IE 7374 ST: Machine Learning in Engineering
HW-4

Rohit Bokade
NUID: 001280767

April 22, 2021

1. **Data:**

a_1	a_2	a_3	a_4	C
1	1	0	0	ω_1
0	0	0	0	ω_1
1	0	1	0	ω_1
0	0	1	1	ω_1
1	1	0	0	ω_2
1	1	1	1	ω_2
1	1	1	0	ω_2
0	1	1	1	ω_2

Entropy of y:

$$\begin{aligned}
H(x) &= -\sum_{i}^{n} p(x_i)\log_2 p(x_i) \\
H(C) &= -\sum_{i=1}^{k} P(c_i|D)\log_2 P(c_i|D) \\
&= -[\frac{4}{8}\log_2\frac{4}{8} + \frac{4}{8}\log_2\frac{4}{8}] \\
&= 1 && \text{(Since both classes are equally represented (max entropy))}
\end{aligned}$$

$$\begin{aligned}
H(a_1^{\omega_1}) &= 1 \\
H(a_1^{\omega_2}) &= -[\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}] = 0.81 \\
H(a_1^{\omega_1}, a_1^{\omega_2}) &= \frac{4}{8}(1) + \frac{4}{8}(0.81) = 0.905 \\
\text{Gain}(C, a_1^{\omega_1}, a_1^{\omega_2}) &= H(C) - H(a_1^{\omega_1}, a_1^{\omega_2}) = 1 - 0.905 = 0.095
\end{aligned}$$

$$
\begin{aligned}
H(a_2^{\omega_1}) &= -[\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}] = 0.81 \\
H(a_2^{\omega_2}) &= -[\frac{4}{4}\log_2\frac{4}{4} + 0] = 0 \\
H(a_2^{\omega_1}, a_1^{\omega_2}) &= \frac{4}{8}(0.81) + \frac{4}{8}(0.0) = 0.405 \\
\text{Gain}(C, a_2^{\omega_1}, a_2^{\omega_2}) &= H(C) - H(a_1^{\omega_1}, a_1^{\omega_2}) = 1 - 0.405 = 0.595
\end{aligned}
$$

$$
\begin{aligned}
H(a_3^{\omega_1}) &= -[\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}] = 1 \\
H(a_3^{\omega_2}) &= -[\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}] = 0.81 \\
H(a_3^{\omega_1}, a_1^{\omega_2}) &= \frac{4}{8}(1) + \frac{4}{8}(0.81) = 0.905 \\
\text{Gain}(C, a_3^{\omega_1}, a_3^{\omega_2}) &= H(C) - H(a_1^{\omega_1}, a_1^{\omega_2}) = 1 - 0.905 = 0.095
\end{aligned}
$$

$$
\begin{aligned}
H(a_4^{\omega_1}) &= -[\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}] = 0.81 \\
H(a_4^{\omega_2}) &= -[\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}] = 1 \\
H(a_4^{\omega_1}, a_1^{\omega_2}) &= \frac{4}{8}(0.81) + \frac{4}{8}(1) = 0.905 \\
\text{Gain}(C, a_4^{\omega_1}, a_4^{\omega_2}) &= H(C) - H(a_1^{\omega_1}, a_1^{\omega_2}) = 1 - 0.905 = 0.095
\end{aligned}
$$

The attribute a_2 has the maximum information gain. Hence, we would select a_2 to be the root node.

2. **Only if we assume the classification scheme as shown in Figure 1:**

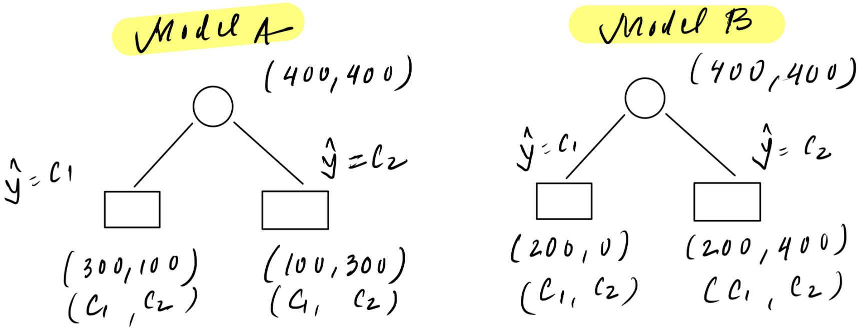

Figure 1: Decision tree classification scheme

$$\delta_{ModelA} = \frac{100 + 100}{400 + 400} = 0.25$$
$$\delta_{ModelB} = \frac{0 + 200}{400 + 400} = 0.25$$

We can show that the misclassification rates are equal.

Cross-entropy:

Let $y_{C_1} = 0$ and $y_{C_2} = 1$

For Model A

$$\begin{aligned}
\mathcal{CE} &= -[\sum_{i=1}^{n} y_i \log p + (1 - y_i) \log(1 - p)] \\
\mathcal{CE}_{C_1} &= -[0 \log(0.75) + (1 - 0) \log(1 - 0.75)] = 1.39 \\
\mathcal{CE}_{C_2} &= -[1 \log(0.25) + (1 - 1) \log(1 - 0.25)] = 1.39 \\
\mathcal{CE} &= 1.39 + 1.39 = 2.78
\end{aligned}$$

For Model B

$$\begin{aligned}
\mathcal{CE}_{C_1} &= -[0 \log(1) + (1 - 0) \log(1 - 1)] = 0 \\
\mathcal{CE}_{C_2} &= -[1 \log(0.33) + (1 - 1) \log(1 - 0.33)] = 1.11 \\
\mathcal{CE} &= 0 + 1.11 = 1.11
\end{aligned}$$

Gini Index:

For Model A

$$\text{Gini(Model A)} = 1 - \left[\left(\frac{400}{800} \right)^2 + \left(\frac{400}{800} \right)^2 \right] = 0.5$$

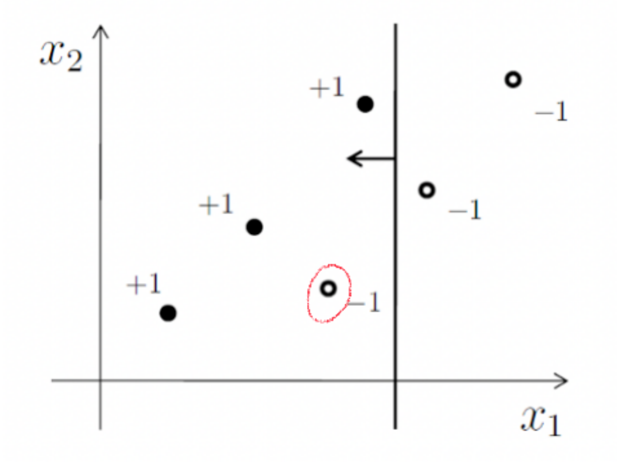

Figure 2: Point -1 (in red circle) is misclassified

For Model B

$$\text{Gini(Model B)} = 1 - \left[\left(\frac{200}{800} \right)^2 + \left(\frac{600}{800} \right)^2 \right] = 0.375$$

Both cross-entropy and gini index are lower for Model B,

3. (a) The point which was misclassified as -1 in the t-th iteration will have the maximum weight in the next iteration as shown in Figure 2.

   (b) We know that the point misclassified in the previous iteration will have the highest probability of being selected in the next iteration. Assuming that the misclassified point was selected the model would try to reduce the error for that point and a possible stump would look like something shown in Figure 3. This also makes sense since now it is misclassifying only one point instead of two, in the previous iteration.

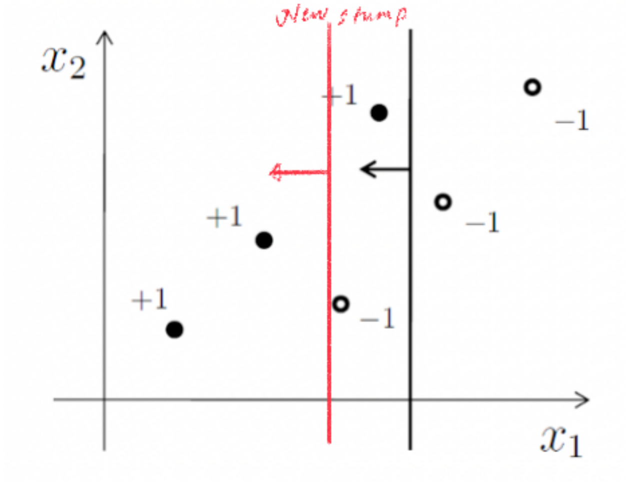

Figure 3: New stump (red line) for the next the boosting iteration

(c) Yes. As seen in the Figure 3, the second stump misclassifies one data point of class $+1$, whereas the first stump misclassified two data points ($+1$, -1). Thus, the relative weight α_2 would be higher.

$$\alpha_2 > \alpha_1$$

4. (a) In boosting, weak learners are stumps which use only one variable to make prediction. The total number of weak learners can be (# of classes) × (# of data points) $= 2m$.

   (b) Yes, AdaBoost can select a weak classifier more than once. In AdaBoost, the previous error is the only criterion for improvement. This can cause the algorithm to revisit a weak learner after it's discarded.

   (c) Mutual Information measures the amount of reduction in uncertainty (in one variable) given the information about the second variable. In this case the $\hat{I}(y; X_j)$ measures the reduction in entropy of y given the information about the variable X_j. Now, AdaBoost would return a ranking of X_j based on how much it is contributing towards the information gain of y *with respect to the other variables* $X_i \forall i \in \{1, \cdots, k\}$ and $i \neq j$. This makes AdaBoost more informative than mutual information since the ranking is subject to the contribution made by other variables.