Boosting:
A weighted crowd of narrowminded experts

Aaron & Dylan
Boosting Hypothesis  

(Kearne, Valiant; 1988-89)

We can make a strong classifier (arbitrarily well at classification) from a collection of weak classifiers (somewhat better than random guess).

Weak Classifiers

- Classifier which may be only slightly correlated with true classification (accuracy $> 50\%$)
- Examples: Naïve Bayes, logistic regression, decision stumps
- Single Level Decision Tree
- Focus on a single feature dimension
- Create a decision boundary along that dimension
Advantages of Boosting:

- Easy and fast to train weak classifiers
- Simple models don’t usually overfit
- Weak classifiers can not solve hard problems
Boosting: The Basic Idea

Think of the weak classifiers as a crowd of experts where each is most familiar with some portion of the dataset.
AdaBoost: Boosting for Binary Classification

Suppose dataset: \((x_1, y_1), \ldots, (x_N, y_N)\)

where \(x_i \in \mathbb{R}^n, y_i \in Y = \{-1, 1\}\)

Let \(D_i(i) = \text{weight of point } x_i\)

Goal: Build classifier \(H(x) = \text{sign}(\alpha_1 h_1(x)+, \ldots, +\alpha_T h_T(x))\)

where \(h_1(x), \ldots, h_T(x)\) are binary classifiers, built on distributions \(D_1, \ldots, D_T\) respectively.

Issue: How to find the best \(\alpha\)’s and \(D\)’s.

Answer: Iteratively minimize exponential loss:

If \(F(x) = \alpha_1 h_1(x)+, \ldots, +\alpha_T h_T(x)\), then

\[
L = \frac{1}{N} \sum_{i=1}^{T} \exp(-y_i F(x_i))
\]
AdaBoost with Decision Stumps as Weak Classifiers  
(Shapire, Freund. 1999)

Round One:

Build $h_1$ on distribution $D_1$

Then calculate:

$$\epsilon_1 = \Pr_{i \sim D_1}(h_1(x) \neq y_i).$$

(sum of misclassified point weights)

Next calculate $\alpha_1$.

Then calculate $D_2$.

For $t = 1, \ldots, T$

Train weak classifier $h_t : \mathbb{R}^n \to R$
on distribution $D_t$

Pick $\alpha_t$

(weight for $h_t$)

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

Set $D_{t+1}(i) :=$

$$D_t(i) \exp \left( -\alpha_t y_i h_t(x_i) \right) / Z_t$$

$$H(x) := \text{sign} \left( \sum_{t=1}^{T} \alpha_t h_t(x) \right)$$
Round One:

Build $h_1$ on distribution $D_1$

$\epsilon_1 = 3/10$

$\alpha_1 = 0.42$

$D_2(i) = 0.166$ for $x_i$ that were misclassified

$D_2(i) = 0.072$ for $x_i$ that were correctly classified

Round Two:

Build $h_2$ on distribution $D_2$

$\epsilon_2 = 0.216$

$\alpha_2 = 0.65$
\( \varepsilon_2 = 0.216, \quad \alpha_2 = 0.65 \)

For each \( X_i \) where:

- \( h_1 \) was wrong, \( h_2 \) was right:
  \( D_3(i) = 0.11, \quad D_2(i) = 0.166 \)

- \( h_1 \) was right, \( h_2 \) was wrong:
  \( D_3(i) = 0.175, \quad D_2(i) = 0.072 \)

- \( h_1 \) was right, \( h_2 \) was right:
  \( D_3(i) = 0.047 \quad D_2(i) = 0.072 \)

**Round Three:**

Train \( h_3 \) on \( D_3 \)

\( \varepsilon_3 = 0.144, \quad \alpha_3 = 0.91 \)
Strong Classifier

\[ H(x) \]

\[
\begin{pmatrix}
0.42 \\
h_1(x)
\end{pmatrix} +
\begin{pmatrix}
+0.65 \\
h_2(x)
\end{pmatrix} +
\begin{pmatrix}
+0.91 \\
h_3(x)
\end{pmatrix}
\]

\[
\forall i, D_t(i) = \frac{1}{N}
\]

Train weak classifier
\[ h_t : \mathbb{R}^n \rightarrow R \]
on distribution \( D_t \)

Pick \( \alpha_t \)
(weight for \( h_t \))
\[
\alpha_t := \frac{1}{2} \ln \left( \frac{1 - \epsilon_t}{\epsilon_t} \right)
\]

Set \( D_{t+1} \) :=
\[
\frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}
\]

\[
H(x) := \text{sign} \left( \sum_{t=1}^{T} \alpha_t h_t(x) \right)
\]
Boosting Demos

Swirly boosting demo

More Swirly boosting demo

AdaBoost in action
References:


Software:

*Wikipedia list from AdaBoost page*

*Boosting Song*