Howework 7

Lingyu Zhou

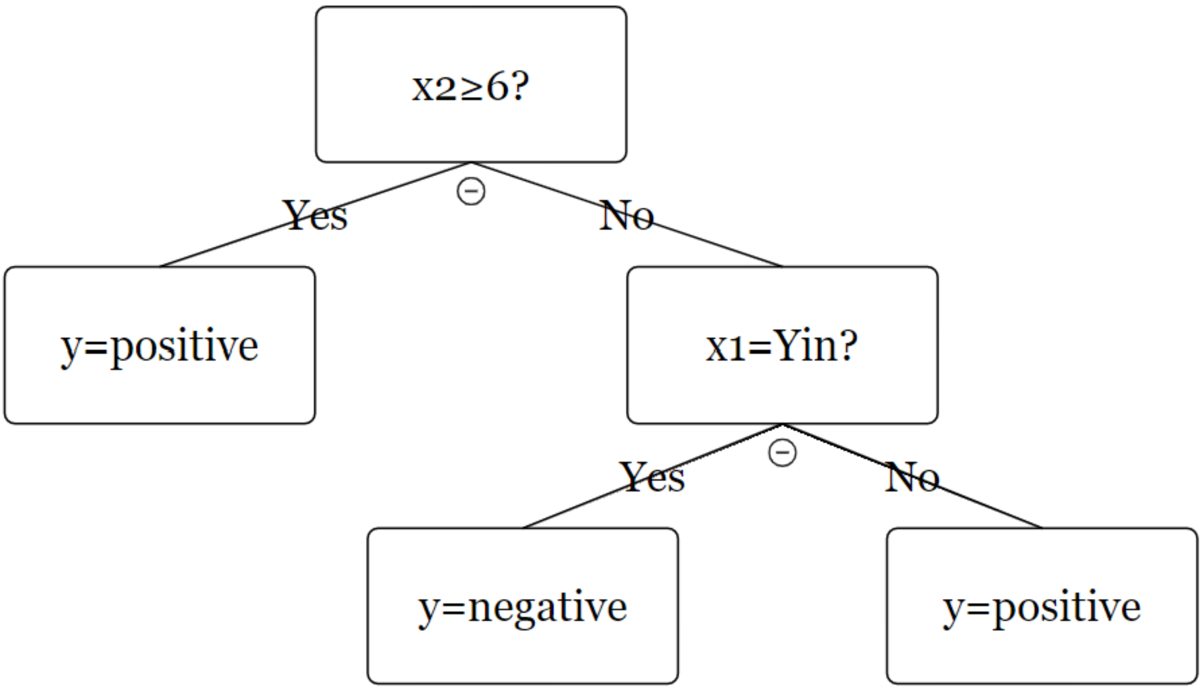
2024.4.28

Problem 1.

a.

Split on x1:
For x1=Yin, there are 2 outcomes, + and - each 2. And for x1=Yang, there is only 1 outcome, + (entropy is 0). Therefore, $S_1 = 4/6(1/2\log(2) + 1/2\log(2)) + 2/6 \times 0 \approx 0.4621$.
Split on x2:
For x2≥ 6, there are 3 outcomes, all + (entropy is 0). For x2< 6, there are 2 - and 1 +. Therefore $S_2 = 3/6(2/3\log(3/2) + 1/3\log(3)) + 3/6 \times 0 \approx 0.3183$
$S_2 < S_1$ hence choose x2.

b.



c.

Split on x1:
For x1=Yin, there are 3 - and 1 +. And for x1=Yang, there are only 2 +. Therefore, $S_1' = 4/6(1/4\log(4) + 3/4\log(4/3)) + 2/6 \times 0 \approx 0.3749$.

Split on x2:
For x2≥ 6, there are 2 + and 1 -. For x2< 6, there are 2 - and 1 +. Therefore $S_2' = 3/6(2/3\log(3/2) + 1/3\log(3)) + 3/6(2/3\log(3/2) + 1/3\log(3)) \approx 0.6365$
$S_2' > S_1'$ hence choose `x1` instead.

Problem 2.

1.

For convenience, ignore all subscript t and replace subscript $t+1$ by $'$. Need to show: $\sum_i w_i \cdot e^{-\alpha' h'(x_i) y_i} = 2\sqrt{\epsilon'(1-\epsilon')}$

$$
\begin{aligned}
\sum_i w_i \cdot e^{-\alpha' h'(x_i) y_i} &= \sum_{i:h'(x_i)=y_i} w_i \cdot e^{-\alpha'} + \sum_{i:h'(x_i)\neq y_i} w_i \cdot e^{\alpha'} \\
&= e^{-\alpha'} \sum_{i:h'(x_i)=y_i} w_i + e^{\alpha'} + \sum_{i:h'(x_i)\neq y_i} w_i \\
&= e^{\frac{1}{2}\ln\left(\frac{1-\epsilon'}{\epsilon'}\right)} \sum_{i:h'(x_i)\neq y_i} w_i + e^{-\frac{1}{2}\ln\left(\frac{1-\epsilon'}{\epsilon'}\right)} \sum_{i:h'(x_i)=y_i} w_i \\
&= e^{\ln\left(\frac{1-\epsilon'}{\epsilon'}^{\frac{1}{2}}\right)} \sum_{i:h'(x_i)\neq y_i} w_i + e^{\ln\left(\frac{\epsilon'}{1-\epsilon'}^{\frac{1}{2}}\right)} \sum_{i:h'(x_i)=y_i} w_i \\
&= \left(\frac{1-\epsilon'}{\epsilon'}\right)^{\frac{1}{2}} \sum_{i:h'(x_i)\neq y_i} w_i + \left(\frac{\epsilon'}{1-\epsilon'}\right)^{\frac{1}{2}} \sum_{i:h'(x_i)=y_i} w_i \\
&= \left(\frac{1-\epsilon'}{\epsilon'}\right)^{\frac{1}{2}} \epsilon' + \left(\frac{\epsilon'}{1-\epsilon'}\right)^{\frac{1}{2}} (1-\epsilon') \\
&= \epsilon'^{\frac{1}{2}}(1-\epsilon')^{\frac{1}{2}} + \epsilon'^{\frac{1}{2}}(1-\epsilon')^{\frac{1}{2}} \\
&= 2\left[\epsilon'(1-\epsilon')\right]^{\frac{1}{2}} \\
&= 2\sqrt{\epsilon'(1-\epsilon')}
\end{aligned}
$$

$\square$

2.

For each iteration in the boosting algorithm, the weight of each data point will be updated based on weak model's performace in that epoch.
More specifically, for data points that are correctly misclassified by the weak models, they will be updated to lower weights so that the algorithm can somewhat ignore them and for data points that are missclassfied by the weak models, they will be updated to higher weights so that we can more focus on those missclassfied points to hence gradually improving algorithm's performance.
Intuitively, the weights are adjusted in such a way that emphasizes the importance of difficult-to-classify points, allowing the algorithm to effectively learn from its mistakes and iteratively improve its overall performance.

Problem 3.

$$\underset{D}{E}\underset{D_1 \dots D_m}{E}\underset{(x,y)}{E}\left[\left[(\hat{h}(x)-y)^2\right]\right] = E \cdots \underset{(x,y)}{E}\left[\underbrace{(y(x)-\bar{y}(x))^2}_{\text{Noise}} + \underbrace{(\bar{y}(x)-\bar{h}(x))^2}_{\text{Bias}} + \underbrace{(\bar{h}(x)-\hat{h}(x))^2}_{\text{Variance}}\right]$$

Since Noise and buis term are both not related to $\underset{D_1 \dots D_m}{E}$, $\bar{h}(x) \in \mathbb{R}$, we only need to consider variance term, that is:

$$\underset{D}{E}\underset{D_1 \dots D_m}{E}\underset{(x,y)}{E}(\bar{h}(x)-\hat{h}(x))^2 = E \cdots \underset{(x,y)}{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m} h_{D_i} - \bar{h}(x)\right)^2\right] \quad \cdots (1)$$

Write $\underset{D}{E}\underset{D_1 \dots D_m}{E}\underset{(x,y)}{E}$ as $\mathbb{E}$ now for convenience

$$
\begin{aligned}
(1) &= \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{n} h_{D_i}\right)^2\right] - \mathbb{E}\left[2\frac{1}{m}\sum_{i=1}^{m} h_{D_i}\bar{h}(x)\right] + \mathbb{E}\left[(\bar{h}(x))^2\right] \\
&= \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m} h_{D_i}\right)^2\right] - 2\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} h_{D_i}\bar{h}(x)\right] + \mathbb{E}\left[\bar{h}(x)^2\right] \\
&= \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m} h_{D_i}\right)^2\right] - \frac{2}{m}\sum_{i=1}^{m}\mathbb{E}[h_{D_i}]\bar{h}(x) + \bar{h}(x)^2, \text{ since } \bar{h}(x) \in \mathbb{R} \\
&= \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m} h_{D_i}\right)^2\right] - \frac{2}{\cancel{m}}\cancel{m}\,\bar{h}(x)\cdot\bar{h}(x) + \bar{h}(x)^2 \\
&= \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m} h_{D_i}\right)^2\right] - \bar{h}(x)^2 \quad \cdots (2)
\end{aligned}
$$

Since $\bar{h}(x)^2 = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} D_i\right]^2$, $(2) = \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m} h_{D_i}\right)^2 - \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} D_i\right]^2\right] = Var\left[\frac{1}{m}\sum_{i=1}^{m} h_{D_i}(x)\right]$

So $(1) = Var\left[\frac{1}{m}\sum_{i=1}^{m} h_{D_i}(x)\right]$

using law of total variance $\rightarrow$ $= Var\left[\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} h_{D_i}(x) \mid \{D, X\}\right]\right] + \mathbb{E}\left[Var\left[\frac{1}{m}\sum_{i=1}^{m} h_{D_i}(x) \mid \{D, X\}\right]\right]$

given $h_{D_i} \perp h_{D_{i+1}}$ $\rightarrow$ $= Var\left[\mathbb{E}\left[h_D(x) \mid \{D, X\}\right]\right] + \mathbb{E}\left[\frac{1}{m^2}Var\left[\sum_{i=1}^{m} h_{D_i}(x) \mid \{D, X\}\right]\right]$

$$= \underbrace{Var\left[\mathbb{E}\left[h_D(x) \mid \{D, X\}\right]\right]}_{\text{doesn't change on increasing } m} + \frac{1}{m^2}\mathbb{E}\left[Var\left[\sum_{i=1}^{m} h_{D_i}(x) \mid \{D, X\}\right]\right]$$

$$= Var\left[\mathbb{E}(h_D(x) \mid \{D, X\})\right] + \mathbb{E}\left[\frac{1}{m}Var\left[h_D(X) \mid \{D, X\}\right]\right]$$

$$= \underbrace{Var\left[\mathbb{E}\left[h_D(x) \mid \{D, X\}\right]\right]}_{\text{not depends on } m} + \underbrace{\mathbb{E}\left[\frac{1}{m}Var\left[h_D(x) \mid D\right]\right]}_{\text{as } m\uparrow, \text{ it will } \downarrow \text{ since } Var(\cdot) \text{ fixed}}$$

Therefore, we've shown that if m increases variances will not also increase, meaning expected squared error of the ensemble will not increases as m increases.