TTIC 31250
An Introduction to the Theory of Machine Learning

Boosting

Avrim Blum

04/23/18

Boosting: a practical algorithmic tool and a statement about learning in the PAC model itself

Boosting, view #1

- Definition: Algorithm A is a weak-learner with edge γ for class C if: for any distribution D over examples labeled by some target $f \in C$, whp A produces a hypothesis h with $err_D(h) \leq 1/2 - \gamma$.

- Note: Ignoring δ parameter throughout the lecture since it can be handled easily (hwk 2).

- Theorem: Given a weak-learner A with edge γ for class C, we can produce an alg A' that achieves a PAC guarantee for class C (whp produces hypothesis with error $\leq \epsilon$) using $O\left(\frac{1}{\gamma^2}\log\frac{1}{\epsilon}\right)$ calls to A. A' is efficient if A is.

"Weak learning ⇒ Strong learning"

Boosting, view #2

- Imagine you want a highly accurate algorithm to predict y from x.

- So, you publish a large dataset S_1 of (x, y) pairs and ask if anyone can find an h_1 of error $\leq 40\%$. (And say we require h_1 to be "simple" so we know it's not overfitting)

- Now, you use h_1 to create a new dataset S_2 (by focusing more on the problematic data for h_1) and ask if anyone can find an h_2 of error $\leq 40\%$ on S_2.

- And so on.

- You can do this and combine the h_i s.t either (a) you drive your error down to 0 or else (b) you reach a hard dataset that nobody can do much better than random guessing on.

Preliminaries

- Assume we want to learn some unknown target function f over distribution D.
- Assume we have a weak-learner A with edge γ that uses hypotheses from some class of VC-dim d. (A should be able to achieve error $\leq 1/2 - \gamma$ for learning f over any reweighting of D)
- We will end up running A for T times producing hypotheses $h_1, \ldots, h_T$ and combining them into a single rule.
- By problem 3 on current hwk, the set of such combinations has VC-dim $O(Td \log Td)$.
- This will allow us to do all this on a sample of size $\tilde{O}\left(\frac{Td}{\epsilon}\right)$.

Preliminaries, contd.

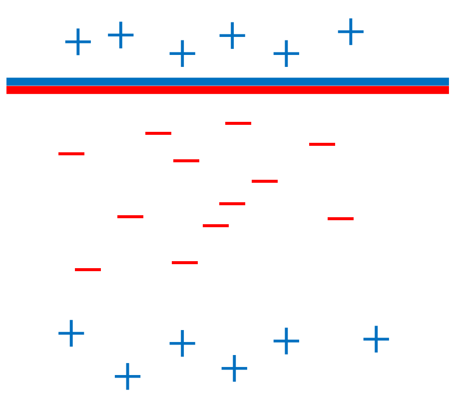
- We will draw a training sample S of size $m = \tilde{O}\left(\frac{Td}{\epsilon}\right)$.
- Assume that given any weighting of the points in S, A will return a hypothesis h of error at most $1/2 - \gamma$ over the distribution induced by that weighting. (ignoring δ)
- Will show can produce h with $err_S(h) = 0$ for $T = O\left(\frac{\log m}{\gamma^2}\right)$.
- Just need $m \gg \frac{d \log m}{\epsilon \gamma^2}$.

Boosting algo (Adaboost-light)

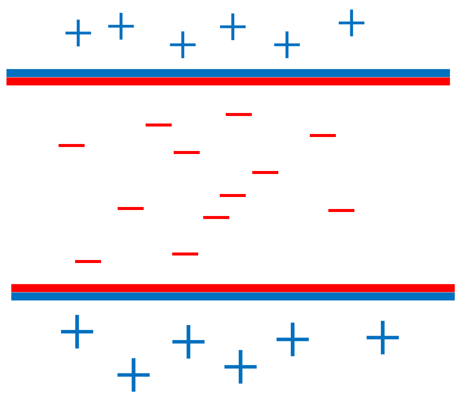
1. Given labeled sample $S = \{x_1, \ldots, x_m\}$, initialize each example x_i to have weight $w_i = 1$. Let $w = (w_1, \ldots, w_n)$.
2. For $t = 1, \ldots, T$ do:
   a. Call A on the distribution D_t over S induced by w.
   b. Receive hypothesis h_t of error $\leq 1/2 - \gamma$ over D_t.
   c. Multiply the weight of each example misclassified by h_t by $\alpha = \frac{0.5+\gamma}{0.5-\gamma}$. Leave the other weights alone.
3. Output the majority-vote classifier $MAJ(h_1, \ldots, h_T)$. Assume T is odd so no ties.

Thm: $T = O\left(\frac{\log m}{\gamma^2}\right)$ is sufficient s.t. $err_S\big(MAJ(h_1, \ldots, h_T)\big) = 0$.

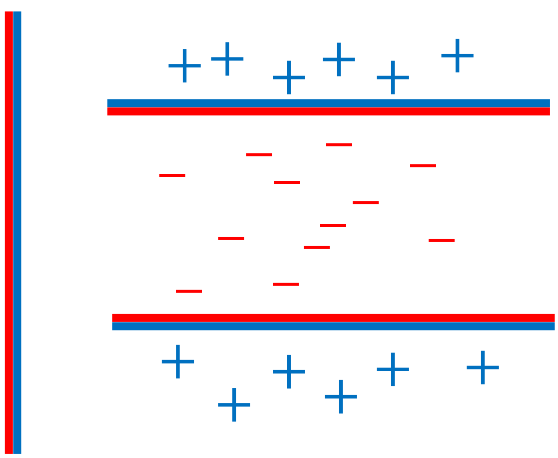
Example



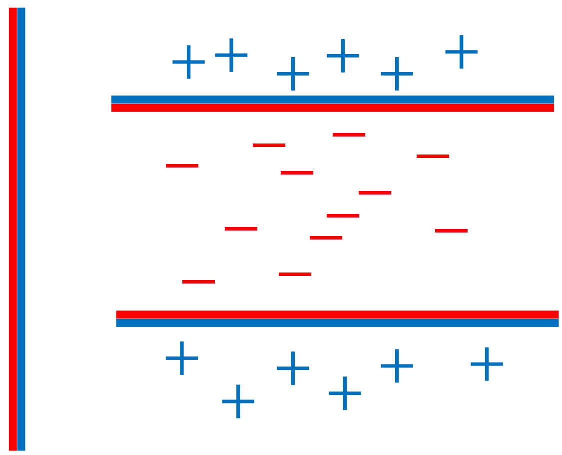
Example



Example



Example



Boosting algo (Adaboost-light)

1. Given labeled sample $S = \{x_1, \ldots, x_m\}$, initialize each example x_i to have weight $w_i = 1$. Let $w = (w_1, \ldots, w_n)$.
2. For $t = 1, \ldots, T$ do:
   a. Call A on the distribution D_t over S induced by w.
   b. Receive hypothesis h_t of error $\leq 1/2 - \gamma$ over D_t.
   c. Multiply the weight of each example misclassified by h_t by $\alpha = \frac{0.5+\gamma}{0.5-\gamma}$. Leave the other weights alone.
3. Output the majority-vote classifier $MAJ(h_1, \ldots, h_T)$. Assume T is odd so no ties.

Thm: $T = O\left(\frac{\log m}{\gamma^2}\right)$ is sufficient s.t. $err_S\big(MAJ(h_1, \ldots, h_T)\big) = 0$.

Boosting algo (Adaboost-light)

	x_1	x_2	x_3	.	.	.				x_m
h_1		X		X		X		X		
h_2		X	X				X	X		X
h_3	X	X				X			X	
…			X		X					
				X		X				
	X		X				X			
				X				X		X

"X" = mistake. Weight of $x_i = \alpha^{\#mistakes\ in\ column\ i}$

BTW, does this remind you of anything we've seen so far?

Proof of Boosting Theorem

Thm: $T = O\left(\frac{\log m}{\gamma^2}\right)$ is sufficient s.t. $err_S\big(MAJ(h_1, \dots, h_T)\big) = 0$.

Proof:

- First, if $MAJ(h_1, \dots, h_T)$ makes a mistake on any x_i then its final weight must be greater than $\alpha^{T/2}$.
- Let W_t be total weight after update t. $W_0 = m$.
- By the weak-learning assumption, h_t has error $\leq 1/2 - \gamma$ on D_t. So, at most $1/2 - \gamma$ fraction of weight multiplied by α.
- So, $W_{t+1} \leq \left(\alpha\left(\frac{1}{2} - \gamma\right) + \left(\frac{1}{2} + \gamma\right)\right) W_t = (1 + 2\gamma) W_t$.
- So if $err_S(\dots) > 0$ then $\alpha^{T/2} \leq W_T \leq (1 + 2\gamma)^T m$.

Proof of Boosting Theorem

Thm: $T = O\left(\frac{\log m}{\gamma^2}\right)$ is sufficient s.t. $err_S\big(MAJ(h_1, \dots, h_T)\big) = 0$.

Proof:

- Substituting $\alpha = \frac{1/2+\gamma}{1/2-\gamma} = \frac{1+2\gamma}{1-2\gamma}$ and rearranging, we get:

  $$1 \le (1-2\gamma)^{T/2}(1+2\gamma)^{T/2}m = (1-4\gamma^2)^{T/2}m \le e^{-2\gamma^2 T}m.$$

- Once $T > \frac{\ln m}{2\gamma^2}$, right-hand-side is less than 1. Done.

- So if $err_S(\dots) > 0$ then $\alpha^{T/2} \le W_T \le (1+2\gamma)^T m$.

Proof of Boosting Theorem

Thm: $T = O\left(\frac{\log m}{\gamma^2}\right)$ is sufficient s.t. $err_S\big(MAJ(h_1, \dots, h_T)\big) = 0$.

Proof:

- Substituting $\alpha = \frac{1/2+\gamma}{1/2-\gamma} = \frac{1+2\gamma}{1-2\gamma}$ and rearranging, we get:

  $$1 \le (1-2\gamma)^{T/2}(1+2\gamma)^{T/2}m = (1-4\gamma^2)^{T/2}m \le e^{-2\gamma^2 T}m.$$

- Once $T > \frac{\ln m}{2\gamma^2}$, right-hand-side is less than 1. Done.

- More generally, after any T steps, the fraction of mistakes is at most $e^{-2\gamma^2 T}$.

Some Reflections

- Suppose each h_t flipped a coin for each example x_i, predicting correctly with probability $1/2 + \gamma$.
  (I.e., suppose they all made *independent* errors)

- Then it's clear that taking majority vote is good. By Hoeffding, for any given x_i, $\Pr[MAJ\ is\ incorrect] \leq e^{-2\gamma^2 T}$.

So we actually just proved Hoeffding bounds, at least for $1/2 + \gamma\ \ vs\ 1/2$. (Take limit as # examples $\rightarrow \infty$, so that fraction of errors for each h_t matches expectation)

- More generally, after any T steps, the fraction of mistakes is at most $e^{-2\gamma^2 T}$.

More Reflections

- Consider a zero-sum game with examples as columns and hypotheses in H as rows.

	x_1	x_2	x_3	.	.	.				x_m
h_1		X		X		X		X		
h_2		X	X				X	X		X
h_3	X	X				X			X	
…			X		X					
				X		X				
	X		X				X			
				X				X		X

Rows represent all h in the class used by A

- If row plays h_i and column plays x_j then row wins if $h_i(x_j)$ is correct, and column wins if $h_i(x_j)$ is incorrect.

More Reflections

- Consider a zero-sum game with examples as columns and hypotheses in H as rows.

	x_1	x_2	x_3	.	.	.				x_m
h_1		X		X		X		X		
h_2		X	X				X	X		X
h_3	X	X				X			X	
…			X		X					
				X		X				
	X		X				X			
				X				X		X

- We are given that for any distrib D over columns (mixed strategy for the column player) there exists a row that wins with prob $\geq 1/2 + \gamma$ (payoff $\geq 1/2 + \gamma$)

More Reflections

- Consider a zero-sum game with examples as columns and hypotheses in H as rows.

	x_1	x_2	x_3	.	.	.				x_m
h_1		X		X		X		X		
h_2		X	X				X	X		X
h_3	X	X				X			X	
…			X		X					
				X		X				
	X		X				X			
				X				X		X

- By Minimax Thm, there exists a distribution P over h_i that wins with prob $\geq 1/2 + \gamma$ for any x_j.
- So, whp a large random sample from P will give correct vote on all x_j. (One way to see boosting is possible in principle)

More Reflections

- Consider a zero-sum game with examples as columns and hypotheses in H as rows.

	x_1	x_2	x_3	.	.	.				x_m
h_1		X		X		X		X		
h_2		X	X				X	X		X
h_3	X	X				X			X	
…			X		X					
				X		X				
	X		X				X			
				X				X		X

- In fact, this is just like RWM versus a best-response oracle, except our focus is on properties of the majority vote over the choices of the best-response oracle.

Margin Analysis

- Empirically noticed that you can keep running the booster past the point of perfect classification of S, and generalization doesn't get worse.

- One way to explain: "L_1 margins" or "margin of the vote"

Margin Analysis

Argument sketch:

- As $T \to \infty$, row player's strategy approaches minimax optimal (for all $x_j \in S$, $1/2 + \gamma$ of h_i vote correctly).
- Define h' as the randomized predictor: "given x, select $O\left(\frac{1}{\gamma^2}\log\frac{1}{\epsilon}\right)$ h_i at random from h and take their maj vote"
- So, $err_S(h') \leq \epsilon/2$.
- Also, $err_D(h') \geq err_D(h)/2$. (If $h(x)$ is wrong, then at least 50% chance that $h'(x)$ is wrong too)
- But h' isn't overfitting since whp no small majority-votes are overfitting and this is just a randomization over them. So h isn't overfitting by much either.