1. **Bayes Optimal Classification**
    1. **Write the Bayes optimal classifier in terms of indicator function and p(Y = 0|X = x)**

For the special case of a binary classifier, i.e. $Y \in \{0, 1\}$, $f*(x)$ can be equivalently written as:

$$f* = \begin{cases} 1, & \text{if } P(Y = 1 \mid X = x) \geq 1/2 \\ 0, & \text{otherwise} \end{cases}$$

We will convert this using indicator function (I[]) and p(Y = 0|X = x) to the following:

$$f* = \begin{cases} 1, & \text{if } I[P(Y = 0 \mid X = x)] < 1/2 \\ 0, & \text{otherwise} \end{cases}$$

2. Show that for any binary classifier *f* and some *x*, the probability of error is
$P(Y \neq f(X) \mid X = x) = (1 - 2P(Y = 0 \mid X = x))\, I[f(x) = 0] + P(Y = 0 \mid X = x)$

Note that $P(Y \neq f(X) \mid X = x) = P(E) = \begin{cases} P(y = 1 \mid X = x), & \text{if } I[f(x) = 0] \\ P(y = 0 \mid X = x), & \text{if } I[f(x) = 1] \end{cases}$
We must add both probabilities and then reduce while including the product of their identity functions.

$P(Y \neq f(X) \mid X = x)$ $= P(y=1|X=x)I[f(x) = 0] + P(y=0|X=x)I[f(x) = 1]$

Note that $I[f(x) = 1] + I[f(x) = 0] = 1$, Thus $I[f(x) = 1]$ can be rewritten as
$I[f(x) = 1] = 1 - I[f(x) = 0]$, we replace this into our previous step to get;

$P(Y \neq f(X) \mid X = x)$ $= P(y=1|X=x)I[f(x) = 0] + P(y=0|X=x)(1 - I[f(x) = 0])$

Similarly, $P(y=1|X=x) + P(y=0|X=x) = 1$, Thus $P(y=1|X=x)$ can be rewritten as
$P(y=1|X=x) = 1 - P(y=0|X=x)$, we replace this into our to get

$P(Y \neq f(X) \mid X = x)$ $= (1 - P(y=0|X=x))I[f(x) = 0] + P(y=0|X=x)(1 - I[f(x) = 0])$
$= I[f(x) = 0] - P(y=0|X=x)I[f(x) = 0] + P(y=0|X=x) - P(y=0|X=x)I[f(x) = 0]$
$= I[f(x) = 0] - 2P(y=0|X=x)I[f(x) = 0] + P(y=0|X=x)$
$= \boldsymbol{(1 - 2P(y=0|X=x))I[f(x) = 0] + P(y=0|X=x)}$

3. **Show that for any binary classifier f and some x, the probability of error is more than Bayes optimal classifier $f*$. In particular show that $P(Y \neq f(X)|X = x) - P(Y \neq f*(X)|X = x) \geq 0$.**
4. **Show that bayes optimal binary classifier has the minimum risk. In particular show that $R(f*) \leq R(f)$ for any binary classifier f .**

2. Decision Tree

1. For each input variable, compute the cost of splitting at that input variable using misclassification rate, and Gini index as impurity measure.

Gini-Index: $G(P) = \sum_{k=1}^{C} P(Y = k)(1 - P(Y = k))$

Cost of splitting at each input:

X_1;

For Branch '0'

$$\frac{25}{50}(1 - \frac{25}{50}) + \frac{25}{50}(1 - \frac{25}{50}) = 0.5$$

For Branch '1'

$$\frac{25}{50}(1 - \frac{25}{50}) + \frac{25}{50}(1 - \frac{25}{50}) = 0.5$$

Total Cost for X_1 = $\frac{50}{100}$*0.5 + $\frac{50}{100}$*0.5 = 0.5

X_2;

For Branch '0'

$$\frac{30}{50}(1 - \frac{30}{50}) + \frac{20}{50}(1 - \frac{20}{50}) = 0.48$$

For Branch '1'

$$\frac{20}{50}(1 - \frac{20}{50}) + \frac{30}{50}(1 - \frac{30}{50}) = 0.48$$

Total Cost for X_2 = $\frac{50}{100}$*0.48 + $\frac{50}{100}$*0.48 = 0.48

X_3;

For Branch '0'

$$\frac{25}{25}(1 - \frac{25}{25}) + \frac{0}{25}(1 - \frac{0}{25}) = 0$$

For Branch '1'

$$\frac{25}{75}(1 - \frac{25}{75}) + \frac{50}{75}(1 - \frac{50}{75}) = 0.45$$

Total Cost for X_3 = $\frac{25}{100}$*0 + $\frac{75}{100}$*0.45 = 0.34

2. Using misclassification rate as criteria, decide which input variable should we first split on to create a two-level decision tree.

We should split on X_3 first since it has the smallest split cost.

3. **Repeat the splitting procedure for the children nodes of tree in previous problem to construct a two level decision tree. Describe the process and draw the decision tree. How many points are still misclassified?**

From our 1st question we know that X3 will be our root node since it has the smallest splitting cost. We repeat the process of computing the split cost for all the other input variables but now the total possible is 75 instead of 100 since we removed 25 instances in a pure Child from X3. The splitting cost are as follows;

X_1;

For Branch '0'

$$\frac{0}{25}\left(1 - \frac{0}{25}\right) + \frac{25}{25}\left(1 - \frac{25}{25}\right) = 0$$

For Branch '1'

$$\frac{25}{50}\left(1 - \frac{25}{50}\right) + \frac{25}{50}\left(1 - \frac{25}{50}\right) = 0.5$$

Total Cost for X_1 = $\frac{25}{75}$***0 +** $\frac{50}{75}$***0.5 = 0.34**

X_2;

For Branch '0'

$$\frac{25}{45}\left(1 - \frac{25}{45}\right) + \frac{20}{45}\left(1 - \frac{20}{45}\right) = 0.49$$

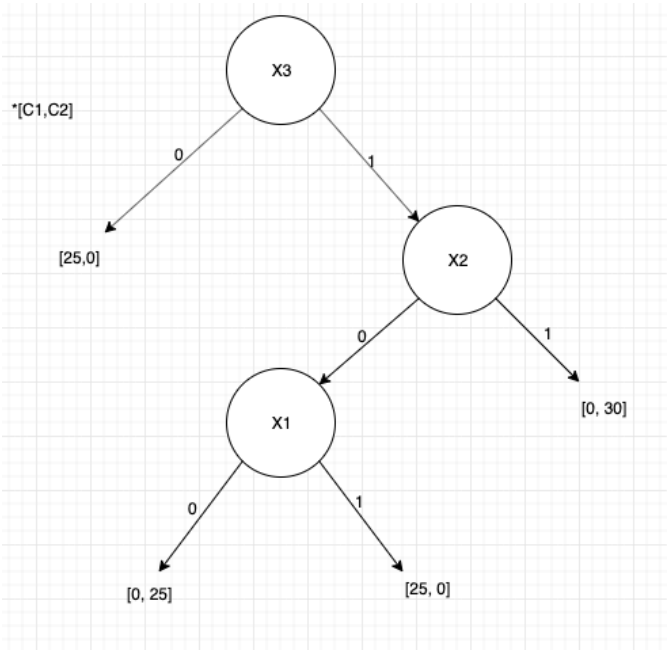
For Branch '1'

$$\frac{0}{30}\left(1 - \frac{0}{30}\right) + \frac{30}{30}\left(1 - \frac{30}{30}\right) = 0$$

Total Cost for X_2 = $\frac{45}{75}$***0.49 +** $\frac{30}{75}$***0 = 0.29**

Since X_2 has a smaller split cost we will split on X_2 next and lastly split on X_1

The tree will look like the following:



There is zero misclassifications as all leafs are pure/certain.

4. **Create a two-level decision tree by first splitting on X_1, and choose the second split by misclassification criteria. How many points are misclassified?**

Since we are splitting from X1 first we will have two non-pure branches with equal probabilities for C1 and C2 classes predictions. Thus in order to get to the next split by misclassification we must compare split cost for X2 and X3 for the case that X1 = 1 and X1 = 0 since we will be splitting both branches.

For the case that X1 is '1'

X_2;

For Branch '0'

$$\frac{25}{25}\left(1 - \frac{25}{25}\right) + \frac{0}{25}\left(1 - \frac{0}{25}\right) = 0$$

For Branch '1'

$$\frac{0}{25}\left(1 - \frac{0}{25}\right) + \frac{25}{25}\left(1 - \frac{25}{25}\right) = 0$$

Total Cost for X_2 = $\frac{25}{50}*0 + \frac{25}{50}*0 = 0$

X_3;

For Branch '0'

$$\frac{0}{0}(1 - \frac{0}{0}) + \frac{0}{0}(1 - \frac{0}{0}) = 0$$

For Branch '1'

$$\frac{25}{25}(1 - \frac{25}{25}) + \frac{25}{25}(1 - \frac{25}{25}) = 0.5$$

Total Cost for X_3 = $\frac{0}{50}$*0 + $\frac{50}{50}$*0.5 = 0.5

X_2 has a smaller split cost thus the branch where X_1 is one will split on X_2 first and then X_3

For the case that X1 is '0'

X_2;

For Branch '0'

$$\frac{5}{25}(1 - \frac{5}{25}) + \frac{20}{25}(1 - \frac{20}{25}) = 0.32$$

For Branch '1'

$$\frac{20}{25}(1 - \frac{20}{25}) + \frac{5}{25}(1 - \frac{5}{25}) = 0.32$$

Total Cost for X_2 = $\frac{25}{50}$*0.32 + $\frac{25}{50}$*0.32 = 0.32

X_3;

For Branch '0'

$$\frac{25}{25}(1 - \frac{25}{25}) + \frac{0}{25}(1 - \frac{0}{25}) = 0$$

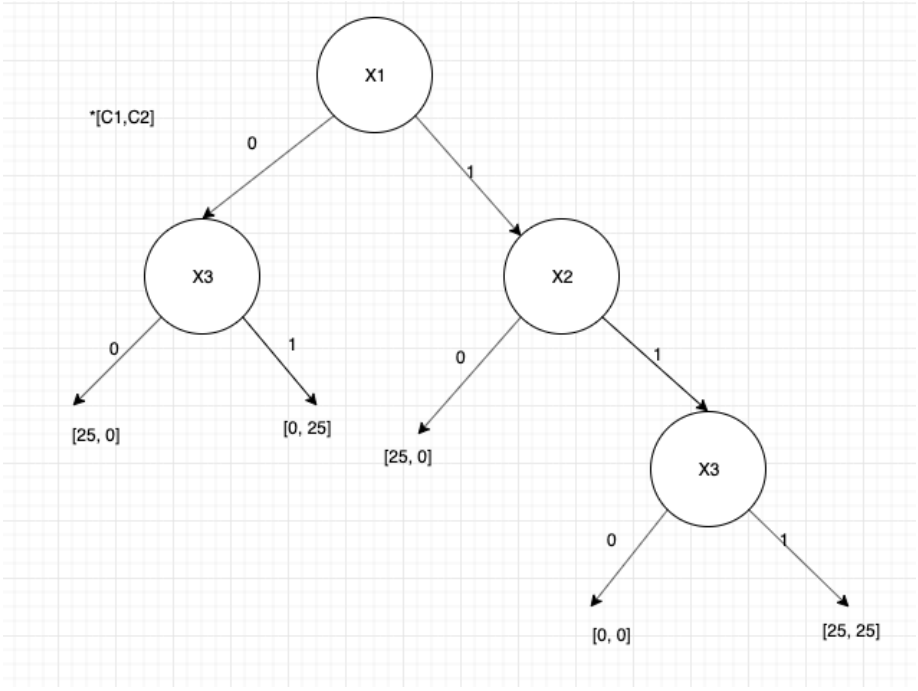
For Branch '1'

$$\frac{0}{25}(1 - \frac{0}{25}) + \frac{25}{25}(1 - \frac{25}{25}) = 0$$

Total Cost for X_3 = $\frac{25}{50}$*0 + $\frac{25}{50}$*0 = 0

X_3 has a smaller split cost thus the branch where X_1 is zero we will split on X_3, we have no non-pure/certain branches from X_3 and thus we stop there.

The full tree will look like the following:



We will have 25 misclassifications.

5. **Compare and comment on different decision trees from problem 2.3 and 2.4. Which decision tree performs better and why?**

Our Tree on question 2.3 is better because there is more certainty when conducting our predictions, we have all pure leafs and so we will get a correct prediction according to our training data. Our tree at 2.4 has some misclassifications and thus may give us a wrong prediction in some cases.

3. Naive Bayes

1. **Assuming X = [X1; X2; X3], write the expression for P(C|X) in terms of P(X1| C), P(X2|C), P(X3|C) and P(C) for a naive bayes model.**

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

Since X = [X1; X2; X3]

$$\text{Then } P(C|X) = \frac{P(C)P(X1|C)P(X2|C)P(X3|C)}{P(X1)P(X2)P(X3)}$$

2. Compute P(X1|C), P(X2|C), P(X3|C) and P(C) from the data.

P(X1|C);

$$P(X1 = 0|\ C = C1) = \frac{25}{50} = \frac{1}{2}$$
$$P(X1 = 0|\ C = C2) = \frac{25}{50} = \frac{1}{2}$$
$$P(X1 = 1|\ C = C1) = \frac{25}{50} = \frac{1}{2}$$
$$P(X1 = 1|\ C = C2) = \frac{25}{50} = \frac{1}{2}$$
$$P(X1|C) = \frac{1}{2} \text{ x } \frac{1}{2} \text{ x } \frac{1}{2} \text{ x } \frac{1}{2} = \frac{1}{16}$$

P(X2|C);

$$P(X2 = 0|\ C = C1) = \frac{30}{50} = \frac{3}{5}$$
$$P(X2 = 0|\ C = C2) = \frac{20}{50} = \frac{2}{5}$$
$$P(X2 = 1|\ C = C1) = \frac{20}{50} = \frac{2}{5}$$
$$P(X2 = 1|\ C = C2) = \frac{30}{50} = \frac{3}{5}$$
$$P(X2|C) = \frac{3}{5} \text{ x } \frac{2}{5} \text{ x } \frac{2}{5} \text{ x } \frac{3}{5} = \frac{36}{625}$$

P(X3|C);

$$P(X3 = 0|\ C = C1) = \frac{25}{50} = \frac{1}{2}$$
$$P(X3 = 0|\ C = C2) = \frac{0}{50} = 0$$
$$P(X3 = 1|\ C = C1) = \frac{25}{50} = \frac{1}{2}$$
$$P(X3 = 1|\ C = C2) = \frac{50}{50} = 1$$
$$P(X3|C) = \frac{1}{2} \text{ x } 0 \text{ x } \frac{1}{2} \text{ x } 1 = 0$$
$$P(C = C1) = \frac{50}{100} = \frac{1}{2}$$
$$p(C = C2) = \frac{50}{100} = \frac{1}{2}$$

3. What are the most probable labels for each input using the naive bayes model. What will be the misclassification error if we use this model to label the inputs?

X3 will have a most probable label of C1 since the probability of C1 is greater than C2 within our computation

X2 and X1 will have equal probability to be either C1 or C2

4. **Naive Bayes model involves assumption that input attributes are independent given label Y. Discuss why this assumption is bad here? How can we improve the model changing the model and how much the model can be improved? Support your answer with arguments, calculations are not required. (Hint: Think of what dependencies between input attributes should be modeled to improve the model.)**

The assumption here is bad because the probability of a given input should changed *depending* on the value of the other values since you must have 3 input values to get a Y value. In example the probability involving X1 to some target label Y could change if the remaining probability of X2 and X3 were either [00,01,10, or 11].