Recap from last time: Boosting

• General method for improving the accuracy of any given learning algorithm.

• Works by creating a series of challenge datasets s.t. even modest performance on these can be used to produce an overall high-accuracy predictor.

• **Adaboost** one of the top 10 ML algorithms.
  • Works well in practice.
  • Backed up by solid foundations.
Adaboost (Adaptive Boosting)

Input: $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$; \hspace{1cm} $x_i \in X$, $y_i \in Y = \{-1, 1\}$

weak learning algo $A$ (e.g., Naïve Bayes, decision stumps)

- For $t=1, 2, \ldots, T$
  - Construct $D_t$ on $\{x_1, \ldots, x_m\}$
  - Run $A$ on $D_t$ producing $h_t: X \rightarrow \{-1, 1\}$

Output $H_{\text{final}}(x) = \text{sign}(\sum_{t=1}^T \alpha_t h_t(x))$

- $D_1$ uniform on $\{x_1, \ldots, x_m\}$ [i.e., $D_1(i) = \frac{1}{m}$]
- Given $D_t$ and $h_t$ set
  $$D_{t+1}(i) = \frac{D_t(i)}{Z_t} e^{-\alpha_t y_i h_t(x_i)}$$

$$\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \epsilon_t}{\epsilon_t} \right) > 0$$

$D_{t+1}$ puts half of weight on examples $x_i$ where $h_t$ is incorrect & half on examples where $h_t$ is correct
Nice Features of Adaboost

• **Very general:** a meta-procedure, it can use **any** weak learning algorithm!!! (e.g., Naïve Bayes, decision stumps)

• **Very fast** (single pass through data each round) & **simple to code,** no parameters to tune.

• **Grounded in rich theory.**
Analyzing Training Error

**Theorem** $\epsilon_t = 1/2 - \gamma_t$ (error of $h_t$ over $D_t$)

$$err_S(H_{final}) \leq \exp\left[-2 \sum_t \gamma_t^2\right]$$

So, if $\forall t, \gamma_t \geq \gamma > 0$, then $err_S(H_{final}) \leq \exp[-2 \gamma^2 T]$.

The training error drops exponentially in $T$!!!

To get $err_S(H_{final}) \leq \epsilon$, need only $T = O \left(\frac{1}{\gamma^2 \log \left(\frac{1}{\epsilon}\right)}\right)$ rounds.

**Adaboost is adaptive**

- Does not need to know $\gamma$ or $T$ a priori
- Can exploit $\gamma_t \gg \gamma$
How about generalization guarantees?

**Original analysis [Freund&Schapire’97]**

- $H$ space of weak hypotheses; $d=\text{VCdim}(H)$

  $H_{\text{final}}$ is a weighted vote, so the hypothesis class is:

  $G=\{\text{all fns of the form } \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x)) \}$

**Theorem [Freund&Schapire’97]**

\[
\forall g \in G, \text{err}(g) \leq \text{err}_S(g) + \tilde{O}\left(\sqrt{\frac{Td}{m}}\right)
\]

$T=\#$ of rounds

**Key reason:** $\text{VCdim}(G) = \tilde{O}(dT)$ plus typical VC bounds.
Generalization Guarantees

Theorem [Freund&Schapire’97]

\[
\forall g \in G, \text{err}(g) \leq \text{err}_S(g) + \tilde{O} \left( \sqrt{\frac{Td}{m}} \right)
\]

where \(d=\text{VCdim}(H)\)}
Generalization Guarantees

• Experiments showed that the test error of the generated classifier usually does not increase as its size becomes very large.

• Experiments showed that continuing to add new weak learners after correct classification of the training set had been achieved could further improve test set performance!!!
Generalization Guarantees

• Experiments showed that the test error of the generated classifier usually does not increase as its size becomes very large.

• Experiments showed that continuing to add new weak learners after correct classification of the training set had been achieved could further improve test set performance!!

• These results seem to contradict FS’97 bound and Occam’s razor (in order to achieve good test error the classifier should be as simple as)

\[
\forall g \in G, err(g) \leq err_S(g) + \tilde{O}\left(\sqrt{\frac{Td}{m}}\right)
\]
How can we explain the experiments?

R. Schapire, Y. Freund, P. Bartlett, W. S. Lee. present in "Boosting the margin: A new explanation for the effectiveness of voting methods" a nice theoretical explanation.

Key Idea:
Training error does not tell the whole story.
We need also to consider the classification confidence!!
Boosting didn’t seem to overfit...(!)

...because it turned out to be increasing the margin of the classifier

Error Curve, Margin Distr. Graph - Plots from [SFBL98]
Classification Margin

- $H$ space of weak hypotheses. The **convex hull** of $H$:

$$co(H) = \{ f = \sum_{t=1}^{T} \alpha_t h_t, \alpha_t \geq 0, \sum_{t=1}^{T} \alpha_t = 1, h_t \in H \}$$

- Let $f \in co(H), f = \sum_{t=1}^{T} \alpha_t h_t, \alpha_t \geq 0, \sum_{t=1}^{T} \alpha_t = 1$.

The majority vote rule $H_f$ given by $f$ (given by $H_f = \text{sign}(f(x))$) predicts wrongly on example $(x, y)$ iff $y f(x) \leq 0$.

**Definition:** **margin** of $H_f$ (or of $f$) on example $(x, y)$ to be $y f(x)$.

$$y f(x) = y \sum_{t=1}^{T} [\alpha_t h_t(x)] = \sum_{t=1}^{T} [y \alpha_t h_t(x)] = \sum_{t:y=h_t(x)} \alpha_t - \sum_{t:y\neq h_t(x)} \alpha_t$$

The margin is positive iff $y = H_f(x)$.

See $|y f(x)| = |f(x)|$ as the strength or the confidence of the vote.

-1 \hspace{10cm} \text{Low confidence} \hspace{10cm} 1

High confidence, incorrect \hspace{20cm} High confidence, correct
Boosting and Margins

Theorem: $\text{VCdim}(H) = d$, then with prob. $\geq 1 - \delta$, $\forall f \in \text{co}(H)$, $\forall \theta > 0$, $\Pr_D[yf(x) \leq 0] \leq \Pr_S[yf(x) \leq \theta] + O\left(\frac{1}{\sqrt{m}} \sqrt{\frac{d \ln^2 m}{\theta^2} + \ln \frac{1}{\delta}}\right)$

Note: bound does not depend on $T$ (the # of rounds of boosting), depends only on the complexity of the weak hyp space and the margin!
Boosting and Margins

Theorem: $\text{VCdim}(H) = d$, then with prob. $\geq 1 - \delta$, $\forall f \in \text{co}(H)$, $\forall \theta > 0$,

$$\Pr_D[yf(x) \leq 0] \leq \Pr_S[yf(x) \leq \theta] + O\left(\frac{1}{\sqrt{m}} \sqrt{\frac{d \ln^2 m}{\theta^2}} + \ln \frac{1}{\delta}\right)$$

• If all training examples have large margins, then we can approximate the final classifier by a much smaller classifier.

• Can use this to prove that better margin $\Rightarrow$ smaller test error, regardless of the number of weak classifiers.

• Can also prove that boosting tends to increase the margin of training examples by concentrating on those of smallest margin.

• Although final classifier is getting larger, margins are likely to be increasing, so the final classifier is actually getting closer to a simpler classifier, driving down test error.
Boosting and Margins

Theorem: \( \text{VCdim}(H) = d \), then with prob. \( \geq 1 - \delta \), \( \forall f \in \text{co}(H), \forall \theta > 0 \),

\[
\Pr[D_y f(x) \leq 0] \leq \Pr[S_y f(x) \leq \theta] + O\left(\frac{1}{\sqrt{m}} \sqrt{\frac{d \ln^2 m}{\theta^2} + \ln \frac{1}{\delta}}\right)
\]

Note: bound does not depend on \( T \) (the # of rounds of boosting), depends only on the complex. of the weak hyp space and the margin!
Boosting, Adaboost Summary

• Shift in mindset: goal is now just to find classifiers a bit better than random guessing.

• Backed up by solid foundations.

• Adaboost work and its variations well in practice with many kinds of data (one of the top 10 ML algos).

• More about classic applications in Recitation.

• Relevant for big data age: quickly focuses on “core difficulties”, so well-suited to distributed settings, where data must be communicated efficiently [Balcan-Blum-Fine-Mansour COLT'12].