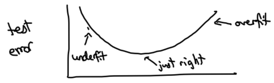
Something Unexpected: AdaBoost tends not to overfit

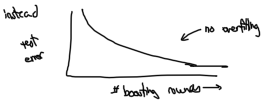
The Bias / Variance Curve we expect:



complexity ⟶
(approx by # boosting rounds)

Luckily, experimentalists ignored this advice!

Drucker & Cortes '96, Quinlan '96, Breiman '98 found curves this instead



Seems to contradict Occam's Razor!

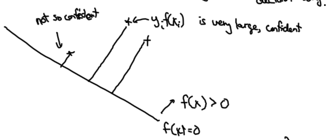
Explaination is in "Boosting the Margin" (Schapire, Freund, Bartlett Lee '98)

complexity ≠ # of boosting rounds

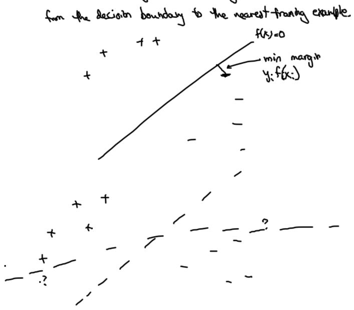
· complexity is related to the margin.

The margin of training example i is proportional to $y_i f(x_i)$.

- confidence measure of a classifier's ability
- "distance" from the training example to the decision bdry.



The minimum margin (or the margin of the classifier f) is the min of the margins over training examples, ie. the "distance" from the decision boundary to the nearest training example.



"Boosting the Margin" has 2 results:

1) Large Margins are good (statistical guarantee)

Pseudo theorem: w. high prob,

$$\text{misclassific error on the test set} \le \text{misclassific. error on training set} + \text{fctn}\left((\text{VCdim})^{\frac{1}{2}}, \frac{1}{\sqrt{m}}, \frac{1}{\text{margin}}\right)$$

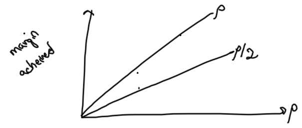
measure of complexity of a set of fctns

This is why it is believed that lg margins are the key to success.

2) AdaBoost achieves large margins

Pseudo Thm:

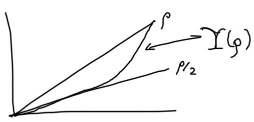
If the max. margin (for our data & weak classifiers) is ρ, AdaBoost acheives a margin of at least $\rho/2$



No more contradiction w/ Occam's Razor. But theory is still incomplete

Experimental results often indicated that AdaBoost maximizes the margin... except for Breiman. Breiman claimed his alg arc-gv achieved a max. margin & that AdaBoost did not (1999).

Rätsch & Warmuth (2005) tightened the bd slightly



If max margin is ρ, R & W proved AdaBoost acheives a margin of at least

$$\Upsilon(\rho) = \frac{-\ln(1-\rho^2)}{\ln\left(\frac{1+\rho}{1-\rho}\right)}.$$

It turns out that R & W were right!

Feb 21-2:37 PM

Recently, it has been shown :

1) AdaBoost does not necessarily maximize the margin (RDS 04)

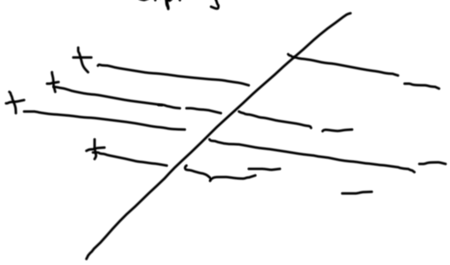
2) Rätsch & Warmuth's bound is tight, i.e. that AdaBoost (in a special case) can achieve a margin of exactly $\Upsilon(\rho)$. (RSD 06,07)

It also turns out that Breiman was right:

1) arc-gv does maximize the margin (Breiman, R & W) with a fast convergence rate (RSD 07)

2) AdaBoost still beats arc-gv experimentally. (Breiman, Reyzin & Schapire 06)

exploring dist of margins



In fact, there have been a number of other algorithms (w/ fast convergence rates) designed to maximize the margin, none of which has been shown to beat AdaBoost experimentally.

arc-gv (Breiman 99)

AdaBoost $_\rho$ & AdaBoost* (Rätsch & Warmuth 05)

Approx. Coord Ascent Boosting (RSD 07)

LP-AdaBoost (Grove & Schuurmans (98))

There are many unsolved problems - this is the current state of theoretical results on generalization of AdaBoost!

Experimental work - there are *lots* of variations of AdaBoost (almost all heuristically based, all claim to beat AdaBoost.)

One more important fact about AdaBoost:

It solves the <u>bipartite ranking</u> problems at the same time as it solves the problem of classification.

$L(\lambda)$

Feb 21-3:44 PM