Boosting

- Easy to come up with rough rules of thumb for classifying data
- E.g., for email,
  - Does it contain "!!!"?
  - Does it contain "buy now!"?
- Each alone isn't great, but better than random.
- **Boosting** converts rough rules of thumb into an accurate classifier.
- (Note: Boosting was invented by Prof. Schapire.)

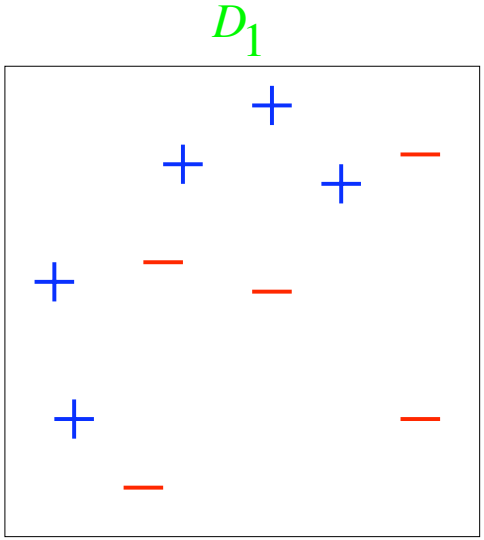
Sketch of the algorithm

- Devise a program for finding a weak hypothesis from data.
- Run it on the training data.
- Obtain the 1st weak hypothesis ("rule of thumb").
- Reweight the examples according to its accuracy
- Repeat T times to obtain 2nd, 3rd, ... weak hypotheses.
- At the end, combine the hypotheses into a classifier.

Boosting can drive the error down to ϵ

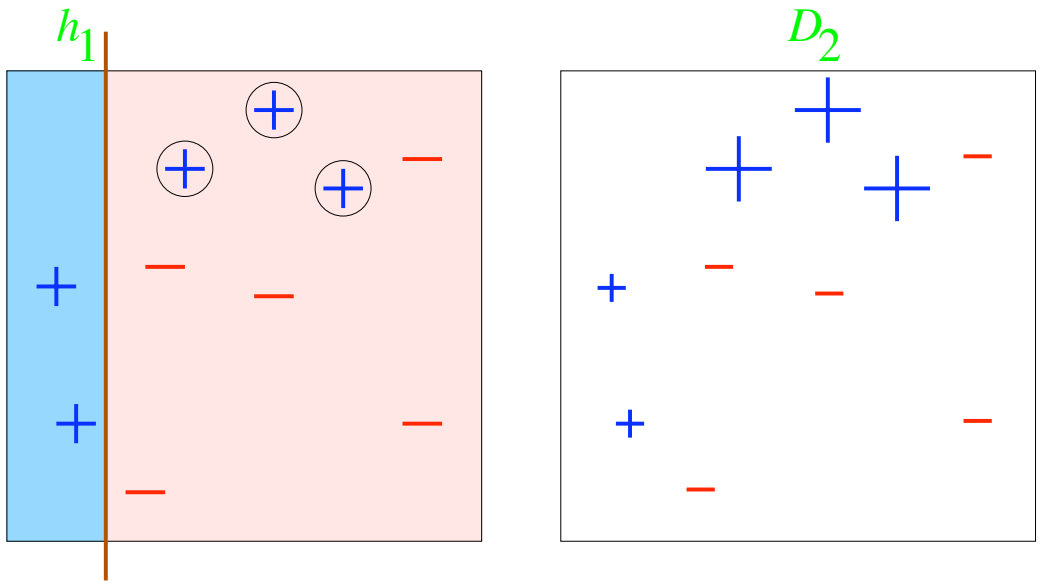
- We do as well as random, or we obtain training error 0.
- Empirically, boosting does well at test time too.

Toy Example

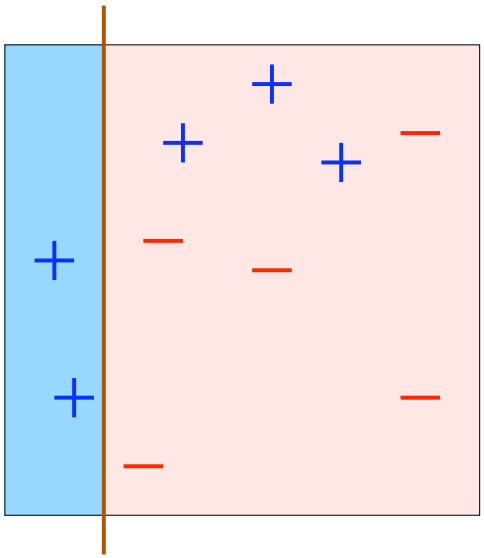


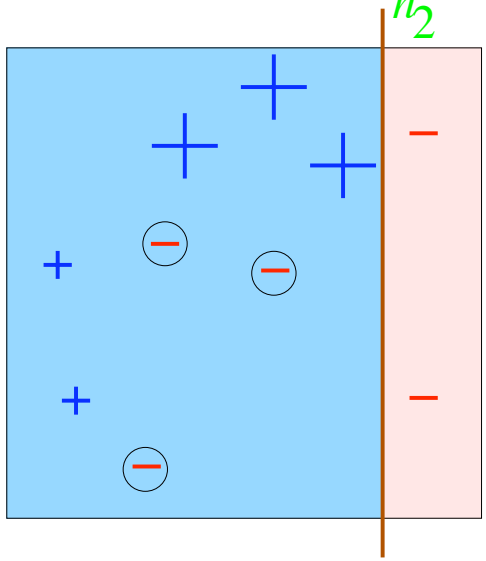
weak classifiers = vertical or horizontal half-planes

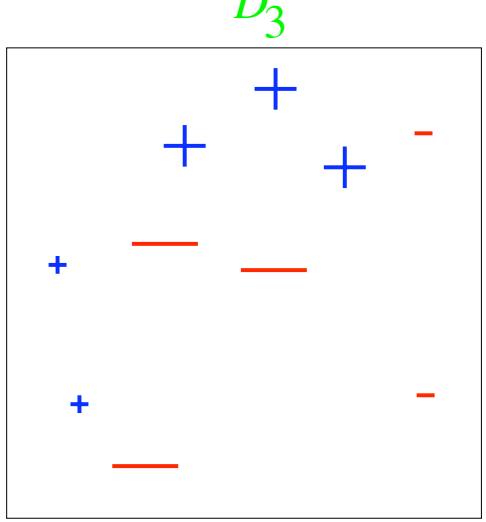
Round 1



Round 2



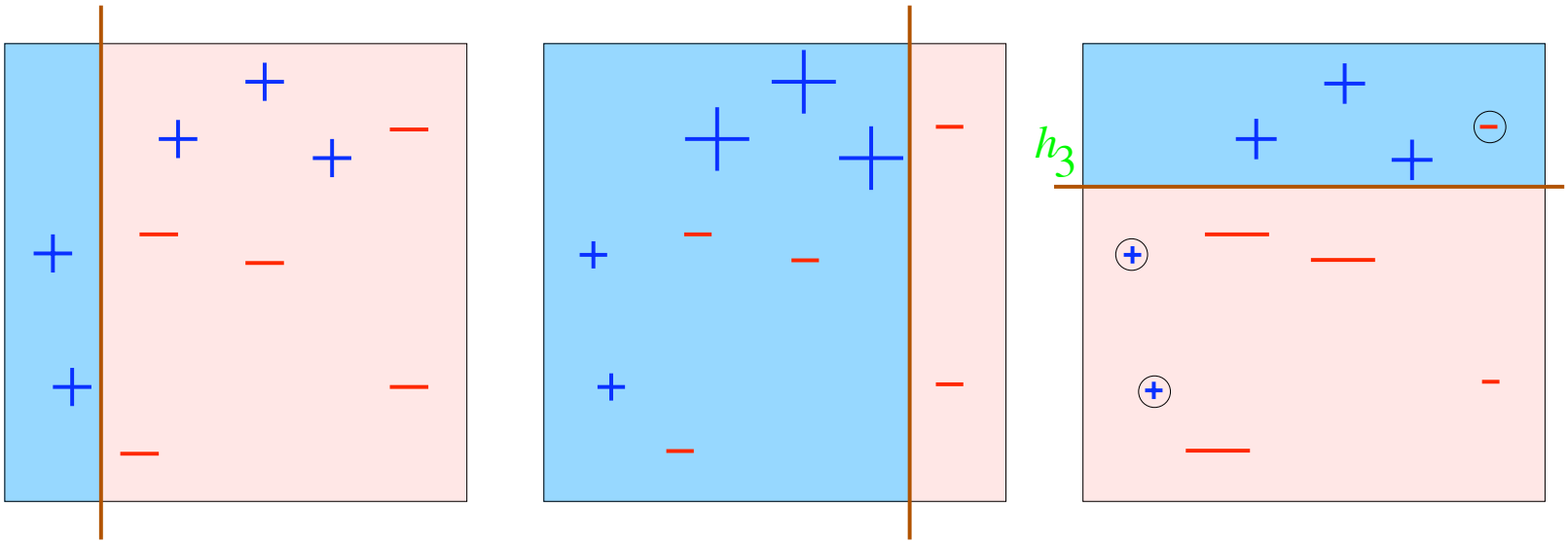




$\varepsilon_2 = 0.21$

$\alpha_2 = 0.65$

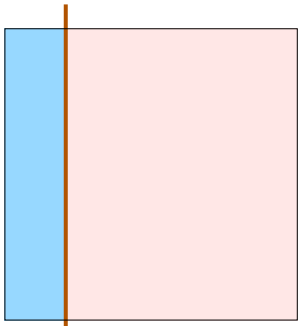
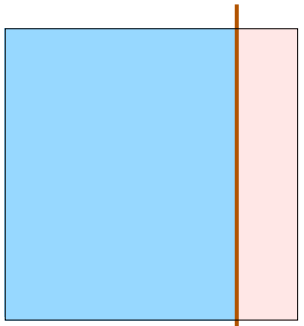
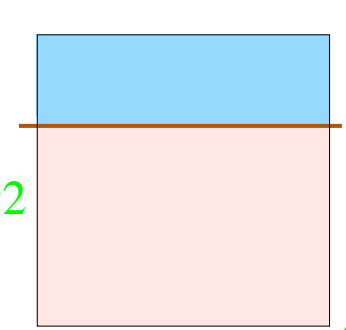
Round 3

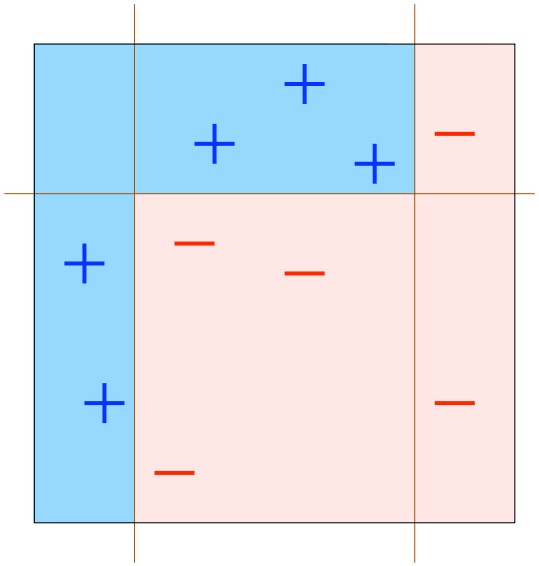


$\varepsilon_3 = 0.14$

$\alpha_3 = 0.92$

Final Classifier

$H_{\text{final}} = \text{sign}\left(0.42 \right.$  $+ 0.65$  $+ 0.92$  $\left. \right)$

$=$ 

Usenet newsgroups

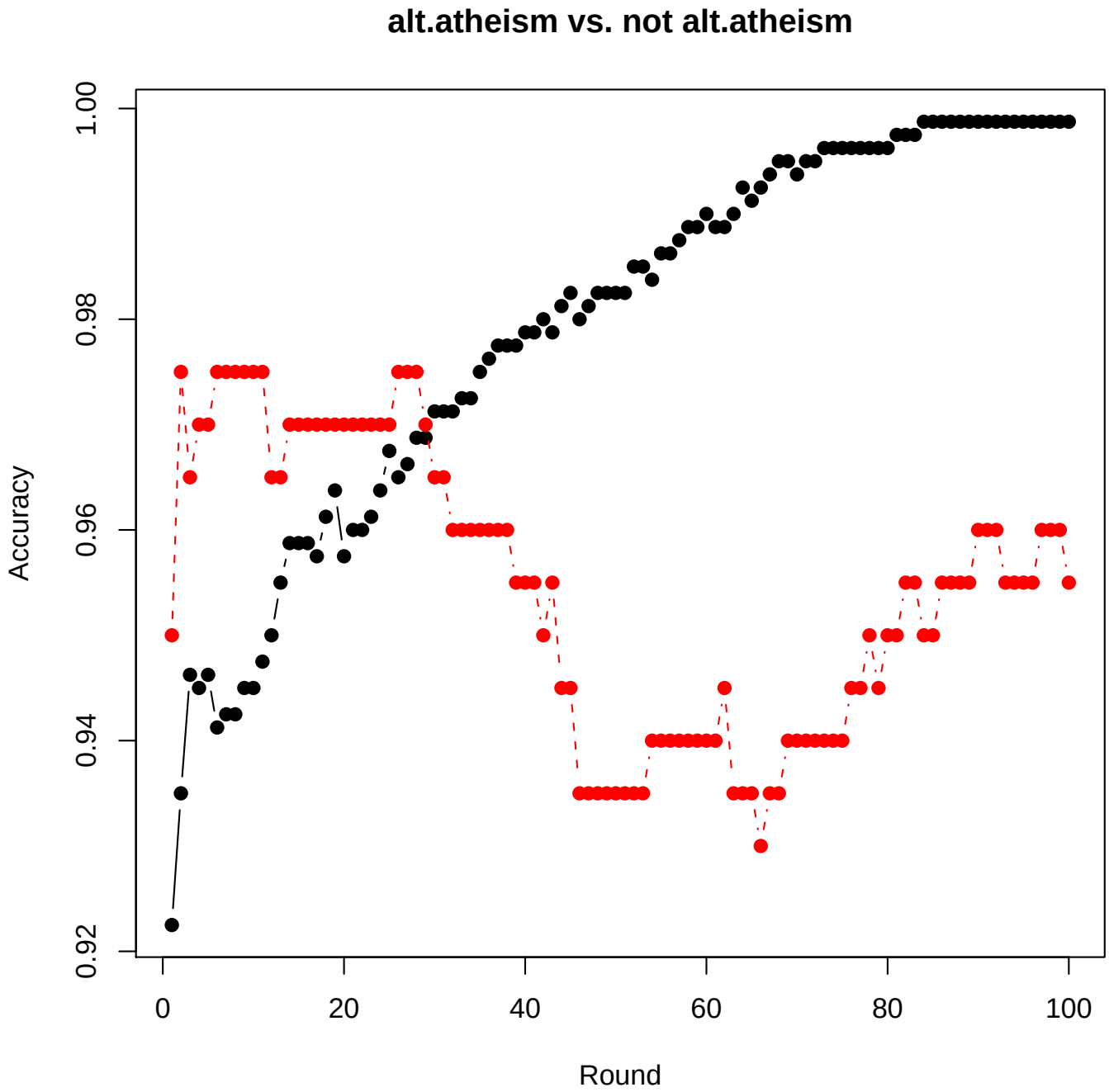
- Discussion forum from the beginning of the internet.
- Benchmark data set in text classification
- Goal: predict the forum of a given message

From keith [at] cco.caltech.edu (Keith Allan Schneider) Subject Re Political Atheists? Organization California Institute of Technology, Pasadena Lines 11 NNTP-Posting-Host punisher.caltech.edu arromdee[at]jyusenkyou.cs.jhu.edu (Ken Arromdee) writes The motto originated in the Star-Spangled Banner. Tell me that this has something to do with atheists. ¿The motto oncoins originated as a McCarthyite smear which equated atheism ¿with Communism and called both unamerican. No it didn't. The motto has been on various coins since the Civil War. It was just required to be on *all* currency in the 50s. keith

From: by028@cleveland.freenet.edu (Gary V. Cavano) Subject: Pantheism & Environmentalism Organization: Case Western Reserve University, Cleveland, Ohio (USA) Lines: 21 Hi... I'm new to this group, and maybe this has been covered already, but does anybody out there see the current emphasis on the environment being turned (unintentionally, of course) into pantheism? I've debated this quite a bit, and while I think a legitimate concern for the planet is a great thing, I can easily see it being perverted into something dangerous. As evidence, may I quote THE WALL STREET JOURNAL (of all things!), April 2 (Editorial page): "We suspect that's because one party to the (environmental) dispute thinks the Earth is sanctified. It's clear that much of the environmentalist energy is derived from what has been called the Religious Left, a SECULAR, or even PAGAN fanaticism that now WORSHIPS such GODS as nature and gender with a reverence formerly accorded real religions." (EMPHASIS MINE). Thoughts? Reactions? Harangues?

Binary classification

- In this example, we consider the 1000 X 1000 matrix M
- M_{ij} is the number of times the jth word occurred in the ith document. (You'll be analyzing these on your next homework.)
- AdaBoost with weak learners being the presence of single words.
- Classified articles from alt.atheism versus not alt.atheism.


alt.atheism vs. not alt.atheism
Accuracy
1.00
0.98
0.96
0.94
0.92
0
20
40
60
80
100
Round

Words added at each round

round 1	term 250 (atheists)	train=0.92 test=0.95
round 2	term 849 (islamic)	train=0.94 test=0.97
round 3	term 390 (moral)	train=0.95 test=0.96
round 4	term 33 (usa)	train=0.94 test=0.97
round 5	term 342 (keith)	train=0.95 test=0.97
round 6	term 19 (say)	train=0.94 test=0.97
round 7	term 316 (experience)	train=0.94 test=0.97
round 8	term 48 (year)	train=0.94 test=0.97
round 9	term 245 (mail)	train=0.94 test=0.97
round 10	term 480 (jon)	train=0.94 test=0.97
round 11	term 763 (germany)	train=0.95 test=0.97
round 12	term 157 (wrote)	train=0.95 test=0.96
round 13	term 470 (bob)	train=0.95 test=0.96
round 14	term 710 (islam)	train=0.96 test=0.97
round 15	term 99 (high)	train=0.96 test=0.97
round 16	term 503 (die)	train=0.96 test=0.97
round 17	term 857 (peace)	train=0.96 test=0.97