15-859(B) Machine Learning Theory

Avrim Blum

01/13/10

Lecture 2: Online learning I

Mistake-bound model:
- Basic results, halving and StdOpt algorithms
- Connections to information theory

Combining "expert advice":
- (Randomized) Weighted Majority algorithm
- Regret-bounds and connections to game-theory

Recap from last time

- Last time: PAC model and Occam's razor.
  - If data set has $m \geq (1/\varepsilon)[s \ln(2) + \ln(1/\delta)]$ examples, then whp any consistent hypothesis with size(h) < s has err(h) < ε.
  - Equivalently, suffices to have $s \leq (\varepsilon m - \ln(1/\delta))/\ln(2)$
  - "compression $\Rightarrow$ learning"
- [KV] book has esp. good coverage of this and related topics.
- Occam bounds $\Rightarrow$ any class is learnable if computation time is no object.

Online learning

- What if we don't want to make assumption that data is coming from some fixed distribution? Or any assumptions at all?
- Can no longer talk about past performance predicting future results.
- Can we hope to say anything interesting??

Idea: mistake bounds & regret bounds.

Mistake-bound model

- View learning as a sequence of stages.
- In each stage, algorithm is given *x*, asked to predict *f(x)*, and then is told correct value.
- Make no assumptions about order of examples.
- Goal is to bound total number of mistakes.

Alg A learns class C with mistake bound M if A makes $\leq$ M mistakes on any sequence of examples consistent with some f $\in$ C.

Mistake-bound model

Alg A learns class C with mistake bound M if A makes $\leq$ M mistakes on any sequence of examples consistent with some f $\in$ C.

- Note: can no longer talk about "how much data do I need to converge?" Maybe see same examples over again and learn nothing new. But that's OK if don't make mistakes either…
- Want mistake bound poly(n, s), where n is size of example and s is size of smallest consistent f $\in$ C.
- C is **learnable** in MB model if exists alg with mistake bound and running time per stage poly(n,s).

Simple example: disjunctions

- Suppose features are boolean: $X = \{0,1\}^n$.
- Target is an OR function, like $x_3 \vee x_9 \vee x_{12}$.
- Can we find an on-line strategy that makes at most n mistakes?
- Sure.
  - Start with $h(x) = x_1 \vee x_2 \vee ... \vee x_n$
  - Invariant: {vars in h} $\supseteq$ {vars in f}
  - Mistake on negative: throw out vars in h set to 1 in x. Maintains invariant and decreases |h| by 1.
  - No mistakes on positives. So at most n mistakes total.

Simple example: disjunctions

- Algorithm makes at most n mistakes.
- No deterministic alg can do better:

  1 0 0 0 0 0 0 + or – ?
  0 1 0 0 0 0 0 + or – ?
  0 0 1 0 0 0 0 + or – ?
  0 0 0 1 0 0 0 + or – ?
  ...

MB model properties

An alg A is "conservative" if it only changes its state when it makes a mistake.

Claim: if C is learnable with mistake-bound M, then it is learnable by a conservative alg.

Why?

- Take generic alg A. Create new conservative A' by running A, but rewinding state if no mistake is made.
- Still $\leq$ M mistakes because A still sees a legal sequence of examples.

MB learnable $\Rightarrow$ PAC learnable

Say alg A learns C with mistake-bound M.

Transformation 1:

- Run (conservative) A until it produces a hyp h that survives $\geq (1/\varepsilon)\ln(M/\delta)$ examples.
- Pr(fooled by any given h) $\leq \delta/M$.
- Pr(fooled ever) $\leq \delta$.

  Uses at most $(M/\varepsilon)\ln(M/\delta)$ examples total.

MB learnable $\Rightarrow$ PAC learnable

Say alg A learns C with mistake-bound M.

Transformation 2: $O(\varepsilon^{-1}[M + \ln(1/\delta)])$ examples

- Run conservative A for $O(\varepsilon^{-1}[M + \ln(1/\delta)])$ examples. Argue that whp at least one of hyps produced has error $\leq \varepsilon/2$.
- Test the M hyps produced on $O(\varepsilon^{-1}\ln(M/\delta))$ new examples and take the best.
- Wait on full analysis until we get to Chernoff bounds…

One more example…

- Say we view each example as an integer between 0 and 2^n-1.
- $C = \{[0,a] : a < 2^n\}$. (device fails if it gets too hot)
- In PAC model we could just pick any consistent hypothesis. Does this work in MB model?
- What would work?

What can we do with unbounded computation time?

- "Halving algorithm": take majority vote over all consistent $h \in C$. Makes at most $\lg(|C|)$ mistakes.
- What if C has functions of different sizes?
- For any (prefix-free) representation, can make at most 1 mistake per bit of target.
  - give each h a weight of $(\frac{1}{2})^{size(h)}$
  - Total sum of weights ≤ 1.
  - Take weighted vote. Each mistake removes at least ½ of total weight left.

What can we do with unbounded computation time?

- "Halving algorithm": take majority vote over all consistent $h \in C$. Makes at most $\lg(|C|)$ mistakes.
- What if we had a "prior" p over fns in C?
  - Weight the vote according to p. Make at most $\lg(1/p_f)$ mistakes, where f is target fn.
- What if f was really chosen according to p?
  - Expected number of mistakes $\leq \sum_h [p_h \cdot \lg(1/p_h)]$ = entropy of distribution p.

Is halving alg optimal?

- Not necessarily (see hwk).
- Can think of MB model as 2-player game between alg and adversary.
  - Adversary picks x to split C into $C_-(x)$ and $C_+(x)$. [fns that label x as – or + respectively]
  - Alg gets to pick one to throw out.
  - Game ends when all fns left are equivalent.
  - Adversary wants to make game last as long as possible.
- OPT(C) = MB when both play optimally.

Is halving alg optimal?

- Halving algorithm: throw out larger set.
- Optimal algorithm: throw out set with larger mistake bound.
- You'll think about this more on the hwk…

What if there is no perfect function?

Think of as $h \in C$ as "experts" giving advice to you. Want to do nearly as well as best of them in hindsight.

These are called "regret bounds".
➢Show that our algorithm does nearly as well as best predictor in some class.

We'll look at a strategy whose running time is $O(|C|)$. So, only computationally efficient when C is small.

Using "expert" advice

Say we want to predict the stock market.

- We solicit n "experts" for their advice. (Will the market go up or down?)
- We then want to use their advice somehow to make our prediction. E.g.,

Expt 1	Expt 2	Expt 3	neighbor's dog	truth
down	up	up	up	up
down	up	up	down	down
…	…	…	…	…

Can we do nearly as well as best in hindsight?

["expert" ≡ someone with an opinion. Not necessarily someone who knows anything.]
[note: would be trivial in PAC (i.i.d.) setting]

Using "expert" advice

If one expert is perfect, can get $\leq \lg(n)$ mistakes with halving alg.

But what if none is perfect? Can we do nearly as well as the best one in hindsight?

Strategy #1:

- Iterated halving algorithm. Same as before, but once we've crossed off all the experts, restart from the beginning.
- Makes at most $\lg(n)[\text{OPT}+1]$ mistakes, where OPT is #mistakes of the best expert in hindsight.

Seems wasteful. Constantly forgetting what we've "learned". Can we do better?

Weighted Majority Algorithm

Intuition: Making a mistake doesn't completely disqualify an expert. So, instead of crossing off, just lower its weight.

Weighted Majority Alg:

- Start with all experts having weight 1.
- Predict based on weighted majority vote.
- Penalize mistakes by cutting weight in half.

					prediction	correct
weights	1	1	1	1		
predictions	Y	Y	Y	N	Y	Y
weights	1	1	1	.5		
predictions	Y	N	N	Y	N	Y
weights	1	.5	.5	.5		

Analysis: do nearly as well as best expert in hindsight

- M = # mistakes we've made so far.
- m = # mistakes best expert has made so far.
- W = total weight (starts at n).
- After each mistake, W drops by at least 25%. So, after M mistakes, W is at most $n(3/4)^M$.
- Weight of best expert is $(1/2)^m$. So,

$$(1/2)^m \leq n(3/4)^M$$
$$(4/3)^M \leq n2^m$$
$$M \leq 2.4(m + \lg n)$$

constant ratio

Randomized Weighted Majority

2.4(m + lg n) not so good if the best expert makes a mistake 20% of the time. Can we do better? Yes.

- Instead of taking majority vote, use weights as probabilities. (e.g., if 70% on up, 30% on down, then pick 70:30) Idea: smooth out the worst case.
- Also, generalize $\frac{1}{2}$ to 1- ε.

Solves to: $M \leq \frac{-m\ln(1-\varepsilon)+\ln(n)}{\varepsilon} \approx (1+\varepsilon/2)m + \frac{1}{\varepsilon}\ln(n)$

M = expected #mistakes

$M \leq 1.39m + 2\ln n \quad \leftarrow \varepsilon = 1/2$

$M \leq 1.15m + 4\ln n \quad \leftarrow \varepsilon = 1/4$

$M \leq 1.07m + 8\ln n \quad \leftarrow \varepsilon = 1/8$

unlike most worst-case bounds, numbers are pretty good.

Analysis

- Say at time t we have fraction F_t of weight on experts that made mistake.
- So, we have probability F_t of making a mistake, and we remove an εF_t fraction of the total weight.
  - $W_{final} = n(1-\varepsilon F_1)(1 - \varepsilon F_2)\ldots$
  - $\ln(W_{final}) = \ln(n) + \sum_t [\ln(1 - \varepsilon F_t)] \leq \ln(n) - \varepsilon \sum_t F_t$ (using $\ln(1-x) < -x$)
    $= \ln(n) - \varepsilon M$. ($\sum F_t$ = E[# mistakes])
- If best expert makes m mistakes, then $\ln(W_{final}) > \ln((1-\varepsilon)^m)$.
- Now solve: $\ln(n) - \varepsilon M > m \ln(1-\varepsilon)$.

$$M \leq \frac{-m\ln(1-\varepsilon)+\ln(n)}{\varepsilon} \approx (1+\varepsilon/2)m + \frac{1}{\varepsilon}\log(n)$$

Summarizing

- E[# mistakes] $\leq (1+\varepsilon)$OPT + $\varepsilon^{-1}\log(n)$.
- If set $\varepsilon = (\log(n)/OPT)^{1/2}$ to balance the two terms out (or use guess-and-double), get bound of E[mistakes] $\leq OPT + 2(OPT \cdot \log n)^{1/2} \leq OPT + 2(T \log n)^{1/2}$
- Define average regret in T time steps as:
  (avg per-day cost of alg) – (avg per-day cost of best fixed expert in hindsight).
  Goes to 0 or better as $T \to \infty$ [= "no-regret" algorithm].

What can we use this for?

- Can use to combine multiple algorithms to do nearly as well as best in hindsight.
- Can apply RWM in situations where experts are making choices that cannot be combined.
  - Choose expert i with probability $p_i = w_i / \sum_i w_i$.
  - Experts could be different strategies for some task, or rows in a matrix game. (Alg generalizes to case where in each time step, each expert gets a cost in [0,1])

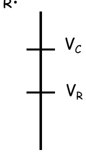
Minimax Theorem (von Neumann 1928)

- Every 2-player zero-sum game has a unique value V.
- Minimax optimal strategy for R guarantees R's expected gain at least V.
- Minimax optimal strategy for C guarantees C's expected loss at most V.

Counterintuitive: Means it doesn't hurt to publish your strategy if both players are optimal. (Borel had proved for symmetric 5x5 but thought was false for larger games)

Nice proof of minimax thm

- Suppose for contradiction it was false.
- This means some game G has $V_C > V_R$:
  - If Column player commits first, there exists a row that gets the Row player at least V_C.
  - But if Row player has to commit first, the Column player can make him get only V_R.
- Scale matrix so payoffs to row are in [-1,0]. Say $V_R = V_C - \delta$.



Proof, contd

- Now, consider playing randomized weighted-majority alg as Row, against Col who plays optimally against Row's distrib.
- In T steps,

  How can we think of RWM as an alg for repeatedly playing a matrix game???

  - Alg gets $\geq (1-\varepsilon/2)$[best row in hindsight] – $\log(n)/\varepsilon$
  - BRiH $\geq T \cdot V_C$ [Best against opponent's empirical distribution]
  - Alg $\leq T \cdot V_R$ [Each time, opponent knows your randomized strategy]
  - Gap is δT. Contradicts assumption if use $\varepsilon=\delta$, once $T > 2\log(n)/\varepsilon^2$.

A natural generalization

- A natural generalization of this setting: say we have a list of n prediction rules, but not all rules fire on any given example.
- E.g., document classification. Rule: "if <word-X> appears then predict <Y>". E.g., if has football then classify as sports.
- Natural goal: simultaneously, for each rule i, guarantee to do nearly as well as it *on the time steps in which it fires.*
  - For all i, want $E[cost_i(alg)] \leq (1+\varepsilon)cost_i(i) + O(\varepsilon^{-1}\log n)$.
- So, if 90% of documents with football *are* about sports, we should have error $\leq$ 11% on them.

"Specialists" or "sleeping experts" problem. Will get to this later…