Boosting: Adaboost

1. Binary classification \((y \in \{+1, -1\})\)
2. We are given a training set \(D = \{(x_1, y_1), \ldots, x_n, y_n\}\)
3. Assume access to a dumb classification algorithm we will call the weak learner that will do barely better than chance (> 50% accuracy). (High bias, low variance method)
4. Boosting answers the question: how can we combine classifiers from this dumb algorithm to get an awesome one!
5. In Bagging we always sampled from \(D\) uniformly at random and fit many high variance models that had low bias.
6. For boosting, in a sequential fashion we will sample from carefully crafted distributions over elements of \(D\) and feed to the weak learner and combine the classifiers received from this weak learner.

**First cut boosting:**

\[\forall i, w_i[I_i] = 1/n.\] \((\text{Initialize uniformly})\)
\[H_0 = 0\]

for \(t = 1\) to \(T\):

- Create sample \(D_t\) by drawing points from \(D\) according to \(w_t\).
- Feed \(D_t\) to the weak learner and obtain classifier \(h_t\).
- Add \(h_t\) to your ensembled classifier \(H_t\).
- Update weights \(w_t\) over points in \(D\).

End

Return Ensembled awesome classifier

\[E_0\]

\[E_0 = \begin{cases} 
+ & \text{if } h_1(x) = \text{positive} \\
- & \text{if } h_1(x) = \text{negative} \\
\end{cases}\]
∀i, w_1[i] = 1/n.  (Initialize uniformly)
\[ H_0 = 0 \]

for t = 1 to T:

Create sample D_t by drawing points from D according to w_t
Feed D_t to the weak learner and obtain classifier h_t
Add h_t to your ensembled classifier H_t

∀i, w_{t+1}[i] = \begin{cases} w_t[i] \times \exp(\alpha_t) & \text{if } h_t(x_i) \neq y_i \\ w_t[i] \times \exp(-\alpha_t) & \text{if } h_t(x_i) = y_i \end{cases}

= w_t[i] \times \exp(-\alpha_t y_i h_t(x_i))

End

Return Ensembled awesome classifier

AddBoost:

∀i, w_1[i] = 1/n.  (Initialize uniformly)
\[ H_0 = 0 \]

for t = 1 to T:

Create sample D_t by drawing points from D according to w_t
Feed D_t to the weak learner and obtain classifier h_t

\[ \varepsilon_t = \sum_{i=1}^{n} w_t[i] \delta(h_t(x_i) \neq y_i) \]  (evaluate hypothesis)
\[ \lambda_t = \frac{1}{2} \ln \left( \frac{1 - \varepsilon_t}{\varepsilon_t} \right) \]  (Better hypothesis the more we like for the ensemble)
\[ H_t = H_{t-1} + \alpha_t h_t \]

∀i, w_{t+1}[i] = \begin{cases} w_t[i] \times \exp(\alpha_t y_i h_t(x_i)) & \text{if } h_t(x_i) = y_i \\ w_t[i] \times \exp(-\alpha_t y_i h_t(x_i)) & \text{if } h_t(x_i) \neq y_i \end{cases}

End

Return classifier \( h_{boost}(x) = \text{sign}(H_T(x)) \)
\[ Z_t = \sum_{i=1}^{n} w_t[i] \exp(-y_i \alpha_t h_t(x_i)) \]
(WLH)

Weak Learning Hypothesis: For any weights over points in $D$ the weak learning algorithm can produce a hypothesis whose weighted classification error for those points is better than $1/2 - \varepsilon$. 

Boosting Theorem: If weak learning hypothesis holds with margin then Adaboost will find classifier with $O$ training error on $D$ in

$$T \leq O\left( \frac{\log n}{\varepsilon^2} \right)$$ iterations

$$\forall \varepsilon, \varepsilon_t = \frac{1}{2} - \delta_t < \frac{1}{2} - \delta$$

Training error Analysis (Boosting theorem proof)

$$\text{Error}_D(h_{\text{boost}}) = \frac{1}{n} \sum_{i=1}^{n} I[h_{\text{boost}}(x_i) y_i < 0] \leq \frac{1}{n} \sum_{i=1}^{n} \exp \left( -h_{\text{boost}}(x_i) y_i \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \exp \left( -\frac{t}{T} \alpha_t h_t(x_i) y_i \right)$$

$$= \frac{1}{n} \prod_{t=1}^{T} \exp \left( -\alpha_t h_t(x_i) y_i \right)$$

$$= \prod_{t=1}^{T} z_t$$

$$z_T = \frac{1}{z_{T-1}} w_1 c_1 \exp \left( -\alpha_t h_t(x_i) y_i \right)$$

$$= \frac{1}{z_{T-1}} w_{T-1} c_{T-1} \exp \left( -\alpha_{T-1} h_{T-1}(x_i) y_i \right)$$

$$\vdots$$

$$z_1 = \frac{1}{z_0} w_1 c_1 \exp \left( -\alpha_1 h_1(x_i) y_i \right)$$

$$= \frac{1}{z_0} w_1 c_1$$

$$z_0 = w_1 c_1$$
In HW 7, you will show:

\[
Z_t = 2 \sqrt{\sum_{t=1}^{T} \varepsilon_t (1 - \varepsilon_t)} \leq \prod_{t=1}^{T} \left( 1 - \varepsilon_t^2 \right) \overset{\downarrow}{\leq} \left( 1 - \frac{1}{T} \right)^{T/2} \overset{\downarrow}{\leq} e^{-T/2} \leq e^{-\frac{1}{2} - \frac{1}{T}}
\]

Error decreasing with \( T \) exponentially if \( T > O\left( \frac{\log n}{\delta^2} \right) \) then

\[
\text{Error}_D(\text{h}^{\text{boost}}) \leq \frac{1}{n} \quad \text{But } \alpha = 1 \text{ loss}
\]

\[
\text{Error}_D(\text{h}^{\text{boost}}) = 0
\]

1. Each weak learner is a very high bias simple classifier and hence has low variance
2. Since we only combine \( \log(n) \) number of these weak learners, the boosted method will not have a very high variance either
3. Boosting can be used with any base classifier and was SOTA off the shelf method for a long time
4. Boosting can be seen as a stage wise (gradient based optimization) of objective in Eq. 1
5. Setting up for other losses and procedure for such stage wise optimization of objective like in Eq. 1, we can obtain other variants of boosting, like gradient Boosted Regression Trees that have been wildly successful as well.
Round 1

\[ h_1 = 0.30 \]

Round 2

\[ h_2 = 0.21 \]

Round 3

\[ h_3 = 0.92 \]

Final Classifier

\[ \text{sign} = \frac{H_{\text{final}} + 0.42 + 0.65 + 0.92}{H_{\text{final}}} \]