Boosting

Say have lots of data (handwritten characters) and OK learning algorithm (achieves 70% accuracy). Want automatic method to either:

(A) boost it up, or

(B) find distribution where it fails badly.

Idea: give algorithm a sequence of "IQ tests."

PAC model question:

weak learning $\Rightarrow$ strong learning?

If exists alg A for class C such that for any dist. D, target $c \in C$, A achieves 51% accuracy, can you boost it to $1 - \epsilon$ accuracy?

Boosting [Schapire '89]

Step 1: Run alg on initial distribution $D_1 = D$ to get hypothesis h_1. Say accuracy = 70%.

Step 2: Filter D_1 to get new D_2 where hypothesis h_1 gets 50/50. (If h_1 predicts correctly, toss out with prob 4/7.)

$$\Pr_{D_2}(x) = \begin{cases} (5/7)\Pr_{D_1}[x] & \text{if } h_1 \text{ predicts correctly,} \\ (5/3)\Pr_{D_1}[x] & \text{if } h_1 \text{ predicts incorrectly.} \end{cases}$$

Get hypothesis h_2 with accuracy 70% on D_2.

Step 3: Filter D only allowing examples x where $h_1(x) \neq h_2(x)$. Get hypothesis h_3 with accuracy 70% on this distribution D_3.

Resulting hypothesis: majority vote.