Improving Margin Bounds

John Langford

jcl@cs.cmu.edu

(work done at IBM Almaden with Nimrod Megiddo)

Mixture of Experts: Averaging is good

Common applied machine learning technique

- Avoid overfitting
- Reduce error

- setup
- standard margin bound
- improvements
- implications

The Setting

- Input space: X
- Hypothesis: $h : X \to \{-1, 1\}$
- Mixture: sign$(\sum_{i=1}^{k} a_i h_i(x))$
- constraint: $\sum_{i=1}^{k} a_i = 1$ with $a_i > 0$

Assumptions

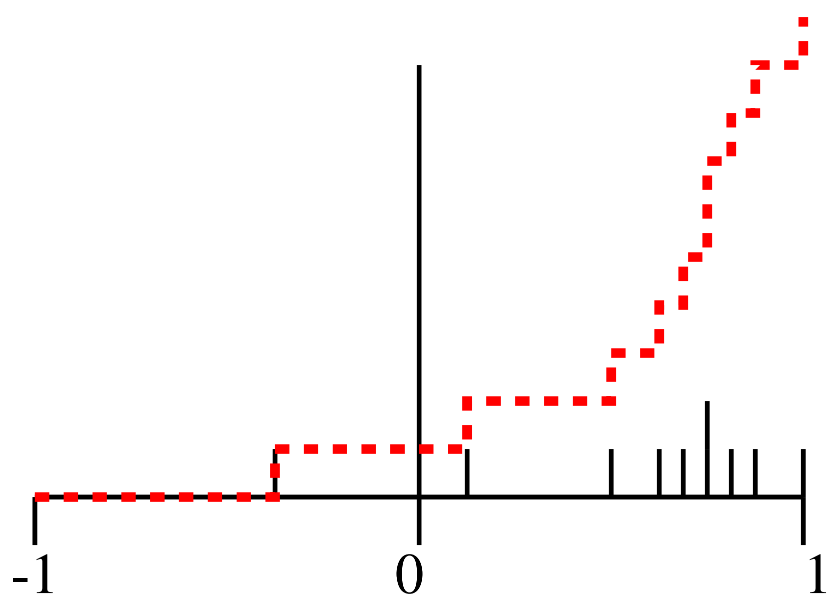
- finite set of base hypotheses: H

  (removable - but here for simplicity)

- m iid examples from unknown distribution D

Margin

Let $f(x) = \sum_{i=1}^{k} a_i h_i(x)$. Then margin $= yf(x)$



Margin Bound

(Schapire, Freund, Bartlett, Lee)

Pick N and pick margin θ

Result:

$\mathrm{Pr}_D(yf(x) < 0) \leq$

$$\Pr_S(yf(x) \leq \theta) + 2e^{-N\frac{\theta^2}{8}} + \sqrt{\frac{1}{2m}(N \ln|H| + \ln\frac{1}{\delta})}$$

Proof sketch

Let $g = \sum_{i=1}^{N} h_i(x)$ with h_i iid according to a_i.

Need to make 3 steps:

$$\mathsf{Pr}_D(yf(x) \leq 0) \leq \mathsf{Pr}_{D,g\sim f}(yg(x) < \tfrac{\theta}{2}) + e^{-N\frac{\theta^2}{8}}$$

$$\mathsf{Pr}_{D,g\sim f}(yg(x) < \tfrac{\theta}{2}) \leq \mathsf{Pr}_{S,g\sim f}(yg(x) < \tfrac{\theta}{2}) + \sqrt{\frac{1}{2m}(N \ln |H| + \ln \frac{1}{\delta})}$$

$$\mathsf{Pr}_{S,g\sim f}(yg(x) < \tfrac{\theta}{2}) \leq \mathsf{Pr}_S(yf(x) < \theta) + e^{-N\frac{\theta^2}{8}}$$

Proof

$$\Pr_D(yf(x) \le 0) = \Pr_{D,g\sim f}(yf(x) \le 0)$$

$$\le \Pr_{D,g\sim f}(yg(x) \le \frac{\theta}{2}) + \Pr_{D,g\sim f}(yg(x) > \frac{\theta}{2}, yf(x) \le 0)$$

$$\le \Pr_{g\sim f,D}(yg(x) \le \frac{\theta}{2}) + \Pr_{D,g\sim f}(yg(x) > \frac{\theta}{2} | yf(x) \le 0)$$

$$\le \Pr_{g\sim f,S}(yg(x) \le \frac{\theta}{2}) + \sqrt{\frac{1}{2m}(N \ln |H| + \ln \frac{1}{\delta})} + e^{-N\frac{\theta^2}{8}}$$

$$\le \Pr_{S,g\sim f}(yf(x) \le \theta) + \Pr_{S,g\sim f}(yg(x) \le \frac{\theta}{2} | yf(x) > \theta) + \ldots$$

$$\le \Pr_S(yf(x) \le \theta) + 2e^{-N\frac{\theta^2}{8}} + \sqrt{\frac{1}{2m}(N \ln |H| + \ln \frac{1}{\delta})}$$

Improved Margin Bound

Let $H(Q_f)$ = entropy on the distribution over H imposed by f.

$$\mathrm{Pr}_D(yf(x) < 0) \le \mathrm{Pr}_S(yf(x) \le \theta) + 2e^{-N\frac{\theta^2}{8}}$$

$$+\sqrt{\frac{1}{2m-1}(N(\ln|H| - H(Q_f)) + \ln\frac{1}{\delta} + \ln m + 2)}$$

The critical step in the proof is relating:

$\Pr_{D,g\sim f}(yg(x) < \frac{\theta}{2})$ to $\Pr_{S,g\sim f}(yg(x) < \frac{\theta}{2})$

From McAllester at Colt99:

Let P be a prior on the hypothesis set H:

$\Pr_{D,h\sim Q}(h(x) \neq y) <$

$$\Pr_{S,h\sim Q}(h(x) \neq y) + \sqrt{\frac{D(Q||P) + \ln\frac{1}{\delta} + \ln m + 2}{2m-1}}$$

$D(Q||P)$ = KL-divergence between Q and P.

We will use P = uniform on all sets of N hypotheses and Q according to N choices from f.

Let l = number of base hypotheses. Then:

$D(Q||P) = \sum_{i_1,\ldots,i_N=1} a_{i_1} \ldots a_{i_N} \ln \frac{a_{i_1} \ldots a_{i_N}}{\frac{1}{|H|^N}}$

$= \sum_{i_1,\ldots,i_N=1} a_{i_1} \ldots a_{i_N} \ln a_{i_1} \ldots a_{i_N} +$

$$\sum_{i_1,\ldots,i_N} a_{i_1} \ldots a_{i_N} N \ln |H|$$

$= \sum_{i_1,\ldots,i_N=1} a_{i_1} \ldots a_{i_N} \ln a_{i_1} \ldots a_{i_N} + N \ln |H|$

$= N(\ln |H| - H(Q_f))$

Implications

$$\Pr_D(yf(x) < 0) \leq \Pr_S(yf(x) \leq \theta)$$

$$+\sqrt{\frac{1}{2m-1}(N(\ln|H| - H(Q_f)) + \ln\frac{1}{\delta} + \ln m + 2)}$$

Larger margin: good (already known)

More base hypotheses: good (known empirically)

More uniform base hypotheses: good

For the future

- How tight can the margin bound be made in practice on real problems?

  (with other improvements, maybe tight enough to be useful on alternating decision trees)

- Bound based upon margin distribution rather than margin?