Part. 2, Questions (20%): 0819823 陳子祈

1. (10%) Consider a data set comprising 400 data points from class C_1 and 400 data points from class C_2. Suppose that a tree model A splits these into (300, 100) at the first leaf node and (100, 300) at the second leaf node, where (n, m) denotes that n points are assigned to C_1 and m points are assigned to C_2. Similarly, suppose that a second tree model B splits them into (200, 400) and (200, 0). **Evaluate the <u>misclassification rates</u> for the two trees and hence show that they are equal.** Similarly, **evaluate the cross-entropy** $Entropy = -\sum_{k=1}^{K} p_k \log_2 p_k$ and **Gini index** $Gini = 1 - \sum_{k=1}^{K} p_k^2$ **for the two trees and show that they are both lower for tree B than for tree A. Define** p_k to be the proportion of data points in region R assigned to class k, where k = 1, . . . , K

(A) (400, 400) — 50% 50%

C1: (300, 100) — 75% 25%

C2: (100, 300) — 25% 75%

(400, 400) — 50% 50%

C2: (200, 400) — 33% 67%

C1: (200, 0) — 100% 0%

Misclassification rate

A left leaf node misclassification rate = $\frac{100}{800}$, A right leaf node misclassification rate = $\frac{100}{800}$

$$P_m(A) = \frac{100}{800} + \frac{100}{800} = \frac{1}{4}$$

B left leaf node misclassification rate = $\frac{200}{800}$, B right leaf node misclassification rate = 0

$$P_m(B) = \frac{200}{800} = \frac{1}{4} \Rightarrow P_m(A) = P_m(B)$$

Entropy information

$$I_E(D_{parent}) = -\left(0.5\, lg(0.5) + 0.5\, lg(0.5)\right) = 1$$

A: $$I_E(D_{left}) = -\left(\frac{3}{4} lg\left(\frac{3}{4}\right) + \frac{1}{4} lg\left(\frac{1}{4}\right)\right) \approx 0.81$$

$$I_E(D_{right}) = -\left(\frac{1}{4} lg\left(\frac{1}{4}\right) + \frac{3}{4} lg\left(\frac{3}{4}\right)\right) \approx 0.81$$

$$I_{EA} = 0.81 + 0.81 = 1.62$$

B: $$I_E(D_{left}) = -\left(\frac{2}{6} lg\left(\frac{2}{6}\right) + \frac{4}{6} lg \frac{4}{6}\right) \approx 0.92$$

$$I_E(D_{right}) = 0$$

$$I_{EB} = 0.92 \Rightarrow I_{EB} < I_{EA}$$

Gini information

$$I_G(D_{parent}) = 1 - (0.5^2 + 0.5^2) = 0.5$$

A: $$I_G(D_{left}) = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = \frac{3}{8} = 0.375$$

$$I_G(D_{right}) = 1 - \left[\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right] = \frac{3}{8} = 0.375$$

$$I_{GA} = 0.375 + 0.375 = 0.75$$

B: $$I_G(D_{left}) = 1 - \left[\left(\frac{2}{6}\right)^2 + \left(\frac{4}{6}\right)^2\right] = \frac{4}{9} = 0.\overline{4}$$

$$I_G(D_{right}) = 1 - (1^2 + 0^2) = 0$$

$$I_{GB} = 0.\overline{4}$$

$$\Rightarrow \quad I_{GB} < I_{GA}$$

※

2. (10%) By making a variational minimization of the expected exponential error function given by (1) with respect to all possible functions $y(x)$, show that the minimizing function is given by (2). Define t **is target variable** $\in$ **{−1, 1}, x is input vector.**

$$E_{x,t}\,[e^{-ty(x)}] = \sum_t \int e^{-ty(x)} p(t|x)p(x)\,dx \qquad (1)$$

$$y(x) = \frac{1}{2} ln \frac{p(t=1|x)}{p(t=-1|x)} \qquad (2)$$

$$\frac{\partial}{\partial y} E_{x,t}[e^{-ty}] = \frac{\partial}{\partial y} \sum_t \int e^{-ty} p(t|x)p(x)dx$$

$$= \frac{\partial}{\partial x} \int [e^{-y} p(t=1|x)p(x) + e^{y} p(t=-1|x)p(x)]\,dx\,\frac{\partial x}{\partial y}$$

$$= [e^{-y} p(t=1|x)p(x) + e^{y} p(t=-1|x)p(x)]\frac{\partial x}{\partial y}$$

令 $[e^{-y} p(t=1|x)p(x) + e^{y} p(t=-1|x)p(x)]\frac{\partial x}{\partial y} = 0$

左右同時對 x 微分

$$\frac{\partial}{\cancel{\partial x}}[e^{-y} p(t=1|x)p(x) + e^{y} p(t=-1|x)p(x)]\frac{\cancel{\partial x}}{\partial y} = 0$$

$$\frac{\partial}{\partial y}[e^{-y} p(t=1|x)p(x) + e^{y} p(t=-1|x)p(x)] = 0$$

$$\Rightarrow -e^{-y} p(t=1|x)p(x) + e^{y} p(t=-1|x)p(x) = 0$$

$$\Rightarrow e^{y} p(t=-1|x)\cancel{p(x)} = e^{-y} p(t=1|x)\cancel{p(x)}$$

$$\Rightarrow e^{2y} = \frac{p(t=1|x)}{p(t=-1|x)}$$

$$\Rightarrow 2y = ln \frac{p(t=1|x)}{p(t=-1|x)} \rightarrow y = \frac{1}{2} ln\left(\frac{p(t=1|x)}{p(t=-1|x)}\right)$$

※