DANIEL WU
1895-2460-22
ddwu@usc.edu

Written Assignment 1
Summer 2020: CSCI 567

2020-07-08

1 Problem 1

$$E(\boldsymbol{x}, \boldsymbol{x'}) = \sum_{d=1}^{D}(x_d - x'_d)^2 = \sum_{d=1}^{D}({x_d}^2 - 2x_d x'_d + {x'_d}^2) = \sum_{d=1}^{D}{x'_d}^2 - \sum_{d=1}^{D}2x_d x'_d + \sum_{d=1}^{D}{x'_d}^2 = 2 - 2\sum_{d=1}^{D}(x_d x'_d)$$

$$C(\boldsymbol{x}, \boldsymbol{x'}) = 1 - \frac{\sum_{d=1}^{D}(x_d x'_d)}{\|\boldsymbol{x}\|_2 \|\boldsymbol{x'}\|_2} = 1 - \sum_{d=1}^{D}(x_d x'_d)$$

$$\implies 2C(\boldsymbol{x}, \boldsymbol{x'}) = 2 - 2\sum_{d=1}^{D}(x_d x'_d) = E(\boldsymbol{x}, \boldsymbol{x'})$$

$$\therefore \forall \boldsymbol{x_i}, \boldsymbol{x_j}, \boldsymbol{x_k} \in \mathbb{R}^d, E(\boldsymbol{x_i}, \boldsymbol{x_j}) > E(\boldsymbol{x_i}, \boldsymbol{x_k})$$

$$\implies \frac{1}{2}E(\boldsymbol{x_i}, \boldsymbol{x_j}) > \frac{1}{2}E(\boldsymbol{x_i}, \boldsymbol{x_k})$$
$$\implies C(\boldsymbol{x_i}, \boldsymbol{x_j}) > C(\boldsymbol{x_i}, \boldsymbol{x_k})$$

WLOG, it can be easily shown for the opposite inequality.

2 Problem 2

2.1

Yes. Let's denote the dataset $\mathcal{D}$ and assume it spans all $\{0, 1\}^{100}$ space. Create a binary decision tree that grows all 100 features to capture all 2^{100} possible combinations of features. Since $\boldsymbol{x_i} \neq \boldsymbol{x_j}, \ \ \forall \boldsymbol{x_i}, \boldsymbol{x_j} \in \mathcal{D}$, no two examples will ever end up in the same leaf. Thus there will be zero classification error.

2.2

Yes. We can just carry our dataset $\mathcal{D}$. Any test example will overlay exactly on a training example (since we assumed a full dataset), resulting in the same classification as the decision tree.

3 Problem 3

3.1

Let $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$ denote the stumps with root node x_1, x_2 and x_3 respectively. The rootnode will test for each stump if the feature is equal to 0. If true, go left. If false, go right.

Let $\epsilon_{i,\{L(R)\}}$, $H_{i,\{L(R)\}}$ and $G_{i,\{L(R)\}}$ denote the misclassification rate, cross entropy and gini index for $\mathcal{T}_i$ in the left(right) leaf node respectively.

Let $\epsilon(i)$, $H(i)$ and $G(i)$ denote the weighted average of the cost measure for T_i.

$\mathcal{T}_1$:

Misclassification Rate:

$$\begin{aligned}
\epsilon_{1,L} &= \frac{25}{25+25} = \frac{1}{2} \\
\epsilon_{1,R} &= \frac{25}{25+25} = \frac{1}{2} \\
\epsilon(1) &= \frac{50}{100} \cdot \frac{1}{2} + \frac{50}{100} \cdot \frac{1}{2} = \boxed{\frac{1}{2}}
\end{aligned}$$

Entropy:

$$\begin{aligned}
H_{1,L} &= -\frac{25}{25+25} \log\left(\frac{25}{25+25}\right) - \frac{25}{25+25} \log\left(\frac{25}{25+25}\right) \\
&= -\frac{1}{2} \log\left(\frac{1}{2}\right) - \frac{1}{2} \log\left(\frac{1}{2}\right) \\
&= -\log\left(\frac{1}{2}\right) = \log(2) \\
H_{1,R} &= -\frac{25}{25+25} \log\left(\frac{25}{25+25}\right) - \frac{25}{25+25} \log\left(\frac{25}{25+25}\right) \\
&= -\log\left(\frac{1}{2}\right) = \log(2) \\
H(1) &= \frac{50}{100} \log(2) + \frac{50}{100} \log(2) = \boxed{\log(2)}
\end{aligned}$$

Gini Impurity:

$$\begin{aligned}
G_{1,L} &= \frac{25}{25+25}\left(1 - \frac{25}{25+25}\right) + \frac{25}{25+25}\left(1 - \frac{25}{25+25}\right) \\
&= \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \\
G_{1,R} &= \frac{25}{25+25}\left(1 - \frac{25}{25+25}\right) + \frac{25}{25+25}\left(1 - \frac{25}{25+25}\right) = \frac{1}{2} \\
G(1) &= \frac{50}{100} \cdot \frac{1}{2} + \frac{50}{100} \cdot \frac{1}{2} = \boxed{\frac{1}{2}}
\end{aligned}$$

$\mathcal{T}_2$:

Misclassification Rate:

$$\epsilon_{2,L} = \frac{15}{35+15} = \frac{3}{10}$$
$$\epsilon_{2,R} = \frac{15}{35+15} = \frac{3}{10}$$
$$\epsilon(2) = \frac{50}{100} \cdot \frac{3}{10} + \frac{50}{100} \cdot \frac{3}{10} = \boxed{\frac{3}{10}}$$

Entropy:

$$\begin{aligned} H_{2,L} &= -\frac{35}{35+15} \log\left(\frac{35}{35+15}\right) - \frac{15}{35+15} \log\left(\frac{15}{35+15}\right) \\ &= -0.7\log(0.7) - 0.3\log(0.3) \\ H_{2,R} &= -\frac{35}{35+15} \log\left(\frac{35}{35+15}\right) - \frac{15}{35+15} \log\left(\frac{15}{35+15}\right) \\ &= -0.7\log(0.7) - 0.3\log(0.3) \\ H(2) &= \cancel{\left(\frac{50}{100} + \frac{50}{100}\right)} \boxed{\big(-0.7\log(0.7) - 0.3\log(0.3)\big)} \end{aligned}$$

Gini Impurity:

$$\begin{aligned} G_{2,L} &= \frac{35}{35+15}\left(1 - \frac{35}{15+35}\right) + \frac{15}{35+15}\left(1 - \frac{15}{15+35}\right) \\ &= 0.7(0.3) + 0.3(0.7) = 0.42 \\ G_{2,R} &= \frac{35}{35+15}\left(1 - \frac{35}{15+35}\right) + \frac{15}{35+15}\left(1 - \frac{15}{15+35}\right) \\ &= 0.7(0.3) + 0.3(0.7) = 0.42 \\ G(2) &= \cancel{\left(\frac{50}{100} + \frac{50}{100}\right)} \boxed{0.42} \end{aligned}$$

$\mathcal{T}_3$:

Misclassification Rate:

$$\epsilon_{3,L} = \frac{0}{40+0} = 0$$
$$\epsilon_{3,R} = \frac{10}{10+50} = \frac{1}{6}$$
$$\epsilon(3) = \frac{40}{100} \cdot 0 + \frac{60}{100} \cdot \frac{1}{6} = \boxed{\frac{1}{10}}$$

Entropy:

$$
\begin{aligned}
H_{3,L} &= -\frac{0}{0+40}\log\left(\frac{0}{0+40}\right) - \frac{40}{0+40}\log\left(\frac{40}{0+40}\right) = 0 \\
H_{3,R} &= -\frac{50}{10+50}\log\left(\frac{50}{10+50}\right) - \frac{10}{10+50}\log\left(\frac{10}{10+50}\right) \\
&= -\frac{5}{6}\log\left(\frac{5}{6}\right) - \frac{1}{6}\log\left(\frac{1}{6}\right) \\
H(3) &= \frac{40}{100}\cdot 0 + \frac{60}{100}\left[-\frac{5}{6}\log\left(\frac{5}{6}\right) - \frac{1}{6}\log\left(\frac{1}{6}\right)\right] \\
&= \boxed{-\frac{1}{2}\log\left(\frac{5}{6}\right) - \frac{1}{10}\log\left(\frac{1}{6}\right)}
\end{aligned}
$$

Gini Impurity:

$$
\begin{aligned}
G_{3,L} &= \frac{0}{40}\cdot\left(1-\frac{0}{40}\right) + \frac{40}{40}\cdot\left(1-\frac{40}{40}\right) = 0 \\
G_{3,R} &= \frac{10}{10+50}\left(1-\frac{10}{10+50}\right) + \frac{50}{10+50}\left(1-\frac{50}{10+50}\right) \\
&= \frac{1}{6}\cdot\frac{5}{6} + \frac{1}{6}\cdot\frac{5}{6} = \frac{5}{18} \\
G(3) &= \frac{40}{100}\cdot 0 + \frac{60}{100}\cdot\frac{5}{18} = \boxed{\frac{1}{6}}
\end{aligned}
$$

3.2

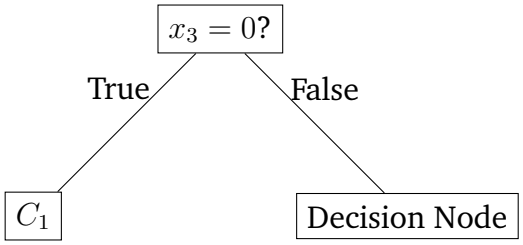
$$G(1) = \frac{1}{2}, \;\; G(2) = 0.42, \;\; G(3) = \frac{1}{6}$$

$$G(3) < G(2) < G(1)$$

Therefore, using gini impurity as our cost measure and the greedy approach, we split on x_3 first.

3.3

We start with the stump and wish to grow a second level.

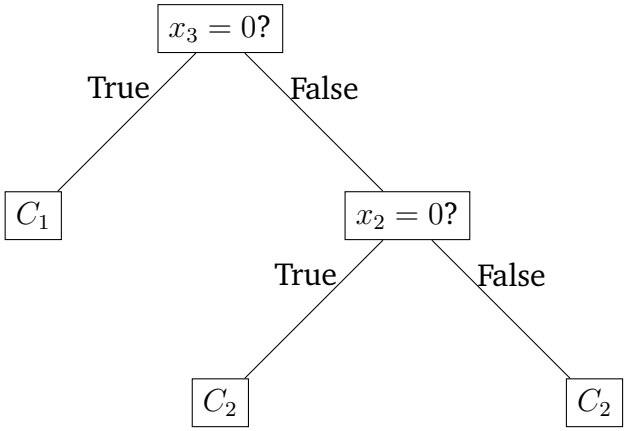


From part 3.2, we saw that the left branch has a perfect classification rate, since there are 40 C1s and 0 C2s in that node, so the left node is a leaf. The right node has 10 C_1s and 50 C_2s. Let's see which feature has a lower weighted gini impurity when splitting on the right branch.

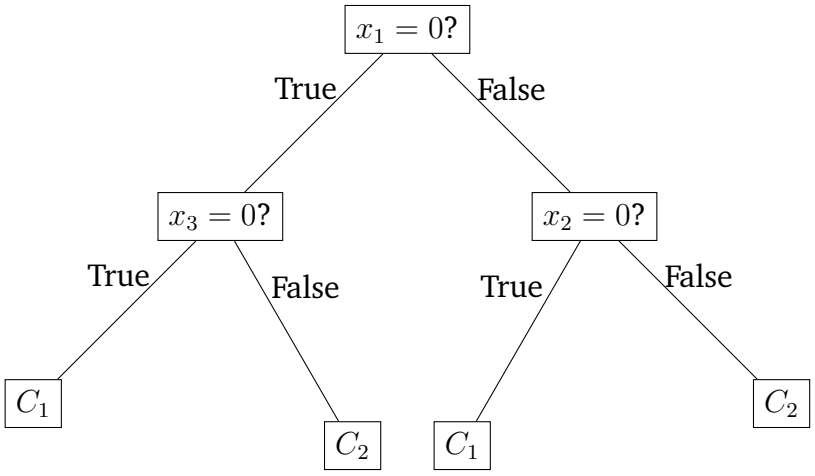
$$G(1) = \frac{25}{60} \cdot \left[\frac{0}{25}\left(1 - \frac{0}{25}\right) + \frac{25}{25}\left(1 - \frac{25}{25}\right)\right] + \frac{35}{60} \cdot \left[\frac{25}{35}\left(1 - \frac{25}{35}\right) + \frac{10}{35}\left(1 - \frac{10}{35}\right)\right]$$
$$= 0 + \frac{7}{12}\left[\frac{5}{7} \cdot \frac{2}{7} + \frac{2}{7} \cdot \frac{5}{7}\right] = \boxed{\frac{5}{21}}$$

$$G(2) = \frac{25}{60} \cdot \left[\frac{10}{25}\left(1 - \frac{10}{25}\right) + \frac{15}{25}\left(1 - \frac{15}{25}\right)\right] + \frac{35}{60} \cdot \left[\frac{0}{35}\left(1 - \frac{0}{35}\right) + \frac{35}{35}\left(1 - \frac{35}{35}\right)\right]$$
$$= \frac{5}{12}\left[\frac{2}{5} \cdot \frac{3}{5} + \frac{3}{5} \cdot \frac{2}{5}\right] + 0 = \boxed{\frac{1}{5}}$$

$G(2) < G(1)$, so we split on x_2 at the second level of the tree. There are 10 points misclassified.



3.4



Left internal node:

$$\epsilon(2) = \frac{25}{50}\left(\frac{10}{25}\right) + \frac{25}{50}\left(\frac{10}{25}\right) = \frac{2}{5}$$
$$\epsilon(3) = \frac{25}{50}\left(\frac{0}{25}\right) + \frac{25}{50}\left(\frac{0}{25}\right) = 0$$

So choose to split on x_3 for the left side.

Right internal node:

$$\epsilon(2) = \frac{25}{50}\left(\frac{0}{25}\right) + \frac{25}{50}\left(\frac{0}{25}\right) = 0$$
$$\epsilon(3) = \frac{15}{50}\left(\frac{0}{15}\right) + \frac{35}{50}\left(\frac{10}{35}\right) = \frac{1}{5}$$

So choose to split on x_2 for the right side.

This tree does not misclassify any point.

3.5

the tree in 3.4 performed better in 3.3, despite x_3 having a lower cost measure at the first level. This shows the greedy approach is indeed suboptimal. The tree in 3.3 did worse because it suffered from low data issues, as there were relatively too few examples with label C_1 to form a majority in any leaf. Thus, there is bound to be misclassification in that case.

4 Problem 4

4.1

$$P(y=1|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y=1)P(y=1)}{P(\boldsymbol{x}|y=1)P(y=1) + P(\boldsymbol{x}|y=0)P(y=0)}$$

$$= \frac{P(\boldsymbol{x}|y=1)P(y=1)/P(\boldsymbol{x}|y=1)P(y=1)}{P(\boldsymbol{x}|y=1)P(y=1)/P(\boldsymbol{x}|y=1)P(y=1) + P(\boldsymbol{x}|y=0)P(y=0)/P(\boldsymbol{x}|y=1)P(y=1)}$$

$$= \frac{1}{1 + \frac{P(\boldsymbol{x}|y=0)P(y=0)}{P(\boldsymbol{x}|y=1)P(y=1)}}$$
$$= \frac{1}{1 + \frac{P(y=0|\boldsymbol{x})}{P(y=1|\boldsymbol{x})}}$$

4.2

$$
\begin{aligned}
P(y=k|\boldsymbol{x}) &= \frac{P(\boldsymbol{x}|y=k)P(y=k)}{p(\boldsymbol{x})} && \text{(Bayes Rule)} \\
&= \frac{\prod_{j=1}^{D} P(x_j|y=k)P(y=k)}{P(\boldsymbol{x})} && \text{(Naive Bayes Assumption)} \\
&= \frac{1}{P(\boldsymbol{x})}\pi_k \prod_{j=1}^{D} P(x_j|y=k) \\
&= \frac{1}{P(\boldsymbol{x})} \exp\left\{ \ln\left(\pi_k \prod_{j=1}^{D} P(x_j|y=k)\right)\right\} && \text{Because } (e^{ln(a)}=a) \\
&= \frac{1}{P(\boldsymbol{x})} \exp\left\{ \ln \pi_k + \sum_{j=1}^{D} \ln\left(P(x_j|y=k)\right)\right\} \\
&= \frac{1}{P(\boldsymbol{x})} \exp\left\{ \ln \pi_k + \sum_{j=1}^{D} \ln\left(\theta_{jk}^{x_j}(1-\theta_{jk})^{1-x_j}\right)\right\} \\
&= \frac{1}{P(\boldsymbol{x})} \exp\left\{ \ln \pi_k + \sum_{j=1}^{D} \left(\ln \theta_{jk}^{x_j} + \ln(1-\theta_{jk})^{1-x_j}\right)\right\} \\
&= \frac{1}{P(\boldsymbol{x})} \exp\left\{ \ln \pi_k + \sum_{j=1}^{D} \left(x_j \ln \theta_{jk} + (1-x_j)\ln(1-\theta_{jk})\right)\right\} \\
&= \frac{1}{P(\boldsymbol{x})} \exp\left\{ \ln \pi_k + \sum_{j=1}^{D} \left(x_j(\ln \theta_{jk} - \ln(1-\theta_{jk})) + \ln(1-\theta_{jk})\right)\right\}
\end{aligned}
$$

4.3

$$
\frac{P(y=0|x)}{P(y=1|x)} = \frac{\cancel{\frac{1}{P(\boldsymbol{x})}} \exp\left\{ \ln(1-\pi) + \sum_{j=1}^{D} \left(x_j(\ln \theta_{j0} - \ln(1-\theta_{j0})) + \ln(1-\theta_{j0})\right)\right\}}{\cancel{\frac{1}{P(\boldsymbol{x})}} \exp\left\{ \ln \pi + \sum_{j=1}^{D} \left(x_j(\ln \theta_{j1} - \ln(1-\theta_{j1})) + \ln(1-\theta_{j1})\right)\right\}}
$$

$$
= \exp\left\{ \ln\left(\frac{1-\pi}{\pi}\right) + \sum_{j=1}^{D} \ln(1-\theta_{j0}) - \sum_{j=1}^{D} \ln(1-\theta_{j1}) + \sum_{j=1}^{D} x_j(\ln \theta_{j0} - \ln(1-\theta_{j0})) - \sum_{j=1}^{D} x_j(\ln \theta_{j1} - \ln(1-\theta_{j1}))\right\}
$$

$$
= \exp\left\{ \underbrace{\ln\left(\frac{1-\pi}{\pi}\right) + \sum_{j=1}^{D} \ln\left(\frac{1-\theta_{j0}}{1-\theta_{j1}}\right)}_{-w_0} + \underbrace{\sum_{j=1}^{D} x_j \ln\left(\frac{\theta_{j0}(1-\theta_{j1})}{\theta_{j1}(1-\theta_{j0})}\right)}_{\boldsymbol{w}^T\boldsymbol{x}} \right\}
$$

Thus, plugging the above expression into the equation in part 4.1, the conditional probability distribution for whether the art is genuine can be expressed by a sigmoid function.

We have:

$$
\begin{aligned}
-w_0 &= \ln\left(\frac{1-\pi}{\pi}\right) + \sum_{j=1}^{D} \ln\left(\frac{1-\theta_{j0}}{1-\theta_{j1}}\right) \\
\boldsymbol{w}^T &= \left(\frac{\theta_{10}(1-\theta_{11})}{\theta_{11}(1-\theta_{10})}, ..., \frac{\theta_{j0}(1-\theta_{j1})}{\theta_{j1}(1-\theta_{j0})}, ..., \frac{\theta_{D0}(1-\theta_{D1})}{\theta_{D1}(1-\theta_{D0})}\right) \\
\boldsymbol{x} &= \begin{pmatrix} x_1 \\ \vdots \\ x_j \\ \vdots \\ x_D \end{pmatrix}
\end{aligned}
$$