Written Exercises

Question 1 - a

For every $x_i \in \{x_1, x_2, ..., x_N\}$ the density function is $g(x_i)$, hence the log-likelihood of the data is,

$$l = \log\left(\prod_{i}^{N} g\left(x_i\right)\right)$$

$$l = \sum_{i}^{N} \log\left(g\left(x_i\right)\right) = \sum_{i}^{N} \log\left(\sum_{k=1}^{K} \pi_k g_k(x_i)\right)$$

Question 1 - b

Question 3 - a

We are given the min-error impurity function:

$$I(r) = min\{r,1-r\}$$

And the weighted impurity:

$$(p_1+n_1)\cdot I\left(\frac{p_1}{p_1+n_1}\right)+(p_2+n_2)\cdot I\left(\frac{p_2}{p_2+n_2}\right)$$

Now, every term can be written as:

$$(p_1+n_1)\cdot I\left(\frac{p_1}{p_1+n_1}\right)=(p_1+n_1)\cdot min\{\frac{p_1}{p_1+n_1},\frac{n_1}{p_1+n_1}\}=min\{p_1,n_1\}$$

Hence the weighted impurity is written as:

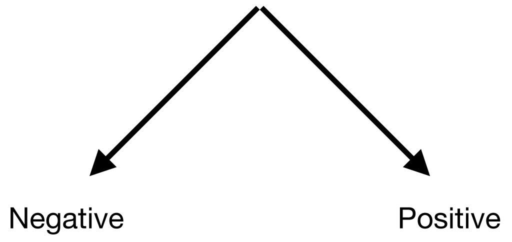
$$min\{p_1,n_1\}+min\{p_2,n_2\}$$

Without loss of generality $p_1 < n_1 \quad p_2 > n_2$, then the left leaf is labeled negative and the right as positive.

That means the number of training mistakes is p_1 on the left leaf and n_2 on the right leaf. A total of p_1+n_2

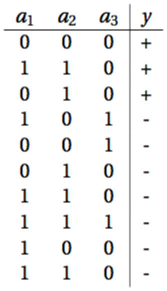
And the weighted impurity is:

$min\{p_1,n_1\}+min\{p_2,n_2\}=p_1+n_2$



Question 3 - b

First, let us look at the 3 possible splits, and then calculate the weighted Gini index and weighted impurity for each split



a_1	a_2	a_3	y
0	0	0	+
1	1	0	+
0	1	0	+
1	0	1	-
0	0	1	-
0	1	0	-
1	1	0	-
1	1	1	-
1	0	0	-
1	1	0	-

$$(2{,}2) \leftarrow a_1 = 0 \text{ [tree 1] } a_1 \neq 0 \rightarrow (1{,}5)$$

$$(1{,}3) \leftarrow a_2 = 0 \text{ [tree 2] } a_2 \neq 0 \rightarrow (2{,}4)$$

$$(3{,}4) \leftarrow a_3 = 0 \text{ [tree 3] } a_3 \neq 0 \rightarrow (0{,}3)$$

The Gini index in a two class case is:

$$I_G(r) = 1 - r^2 - (1-r)^2 = 2r - 2r^2$$

Therefore, the weighted Gini indices of each tree are:

$$\text{tree}_i = (n_1 + p_1) \cdot I_G\left(\frac{p_1}{n_1+p_1}\right) + (n_2+p_2) \cdot I_G\left(\frac{p_2}{n_2+p_2}\right)$$

$$\text{tree}_i = 2(n_1+p_1) \cdot \left(\frac{p_1}{n_1+p_1}\right)\left(\frac{n_1}{n_1+p_1}\right) + 2(n_2+p_2) \cdot \left(\frac{p_2}{n_2+p_2}\right)\left(\frac{n_2}{n_2+p_2}\right)$$

$$\text{tree}_i = \frac{2p_1n_1}{n_1+p_1} + \frac{2p_2n_2}{n_2+p_2}$$

Plug in the numbers drawn above,

$$\text{tree}_1 = \frac{2 \cdot 2 \cdot 2}{2+2} + \frac{2 \cdot 1 \cdot 5}{1+5} = 3.667$$

$$\text{tree}_2 = \frac{2 \cdot 1 \cdot 3}{1+3} + \frac{2 \cdot 2 \cdot 4}{2+4} = 4.167$$

$$\text{tree}_3 = \frac{2 \cdot 3 \cdot 4}{3+4} + \frac{2 \cdot 0 \cdot 3}{0+3} = 3.42$$

As expected, tree 3 classifies 3 negative samples flawlessly in the right side of the split, which minimizes the gini index.
Tree 3 is the split chosen by the Gini index

For the weighted impurity of each tree:

$$\text{tree}_1 = 1 + 2 = 3 \quad \text{tree}_2 = 2 + 1 = 3 \quad \text{tree}_3 = 3 + 0 = 3$$

All the trees are equivalently performant for the min-error impurity function

Question 3 - c

$$N = n_1 + p_1 + n_2 + p_2$$

N samples when n_1, n_2 are negative samples split between the two different branches of the tree and p_1, p_2 are positive samples split between the two different branches of the tree
We want to attain,

min-error impurity before the split > weighted min-error after the split

That is,

$$N \cdot min\left(\frac{p_1+p_2}{N}, \frac{n_1+n_2}{N}\right) > (n_1+p_1) \cdot min\left(\frac{p_1}{n_1+p_1}, \frac{n_1}{n_1+p_1}\right) + (n_2+p_2) \cdot min\left(\frac{p_2}{n_2+p_2}, \frac{n_2}{n_2+p_2}\right)$$

Which reduces to,

$$min\left(p_1+p_2, n_1+n_2\right) > min\left(p_1, n_1\right) + min\left(p_2, n_2\right)$$

Now, if $n_i < p_i$ (or the opposite) for all i , this is impossible. So we constrain only for cases where the minimal class in both sides of the splits is different.
Arbitrarily, $p_1 < n_1 \quad p_2 > n_2$

The inequality is now,

$$min\left(p_1+p_2, n_1+n_2\right) > p_1 + n_2$$

Due to our arbitrary selection, we must set $n_1 + n_2 < p_1 + p_2$
And now,

$$n_1 + n_2 > p_1 + n_2 \rightarrow n_1 > p_1$$

As previously constrained.
So the general conditions are either

$$p_1 < n_1, \quad p_2 > n_2, \quad p_1 + p_2 > n_1 + n_2$$

or

$$p_1 > n_1, \quad p_2 < n_2, \quad p_1 + p_2 < n_1 + n_2$$

Question 3 - d

Min-error impurity seems to not perform very well, it ignores splits that classify accurately (because it always weighs together the number of total errors). Like the split of tree 3 - when one side was a clear no false classifications.

Min-error impurity does not give any accurate indication when the data contains a relatively small sample of a specific class (as the error is already significantly small)