Margin-Based Generalization Lower Bounds for Boosted Classifiers

Allan Grønlund, Lior Kamma, Kasper Green Larsen, Alexander Mathiasen and Jelani Nelson

AARHUS UNIVERSITY

Berkeley UNIVERSITY OF CALIFORNIA

Boosting Algorithms

- Construct strong classifiers out of weak ones.

Accurate

Slightly better than guessing

Boosting Algorithms

- Construct strong classifiers out of weak ones.

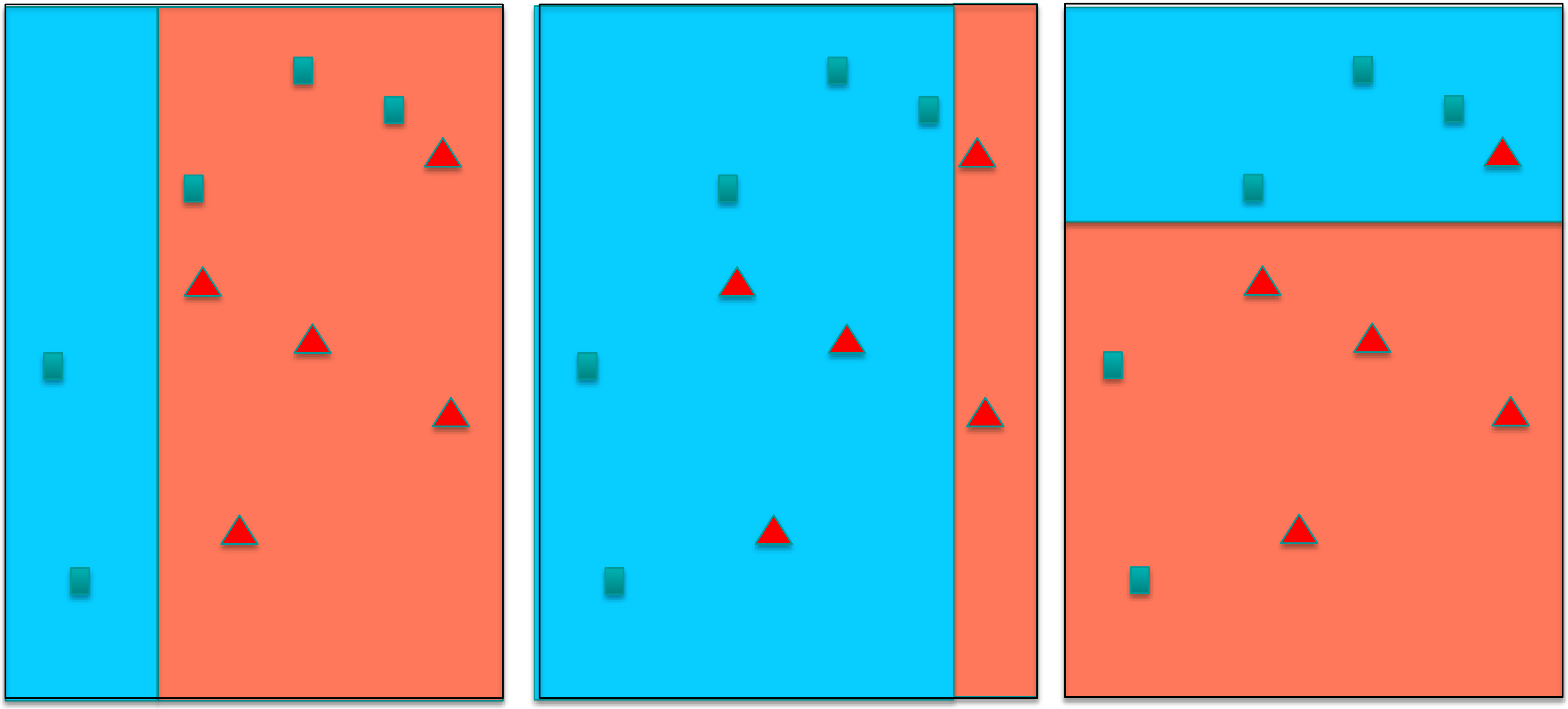
By combining them into a powerful “ensemble”

Boosting Algorithms

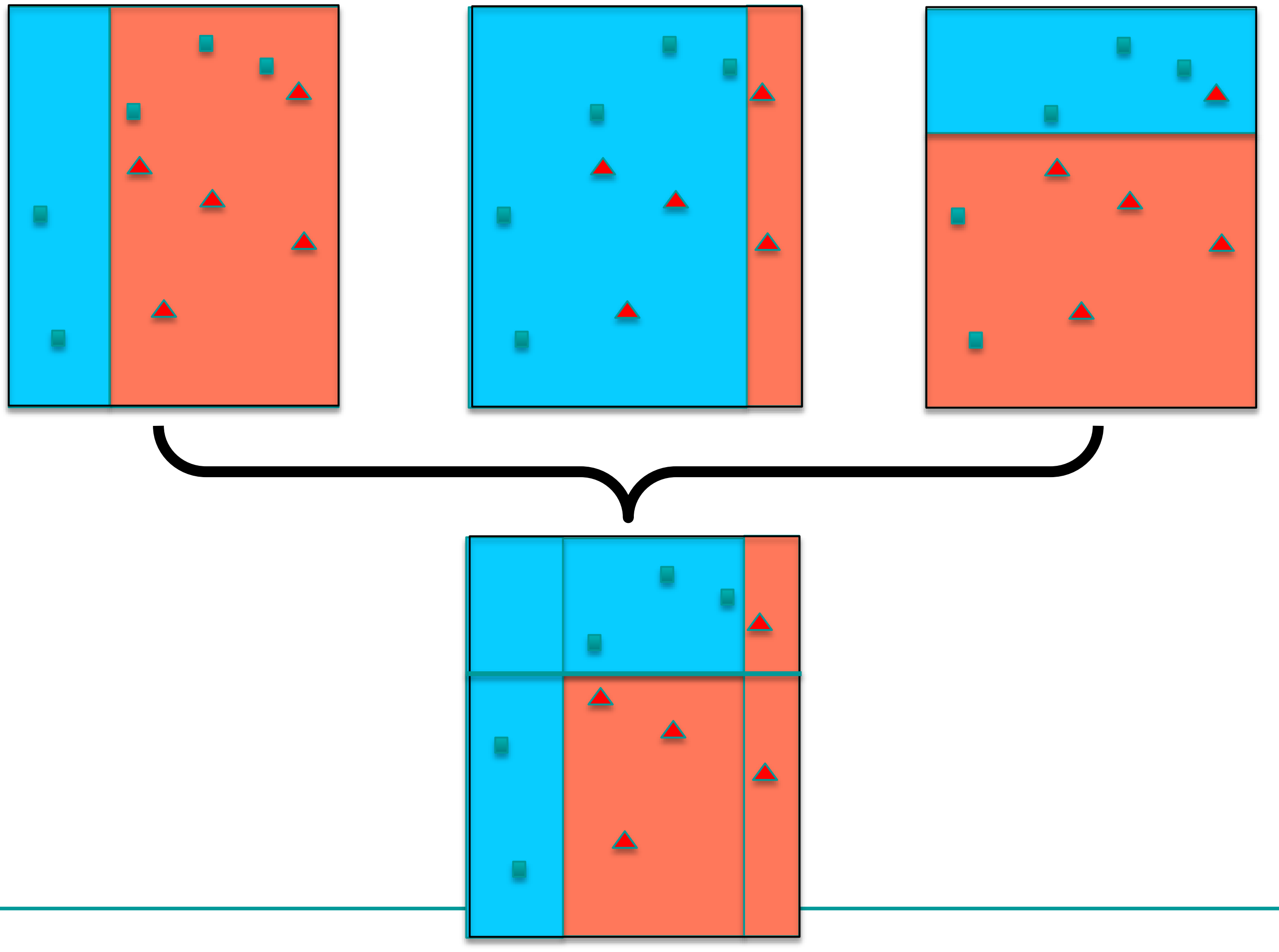
- Construct strong classifiers out of weak ones.
- Intuition: Train many weak classifiers, each "focusing" on a different part of the input space.

Achieved by re-weighing the input sample

Example : Axis Aligned Lines



Example : Axis Aligned Lines



Boosting Algorithms and Margins

- Surprising phenomenon : Even though the strong classifier gets more complicated, it does not overfit.

Boosting Algorithms and Margins

- Surprising phenomenon : Even though the strong classifier gets more complicated,

> Observed in experiments by Schapire *et al.*

Boosting Algorithms and Margins

- Surprising phenomenon : Even though the strong classifier gets more complicated, it does not overfit.

That is, more weak classifiers are involved

Boosting Algorithms and Margins

- Surprising phenomenon : Even though the strong classifier gets more complicated, it does not overfit.

- Prominent explanation : Margin Theory

Loosely speaking, the “confidence” of the classifier on a point.

Margin Theory

- Formally, let $\mathcal{H} \subseteq \mathcal{X} \rightarrow \{-1,1\}$ be the space of weak classifiers, and $S = \{(x_j, y_j)\}_{j=1}^{m}$ is the sample used to train a strong classifier $f = \sum_{h \in \mathcal{H}} \alpha_h h$.
- The margin of f on the j^th sample point is defined as $\theta_j \coloneqq y_j f(x_j)$

Margin Theory

- Formally, let $\mathcal{H} \subseteq \mathcal{X} \to \{-1,1\}$ be the space of weak classifiers, and $S = \{(x_j, y_j)\}_{j=1}^{m}$ is the sample used to train a strong classifier $f = \sum_{h \in \mathcal{H}} \alpha_h h$.

A convex combination of weak classifiers.

- The mar [illegible] point is defined as [illegible]

Margin Theory

- Formally, let $\mathcal{H} \subseteq \mathcal{X} \to \{-1,1\}$ be the space of weak classifiers, and $S = \{(x_j, y_j)\}_{j=1}^{m}$ is the sample used to train a strong classifier $f = \sum_{h \in \mathcal{H}} \alpha_h h$.

f is called a voting-classifier

- The mar… point is defined as $\theta_j := y_j f(x_j)$

Margin Theory

- Formally, let $\mathcal{H} \subseteq \mathcal{X} \to \{-1,1\}$ be the space of weak classifiers, and $S = \{(x_j, y_j)\}_{j=1}^{m}$ is the sample used to train … $_{\in \mathcal{H}} \alpha_h h$.

If θ_j is positive, then $\mathrm{sign}(f)$ classifies (x_j, y_j) correctly.

- The margin of f on the jth sample point is defined as $\theta_j := y_j f(x_j)$

Margin Theory

- Formally, let $\mathcal{H} \subseteq \mathcal{X} \rightarrow \{-1,1\}$ be the space of weak classifiers, and $S = \{(x_j, y_j)\}_{j=1}^{m}$ is the sample used to train ... $\alpha_h h$.

Intuitively, the closer θ_j is to 1, the more "confident" f is.

- The margin of f on the j^{th} sample point is defined as $\theta_j := y_j f(x_j)$

Margin-Based *Upper* Bounds

- Schapire *et al.* (1998) showed the following bound on the error probability of voting classifiers.

$$\Pr_{(x,y)\sim\mathcal{D}}[yf(x) \le 0]$$

$$\le \Pr_{(x,y)\sim S}[yf(x) \le \theta] + O\left(\sqrt{\frac{\ln|\mathcal{H}|\ln m}{m\theta^2}}\right)$$

Margin-Based *Upper* Bounds

- Schapire *et al.* (1998) showed the following bound on the error probability of voting classifiers.

$$\Pr_{(x,y)\sim\mathcal{D}}[yf(x) \le 0]$$

$$\left(\sqrt{\frac{\;|\mathcal{H}|\ln m}{m\theta^2}} \right)$$

The error probability of f with respect to the unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1,1\}$.

Margin-Based *Upper* Bounds

- Schapire *et al.* (1998) showed the following bound on the error probability of voting classifiers.

$$\Pr_{(x,y)\sim\mathcal{D}}[yf($$

The fraction of sample points with margin at most θ.

$$\le \Pr_{(x,y)\sim S}[yf(x) \le \theta] + O\left(\sqrt{\frac{\ln|\mathcal{H}|\ln m}{m\theta^2}}\right)$$

Margin-Based *Upper* Bounds

- Schapire *et al.* (1998) showed the following … ity of voting cla…

Holds for all voting classifiers f and margins $\theta \in (0,1]$

$$\Pr_{(x,y)\sim\mathcal{D}}[yf(x) \le 0]$$

$$\le \Pr_{(x,y)\sim S}[yf(x) \le \theta] + O\left(\sqrt{\frac{\ln|\mathcal{H}|\ln m}{m\theta^2}}\right)$$

Margin-Based *Upper* Bounds

This holds with high probability over the choice of the m sample points

… wed the … or probability of vo… assifiers.

$$\Pr_{(x,y)\sim\mathcal{D}}[\ldots(x) \le 0]$$

$$\le \Pr_{(x,y)\sim S}[yf(x) \le \theta] + O\left(\sqrt{\frac{\ln|\mathcal{H}|\ln m}{m\theta^2}}\right)$$

Margin-Based *Upper* Bounds

- Schapire ... following ... ty of voting cla...

The result gave rise to boosting algorithms that intentionally aim to optimize margins

$$\Pr_{(x,y)\sim\mathcal{D}}[yf(x) \le 0]$$

$$\le \Pr_{(x,y)\sim S}[yf(x) \le \theta] + O\left(\sqrt{\frac{\ln|\mathcal{H}|\ln m}{m\theta^2}}\right)$$

Margin-Based *Upper* Bounds

- Breimann (1999) showed the following bound on the error probability of voting classifiers.

$$\Pr_{(x,y)\sim\mathcal{D}}[yf(x) \leq 0] \leq O\left(\frac{\ln|\mathcal{H}|\ln m}{m\hat{\theta}^2}\right)$$

Holds for all voting classifiers f where $\hat{\theta}$ is the minimum margin

Margin-Based *Upper* Bounds

- Breimann… bound on… classifiers…

This holds with high probability over the choice of the m sample points

$$\Pr_{(x,y)\sim\mathcal{D}}[yf(x) \le 0] \le O\left(\frac{\ln|\mathcal{H}|\ln m}{m\hat{\theta}^2}\right)$$

Holds for all voting classifiers f where $\hat{\theta}$ is the minimum margin

Margin-Based *Upper* Bounds

- State-of-the-Art bounds were given by Gao and Zhou (2013)

$$\Pr_{(x,y)\sim\mathcal{D}}[yf(x) \leq 0] \leq \Pr_{(x,y)\sim S}[yf(x) \leq \theta] + O\left(\frac{\ln|\mathcal{H}|\ln m}{m\theta^2} + \sqrt{\frac{\ln|\mathcal{H}|\ln m}{m\theta^2} \Pr_{(x,y)\sim S}[yf(x) \leq \theta]}\right)$$

Margin-Based *Upper* Bounds

This holds with high probability over the choice of the m sample points

... ere given by

$$\Pr_{(x,y)\sim\mathcal{D}}[yf(x) \le 0] \le \Pr_{(x,y)\sim S}[yf(x) \le \theta]$$

$$+O\left(\frac{\ln|\mathcal{H}|\ln m}{m\theta^2} + \sqrt{\frac{\ln|\mathcal{H}|\ln m}{m\theta^2}\Pr_{(x,y)\sim S}[yf(x) \le \theta]}\right)$$

Holds for all voting classifiers f and margins $\theta \in (0,1]$

Margin-Based *Lower* Bounds?

- Despite being studied for over two decades, the tightness of margin-based generalization bounds was not settled.
- In fact, no margin-based lower bounds were known.

Margin-Based *Lower* Bounds!

- Our main result shows that any algorithm $\mathcal{A}$ optimizing margins cannot do much better than the known upper bounds.

Margin-Based *Lower* Bounds

- Formally, $\forall N, \theta, \tau$ There exist a set $\mathcal{X}$ and a hypothesis set $\mathcal{H}$ such that for every large enough m and algorithm $\mathcal{A}$ that optimizes margins there exists a distribution $\mathcal{D}$ for which

$$\Pr_{(x,y)\sim\mathcal{D}}[yf_{\mathcal{A}}(x) \leq 0] \geq \Pr_{(x,y)\sim S}[yf_{\mathcal{A}}(x) \leq \theta]$$

$$+O\left(\frac{\ln|\mathcal{H}|}{\theta^2} + \sqrt{\frac{\ln|\mathcal{H}|}{\theta^2}\Pr_{(x,y)\sim S}[yf_{\mathcal{A}}(x) \leq \theta]}\right)$$

Margin-Based *Lower* Bounds

- Formally, $\forall N, \theta, \tau$ There exist a set $\mathcal{X}$ and a hypothesis set ... uch that for every ... $\mathcal{A}$ that ... s a distribution $\mathcal{D}$ for which

Where $\theta \in \left(\frac{1}{N}, \frac{1}{40}\right)$ and $\tau \in \left[0, \frac{49}{100}\right]$ are not too large.

$$\Pr_{(x,y)\sim\mathcal{D}}[yf_{\mathcal{A}}(x) \leq 0] \geq \Pr_{(x,y)\sim S}[yf_{\mathcal{A}}(x) \leq \theta]$$

$$+ O\left(\frac{\ln|\mathcal{H}|}{\theta^2} + \sqrt{\frac{\ln|\mathcal{H}|}{\theta^2} \Pr_{(x,y)\sim S}[yf_{\mathcal{A}}(x) \leq \theta]}\right)$$

Margin-Based *Lower* Bounds

- Formally, $\forall N, \theta, \tau$ There exist a set $\mathcal{X}$ and a hypothesis set $\mathcal{H}$ such that for every large enough ... nd algorithm $\mathcal{A}$ that optimi... distrib...

Small set of weak classifiers, $\ln|\mathcal{H}| = \Theta(\ln N)$

$$\Pr_{(x,y)\sim\mathcal{D}}[y f_{\mathcal{A}}(x) \le 0] \ge \Pr_{(x,y)\sim S}[y f_{\mathcal{A}}(x) \le \theta]$$

$$+ O\left(\frac{\ln|\mathcal{H}|}{\theta^2} + \sqrt{\frac{\ln|\mathcal{H}|}{\theta^2} \Pr_{(x,y)\sim S}[y f_{\mathcal{A}}(x) \le \theta]}\right)$$

Margin-Based *Lower* Bounds

- Formally, $\forall N, \theta, \tau$ There exist a set $\mathcal{X}$ and … at for every … large enough … algorithm $\mathcal{A}$ that optimizes mar… there exists a distribution $\mathcal{D}$ for which

Over $\mathcal{X} \times \{-1,1\}$.

$$\Pr_{(x,y)\sim\mathcal{D}}[y f_{\mathcal{A}}(x) \leq 0] \geq \Pr_{(x,y)\sim S}[y f_{\mathcal{A}}(x) \leq \theta]$$

$$+ O\left(\frac{\ln|\mathcal{H}|}{\theta^2} + \sqrt{\frac{\ln|\mathcal{H}|}{\theta^2} \Pr_{(x,y)\sim S}[y f_{\mathcal{A}}(x) \leq \theta]}\right)$$

Margin-Based *Lower* Bounds

- Formally, $\forall N, \theta, \tau$ There exist a set $\mathcal{X}$ and a hypothesis set $\mathcal{H}$ such that for every large enough m and algorithm $\mathcal{A}$ that [illegible] argins there exists a [illegible] $\mathcal{D}$ for which

The classifier returned by $\mathcal{A}$.

$$\Pr_{(x,y)\sim\mathcal{D}}[y f_{\mathcal{A}}(x) \le 0] \ge \Pr_{(x,y)\sim S}[y f_{\mathcal{A}}(x) \le \theta] + O\left(\frac{\ln|\mathcal{H}|}{\theta^2} + \sqrt{\frac{\ln|\mathcal{H}|}{\theta^2} \Pr_{(x,y)\sim S}[y f_{\mathcal{A}}(x) \le \theta]}\right)$$

Margin-Based *Lower* Bounds

- Formally, $\forall N, \theta, \tau$ There exist a set $\mathcal{X}$ and a hypothesis set $\mathcal{H}$ such that for every large enoug … thm $\mathcal{A}$ that optimizes m … ists a distribution $\mathcal{D}$ for whic …

Assuming this is at most τ.

$$\Pr_{(x,y)\sim\mathcal{D}}[yf_{\mathcal{A}}(x) \leq 0] \geq \Pr_{(x,y)\sim S}[yf_{\mathcal{A}}(x) \leq \theta]$$

$$+O\left(\frac{\ln|\mathcal{H}|}{\theta^2} + \sqrt{\frac{\ln|\mathcal{H}|}{\theta^2} \Pr_{(x,y)\sim S}[yf_{\mathcal{A}}(x) \leq \theta]}\right)$$

Margin-Based *Lower* Bounds

- Formally, $\forall N, \theta, \tau$ There exist a set $\mathcal{X}$ and a hypothesis set $\mathcal{H}$ such that for every large enoug … thm $\mathcal{A}$ that optimizes m … ists a distribution $\mathcal{D}$ for whic …

Assuming this is at most τ.

$$\Pr_{(x,y)\sim\mathcal{D}}[yf_{\mathcal{A}}(x) \le 0] \ge \Pr_{(x,y)\sim S}[yf_{\mathcal{A}}(x) \le \theta]$$

$$+O\left(\frac{\ln|\mathcal{H}|}{\theta^2} + \sqrt{\frac{\ln|\mathcal{H}|}{\theta^2}\Pr_{(x,y)\sim S}[yf_{\mathcal{A}}(x) \le \theta]}\right)$$

Summary

- We show margin-based generalization lower bounds which almost match the best known upper bounds.
- These bounds essentially complete the theory of generalization bounds based ob margins alone.
- Open Question : Are there parameters other than margin that can be used to better explain the practical properties of voting classifiers?