log_2 n) \geq \text{\# mistakes by Wtd Maj}$$



---

let M be # of mistakes made by Wtd Maj. Alg using n algs.

k # " " by best $a_i \in A$

$W = \sum_i w_i$

What is final wt of alg a_i? $\left(\frac{1}{2}\right)^k$

What is final $\sum_{j=1}^{n} w_j$

What is initial $W = n$

after mistake #1, $W \le \frac{3}{4} n$

after mistake $M \rightarrow \left(\frac{1}{2}\right)^k \le W \le \left(\frac{3}{4}\right)^M n$

$$\boxed{w_i \le W}$$

$$\boxed{\left(\frac{1}{2}\right)^k \le \left(\frac{3}{4}\right)^M n}$$

What You Should Know

- Sample complexity varies with the learning setting
  - Learner actively queries trainer
  - Examples arrive at random
  - …
- Within the PAC learning setting, we can bound the probability that learner will output hypothesis with given error
  - For ANY consistent learner (case where $c \in H$)
  - For ANY "best fit" hypothesis (agnostic learning, where perhaps c not in H)
- VC dimension as measure of complexity of H
- Mistake bounds
- Conference on Learning Theory: http://www.learningtheory.org
- Avrim Blum's course on Machine Learning Theory:
  - http://www.cs.cmu.edu/~avrim/ML09/index.html