AI 534: Machine Learning

Assignment III

Name: Nam Nguyen

ID: 934-422-327

Problem 1: Naive Bayes Classifier

a) • $p(y=1) = \frac{1}{2}$

• $p(A=0 \mid y=1) = \frac{1}{3}$

• $p(B=0 \mid y=1) = \frac{1}{3}$

• $p(C=0 \mid y=1) = \frac{2}{3}$

• $p(A=0 \mid y=0) = \frac{2}{3}$

• $p(B=0 \mid y=0) = \frac{1}{3}$

• $p(C=0 \mid y=0) = \frac{1}{3}$

b) Let X is event that $A=1$, $B=0$, and $C=0$.

• $$p(y=1 \mid X) = \frac{p(y=1)\, P(A=1 \mid y=1)\, P(B=0 \mid y=1)\, P(C=0 \mid y=1)}{P(A=1, B=0, C=0)}$$

$$= \frac{(0.5)(2/3)(1/3)(2/3)}{P(A=1, B=0, C=0)} = \frac{2/27}{P(A=1, B=0, C=0)}$$

• $$p(y=0 \mid X) = \frac{p(y=0)\, P(A=1 \mid y=0)\, P(B=0 \mid y=0)\, P(C=0 \mid y=0)}{P(A=1, B=0, C=0)}$$

$$= \frac{(0.5)(1/3)(1/3)(1/3)}{P(A=1, B=0, C=0)} = \frac{1/54}{P(A=1, B=0, C=0)}$$

- We have, $p(y=1|X) + p(y=0|X) = 1$

$\Rightarrow P(A=1, B=0, C=0) = \frac{2}{27} + \frac{1}{54} = \frac{5}{54}$

- Hence,

$$p(y=1|X) = p(y=1 \mid A=1, B=0, C=0)$$

$$= \frac{2/27}{P(A=1, B=0, C=0)} = \frac{2/27}{5/54} = \frac{4}{5}$$

c)

- No
- The indepdent $\neq$ conditional independent

$p(A, B, C) = p(A)\, p(B)\, p(C) \rightarrow$ independent

$p(A, B, C \mid y) = p(A|y)\, p(B|y)\, p(C|y) \rightarrow$ conditional independent

Problem 2: (Naive Bayes learns linear decision boundary)

a) Bernoulli Naive Bayes Model:

- We have,

$$\log \frac{P(y=1|\underline{X})}{P(y=0|\underline{X})} = 0 \Rightarrow \log \frac{P(\underline{X}|y=1)P(y=1)}{P(\underline{X}|y=0)P(y=0)} = 0$$

$$\Rightarrow \log \frac{P(y=1)\prod_{i=1}^{d} P(x_i|y=1)}{P(y=0)\prod_{i=1}^{d} P(x_i|y=0)} = 0$$

$$\Rightarrow \log \frac{P(y=1)\prod_{i=1}^{d} P(x_i=1|y=1)^{x_i}\left(1-P(x_i=1|y=1)\right)^{1-x_i}}{P(y=0)\prod_{i=1}^{d} P(x_i=1|y=0)^{x_i}\left(1-P(x_i=1|y=0)\right)^{1-x_i}} = 0$$

$$\Rightarrow \log \frac{P(y=1)}{P(y=0)} + \log \prod_{i=1}^{d}\left(\frac{P(x_i=1|y=1)}{P(x_i=1|y=0)}\right)^{x_i} + \log \prod_{i=1}^{d}\left(\frac{1-P(x_i=1|y=1)}{1-P(x_i=1|y=0}\right)^{1-x_i} = 0$$

$$\Rightarrow \log \frac{P(y=1)}{P(y=0)} + x_i \sum_{i=1}^{d} \log\left(\frac{P(x_i=1|y=1)}{P(x_i=1|y=0)}\right)$$

$$+ (1-x_i) \sum_{i=1}^{d} \log \left(\frac{1 - P(x_i = 1 \mid y = 1)}{1 - P(x_i = 1 \mid y = 0)} \right) = 0$$

$$\Rightarrow \log \frac{P(y=1)}{P(y=0)} + \sum_{i=1}^{d} x_i \left(\log \frac{P(x_i = 1 \mid y=1)}{P(x_i=1 \mid y=0)} - \log \frac{1-P(x_i=1 \mid y=1)}{1-P(x_i=1 \mid y=0)} \right)$$

$$+ \sum_{i=1}^{d} \log \frac{1 - P(x_i = 1 \mid y=1)}{1 - P(x_i = 1 \mid y = 0)} = 0$$

- Hence,

$$\begin{cases} w_0 = \log \dfrac{P(y=1)}{P(y=0)} + \displaystyle\sum_{i=1}^{d} \log \dfrac{1 - P(x_i = 1 \mid y=1)}{1 - P(x_i = 1 \mid y=0)} \\[2em] w_i = \log \dfrac{P(x_i = 1 \mid y=1)}{P(x_i = 1 \mid y=0)} - \log \dfrac{1 - P(x_i = 1 \mid y=1)}{1 - P(x_i = 1 \mid y = 0)} \end{cases}$$

b)

- We have,

$$\log \frac{P(y=1 \mid \underline{x})}{P(y=0 \mid \underline{x})} = 0 \Rightarrow \log \frac{P(\underline{x} \mid y=1)\, P(y=1)}{P(\underline{x} \mid y=0)\, P(y=0)} = 0$$

$$\Rightarrow \log \frac{P(y=1) \prod_{i=1}^{d} P(x_i \mid y=1)^{x_i}}{P(y=0) \prod_{i=1}^{d} P(x_i \mid y=0)^{x_i}} = 0$$

$$\Rightarrow \log \frac{P(y=1)}{P(y=0)} + \log \prod_{i=1}^{d} \left(\frac{P(w_i | y=1)}{P(w_i | y=0)} \right)^{x_i} = 0$$

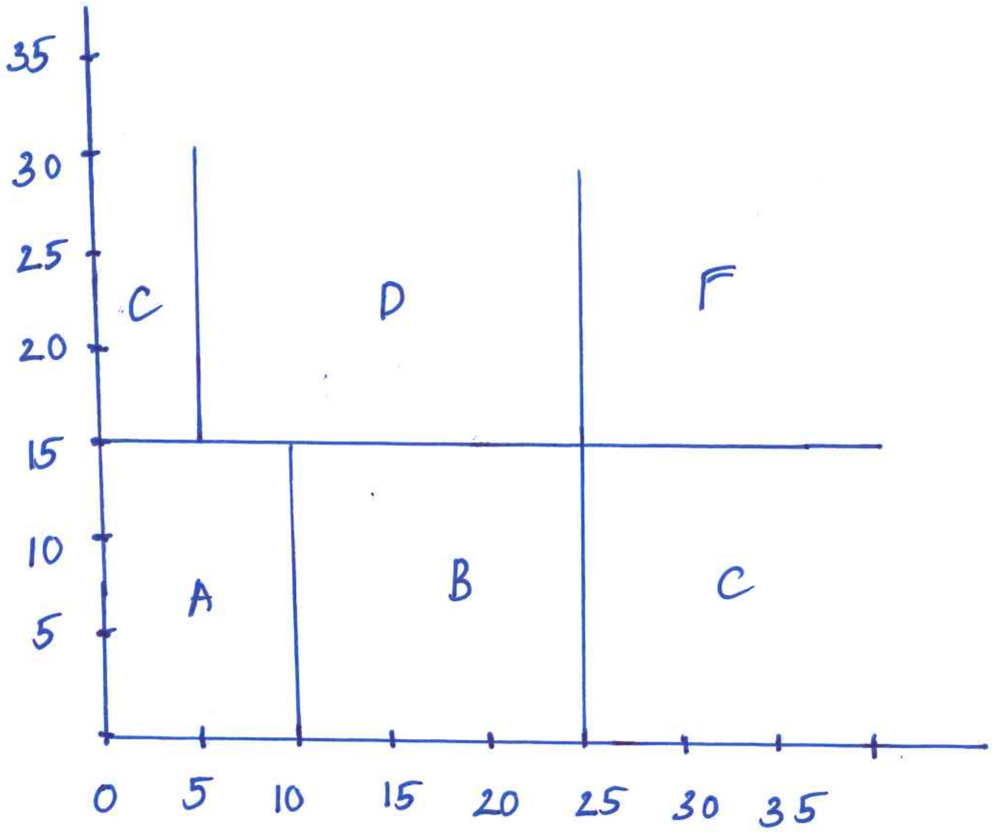
$$\Rightarrow \log \frac{P(y=1)}{P(y=0)} + x_i \sum_{i=1}^{d} \frac{P(w_i | y=1)}{P(w_i | y=0)} = 0$$

- Hence,

$$\begin{cases} w_0 = \log \dfrac{P(y=1)}{P(y=0)} \\ w_i = \displaystyle\sum_{i=1}^{d} \dfrac{P(w_i | y=1)}{P(w_i | y=0)} \end{cases}$$

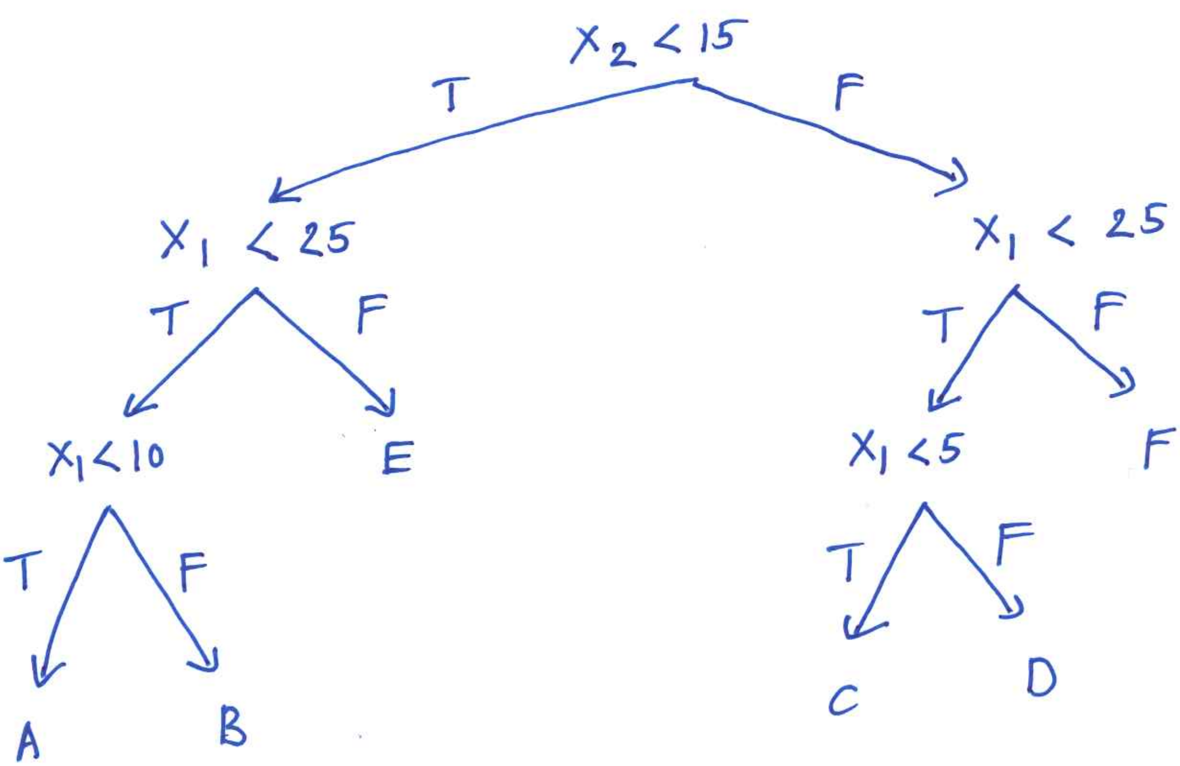
Problem 3:

a)



Decision boundary defined by that tree.

b)



c)

- The redundancy likely provides a computational benefit, as it simplifies the task for an imperfect greedy heuristic in discovering a viable solution. The presence of redundancy enhances the likihood that any random sequence of node expansion's will result in a good tree.

Problem 4:

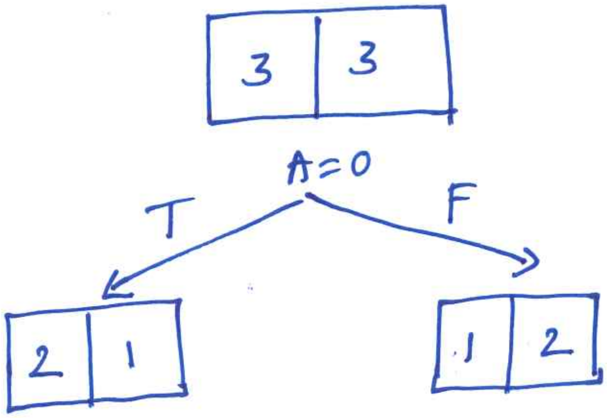
a) The maximum number of leaf nodes is m (leaves) If each training examples were isolated in its own leaf.

b) In the worst-case scenario, we could have m maximum number of leaf nodes by employing maximal mutual information However, on average, we could expect a smaller number of leaf nodes.

c). On average, the use of randomized splits would yield lower accuracy, especially in the presence of irrelevant or noisy features in the data.

- Random splits are more prone to dividing based on irrelevant features, creating two learning problems equivalent to the original one, but only with half as much data.
- Hence, a random split is more likely to introduce inaccurate of the decision trees due to the reduced amount of data available for each learning problem.

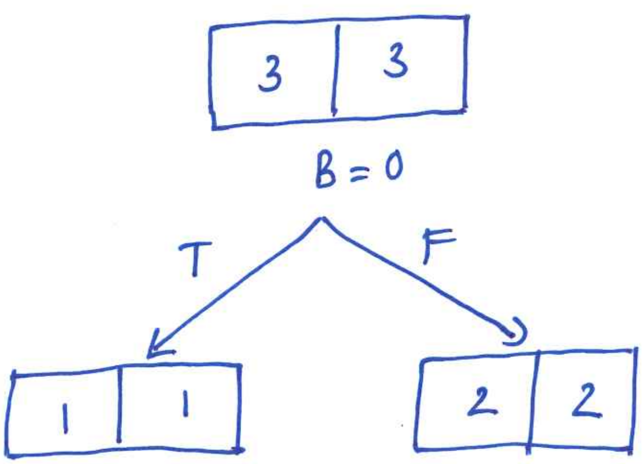
Problem 5:



$$\begin{cases} H(y \mid A=0) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.918 \\ H(y \mid A=1) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.918 \end{cases}$$

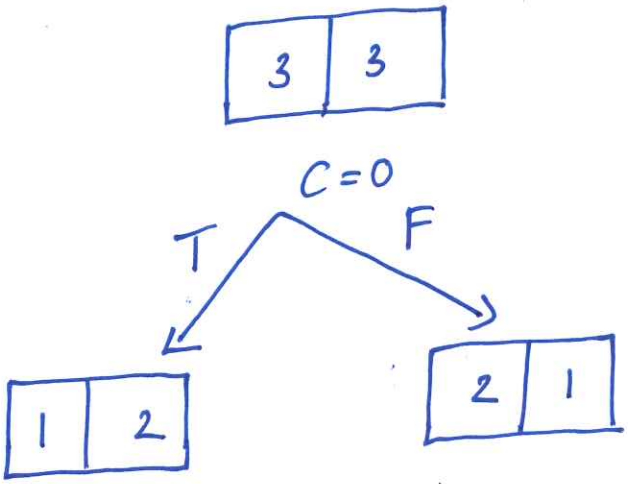
$$\Rightarrow H(y \mid A) = p(A=0)\,H(y \mid A=0) + p(A=1)\,H(y \mid A=1)$$

$$= 0.918$$



$$\begin{cases} H(y \mid B=0) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1 \\ H(y \mid B=1) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1 \end{cases}$$

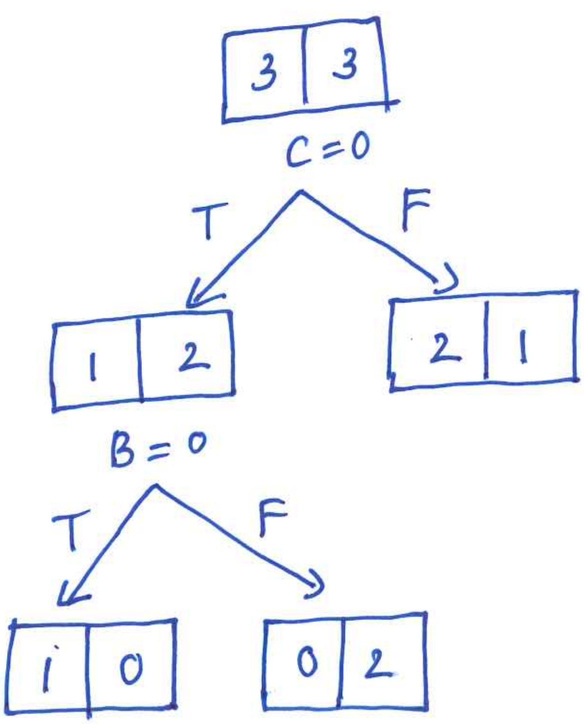
$$\Rightarrow H(y|B) = p(B=0)\,H(y|B=0) + p(B=1)\,H(y|B=1) = 1$$

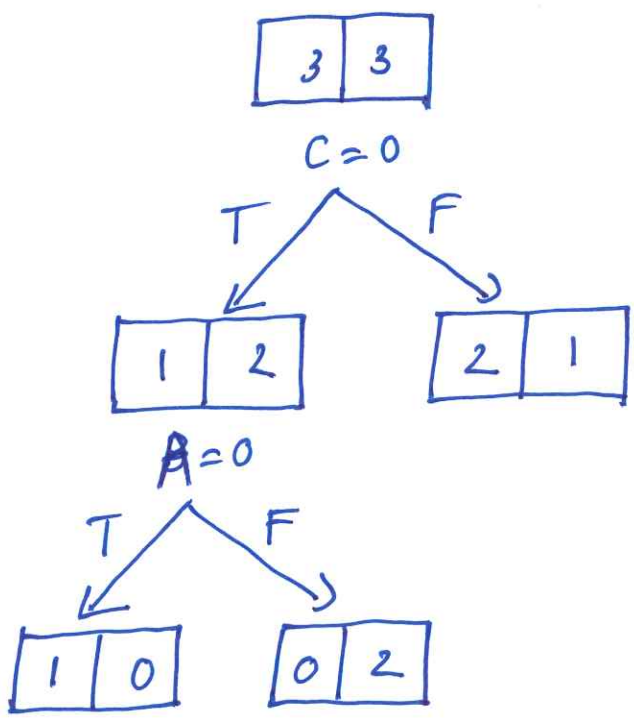


$$\cdot \begin{cases} H(y|C=0) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.918 \\ H(y|C=1) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.918 \end{cases}$$

$$\Rightarrow H(y|C) = p(C=0)\,H(y|C=0) + p(C=1)\,H(y|C=1)$$
$$= 0.918$$

- We pick C as the root node test randomly ($H(Y|A) = H(Y|C)$)





- B is better choice than A. We choose the figure on the left hand side.

Problem 6:

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

- Let's show

$$H(X,Y) = H(X) + H(Y|X)$$

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$$

$$= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x)\,p(y|x)$$

$$= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)$$

$$= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)$$

$$= H(X) + H(Y|X)$$

$$\Rightarrow H(X,Y) = H(X) + H(Y|X) \quad (1)$$

- Let's show

$$H(X,Y) = H(Y) + H(X|Y)$$

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$$

$$= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(y)\, p(x|y)$$

$$= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(y) - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(x|y)$$

$$= -\sum_{y\in\mathcal{Y}} p(y) \log p(y) - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(x|y)$$

$$= H(Y) + H(X|Y)$$

$$\Rightarrow H(X,Y) = H(Y) + H(X|Y) \quad ②$$

- From ① and ②, we have

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \blacksquare$$