Lecture 23 - 08-06-2020

Bagging

$$h_1, ..., h_t \qquad \hat{\ell}(f) \leq e^{-2T\gamma^2} \qquad \gamma_t > \gamma > 0$$

Under the assumption that $\{h_t(x_z) \neq y_z\} \qquad \gamma_t = \frac{1}{2} - \hat{\ell}_s(h_t)$ are independent

$$f = sgn(\sum_{i=1}^{T} h_i) \qquad Bagging$$

1.1 Boosting

$$f = sgn(\sum_{i=1}^{T} w_i h_i) \qquad Boosting$$

The hard thing here is how to compute the weights.
$h_1, ..., h_t \qquad X \rightarrow \{-1, +1\}$

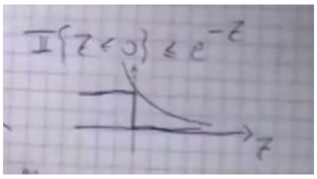

Figure 1.1:

$$\hat{\ell}(f) \sum_{t=1}^{m} I\{y_t g(x_t) \leq 0\} \leq \frac{1}{m} \sum_{t=1}^{m} e^{-g(x_t) y_t} =$$

$$g = \sum_{i=1}^{T} w_i h_i and\ we\ substitute\ g \qquad and \qquad f = sgn(g)$$

$$= \frac{1}{m} \sum_{t=1}^{m} e^{-y_t \sum_{t=1}^{T} w_i h_i(x_i)} \qquad L_i(t) = h_i(x_t) y_t \in \{-1, +1\} i = 1, ..., T$$

$L_i(z)$ where Z uniform over $\{1, ..., m\}$

$$\hat{(\ell)}(f) \leq \frac{1}{m} \sum_{t=1}^{m} e^{-\sum_{t=1}^{T} w_i L_i(t)} = \mathbb{E}\left[e^{-\sum_{t=1}^{T} w_i L_i(t)}\right]$$

$$\mathbb{E}\left[\prod_{t=1}^{T} e^{-w_i L_i}\right] =^{?} \prod_{t=1}^{T} \mathbb{E}\left[e^{-w_i Li}\right]$$

Ok if Li are independent
$E\,[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$
X, Y are independent
$\mathbb{E}_i$ is a probability P_i and P_i is sum $\{1, ...m\}$

$$\hat{\ell}(f) \;\le\; \prod_{i=1}^{T} \mathbb{E}_i\left[e^{-w_i L_i}\right] \;=\; \prod_{i=1}^{T} \left(e^{w_i} P_i(L_i = 1) + e^{w_i} P_i\,(L_i = 1)\right) \;=\; \prod_{i=1}^{T} \left(e^{-w_i}(1-\epsilon_i) + e^{-w_i}\varepsilon_i\right)$$

$L_i(z) \qquad z \sim P_i$

$$\varepsilon_i = P_i(L_i = -1) = \sum_{t=1}^{m} I\{y_t h_i(x_t) \le 0\} P_i(t) \qquad \textbf{weighted training error of } h_i$$

$$F(w) = e^{-w}(1-\varepsilon) + e^{w}\varepsilon \qquad F'(w) = 0 \Leftrightarrow w = \frac{1}{2}\ln\frac{1-\varepsilon}{\varepsilon} \qquad 0 < \varepsilon < 1$$

$$P_i(t) > 0 \quad \forall i, t \qquad \varepsilon_i = \frac{1}{2} \Rightarrow w_i = 0$$

$$\varepsilon_i > \frac{1}{2} \Rightarrow w_i < 0 \qquad\qquad \varepsilon_i < \frac{1}{2} \Rightarrow w_i > 0$$

$$\hat{\ell}(f) \;\le\; \prod_{i=1}^{T} \sqrt{4\,\varepsilon_i(1-\varepsilon_i)}$$

$$\gamma_i = \frac{1}{2} - \varepsilon_i \qquad \textit{edge over random guessing } 0 < \varepsilon_i < 1$$

$$1 + x \le e^x \quad \forall x \in \mathbb{R} \quad \hat{\ell}(f) \le \prod_{i=1}^{T} \sqrt{4\,\varepsilon_i(1-\varepsilon_i)} \;=\; \prod_{i=1}^{T} \sqrt{1 - 4\gamma^2} =$$

$$= \prod_{i=1}^{T} 4\left(\frac{1}{2} - \gamma_i\right)\left(\frac{1}{2} + \gamma_i\right) \;= \prod_{i=1}^{T} e^{-2\gamma_i^2} = e^{-2\sum_{i=1}^{T}\gamma_i^2} \le e^{-2T\gamma^2}$$

$If \quad |\gamma_i| > \gamma > 0 \quad i = 1, ..., T$

$$\hat{\ell}_s(f) = 0 \;\Leftrightarrow\; e^{-\varepsilon T \gamma^2} < \frac{1}{m} \;\Leftrightarrow\; T > \frac{\ln m}{2\gamma^2}$$

$$E\left[\prod_i e^{-w_i L_i}\right] \;=\; \prod_i E\left[e^{-w_i L_i}\right]$$

$P_i, ..., P_T \quad P_1(t) = \frac{1}{m} \qquad t = 1, ...m$

$$P_{i+1}(t) = \frac{P_i(t)e^{-w_iL_i(t)}}{E_i\left[e^{-w_iL_i}\right]} \qquad \sum_t P_{i+1}(t) = \frac{1}{E_i[e^{-w_iL_i}]}\sum_t P_i(t)e^{\cdots}$$

MANCAaaa

$$e^{-w_iL_i(t)} = E_i[e^{-w_iL_i}]\frac{P_{i+1}}{P_i(t)}$$

$$E[\prod_{i=1}^{T} e^{-w_iL_i}] = \frac{1}{m}\sum_{t=1}^{m}\prod_{i=1}^{T} e^{-w_iL_i(t)} = \frac{1}{m}\sum_t l\left(\prod_i E[e^{-w_iL_i}]\frac{P_{t+1}(t)}{P_i(t)}\right) =$$

$$\frac{1}{m}\sum_t\left(\prod_i E_i[e^{-w_iL_i}\right)\frac{P_{t+1}(t)}{P_1(t)} = \left(\prod_i E_i[e^{-w_iL_i}]\right)\color{red}{\frac{1}{m}\sum_t\frac{P_{t+1}(t)}{ym}}$$

where **red** cancel out since $= 1$

1.2 Adaboost

It is a meta learning algorithm.

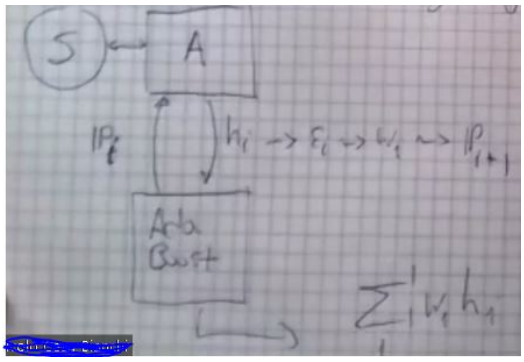

Figure 1.2:

Initialize $P_i(t) = \frac{1}{m} \quad t = 1, ..., m$
For $i = 1, ..., T$
1) Feed A with S wrighted by P_i and get h_1
2) $w_i = \frac{1}{2} \ln \frac{\varepsilon_i}{1-\varepsilon_i}$
3) Compute P_{i+1}
Output $\sum_i w_i h_i$

What should A do?
1) A should pay attention to P_i
2) More precisely A should output h_i s.t. $|\gamma_i|$ is as big as possible
where $|\gamma| \to \frac{1}{2}\varepsilon_i$

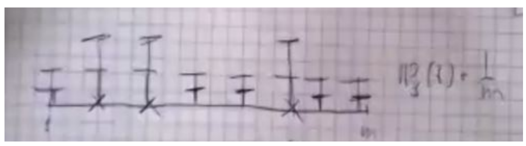

Figure 1.3:

$$P_{i+1} = \frac{P_i(t)e^{-w_i L_i(t)}}{E_i[\quad]}$$
$$L_i(t) = 1 \;\Leftrightarrow h_t(x_t) = y_t \qquad w_i > 0$$

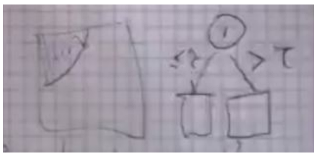

Figure 1.4:

Typically h_i (classifiers) are simple
Decision stamps:

$$h(x) = \pm\, sgn(x_i - \tau)$$

i if is feature index, $\tau \in \mathbb{R}$
At the end boosting is gonna look like this

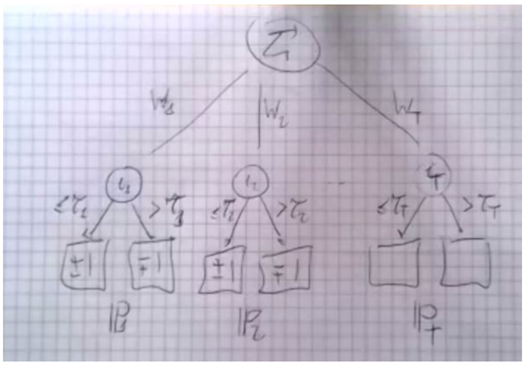

Figure 1.5: