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:

$\forall i, w_1[i] = 1/n.$ (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

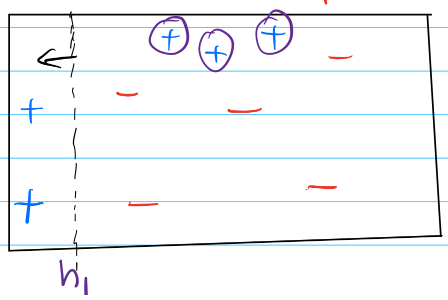
Add h_t to your ensembled classifier H_t

Update weights w_t, over points in D

End

Return Ensembled awesome classifier

Eg



How should we pick W_2?

Second Cut

$\forall i, w_1[i] = 1/n.$ (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

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

$$= \frac{w_t[i] \times \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

End

Return Ensembled awesome classifier

Adaboost:

$\forall i, w_1[i] = 1/n.$ (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

$$\varepsilon_t = \underset{i \sim w_t}{E}\left[\delta\{h_t(x_i) \neq y_i\}\right]$$ (evaluate hypothesis)

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$ (Better hypothesis the more we like for the ensemble)

$$H_t = H_{t-1} + \alpha_t h_t$$

$$\forall i, \quad w_{t+1}[i] = \frac{w_t[i] \times \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

End

Return classifier $h_{Boost}(x) = \mathrm{sign}(H_T(x))$

$$Z_t = \sum_{i=1}^{n} w_t[i] \exp(-y_i \alpha_t h_t(x_i))$$

(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 $1/2 - \gamma$

Boosting Theorem: If weak learning hypothesis holds with margin then Adatoost will find classifier with 0 training error on D in

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

[Freund Schapire]

$$\text{WLH} \Rightarrow \quad \forall t, \quad \varepsilon_t = \frac{1}{2} - \gamma_t < \frac{1}{2} - \gamma$$

Training error Analysis (Boosting thm proof)

$$\begin{aligned}
\text{Error}_D(h_{Boost}) &= \frac{1}{n}\sum_{i=1}^{n} 1\{h_{Boost}(x_i)\, y_i < 0\} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \exp\left(-h_{Boost}(x_i)\, y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\sum_{t=1}^{T} \alpha_t\, h_t(x_i)\, y_i\right) \\
\textcircled{1} \leftarrow \quad &\boxed{= \frac{1}{n}\sum_{i=1}^{n} \prod_{t=1}^{T} \exp\left(-\alpha_t\, h_t(x_i)\, y_i\right)} \\
&= \prod_{t=1}^{T} Z_t
\end{aligned}$$

$1\{a<0\} \leq \exp(-a)$

h_{Boost} defn.

sum in exp = product outside

$$\begin{aligned}
Z_T &= \sum_{i=1}^{n} w_T[i]\, e^{-\alpha_T h_T(x_i) y_i} \\
&= \sum_{i=1}^{n} \frac{w_{T-1}[i]}{Z_{T-1}}\, e^{-\alpha_{T-1} h_{T-1}(x_i) y_i} \times e^{-\alpha_T h_T(x_i) y_i} \\
&= \cdots \\
&= \frac{\sum_{i=1}^{n} w_1[i] \prod_{t=1}^{T} e^{-\alpha_t h_t(x_i) y_i}}{Z_1 \cdots Z_{T-1}}
\end{aligned}$$

In HW 7
You will show
$Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$

$$= \prod_{t=1}^{T} 2\sqrt{\epsilon_t(1-\epsilon_t)} \qquad \epsilon_t = \frac{1}{2} - \gamma_t$$

$$= \prod_{t=1}^{T} 2\sqrt{\frac{1}{4} - \gamma_t^2}$$

$$= \prod_{t=1}^{T} \sqrt{1 - \frac{\gamma_t^2}{4}}$$

$$\leq \left(1 - \frac{\gamma^2}{4}\right)^{T/2} \qquad \gamma_t \geq \gamma$$

$\Downarrow$

Error decreasing with T exponentially

if $T > O\left(\frac{\log n}{\gamma^2}\right)$ then

$\text{Error}_D(h_{Boost}) < \frac{1}{n}$

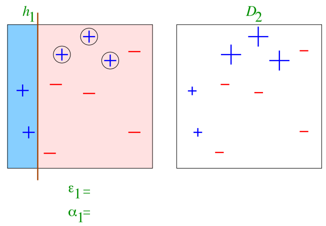
But 0-1 loss so if err $< \frac{1}{n}$

$\Downarrow$

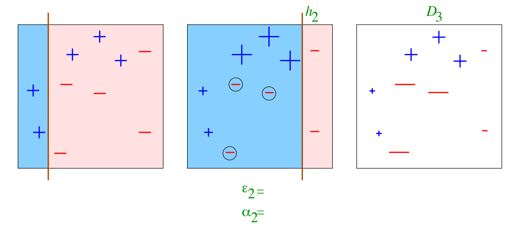
$\text{Error}_D(h_{Boost}) = 0$

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 wont 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 Reregssion Trees that have been wildly successful as well.

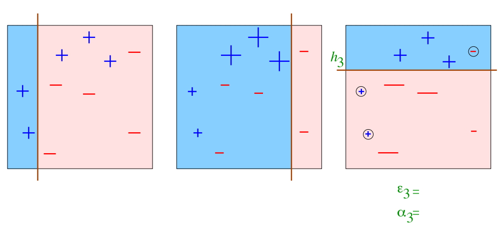
Round 1



Round 2



Round 3



Final Classifier