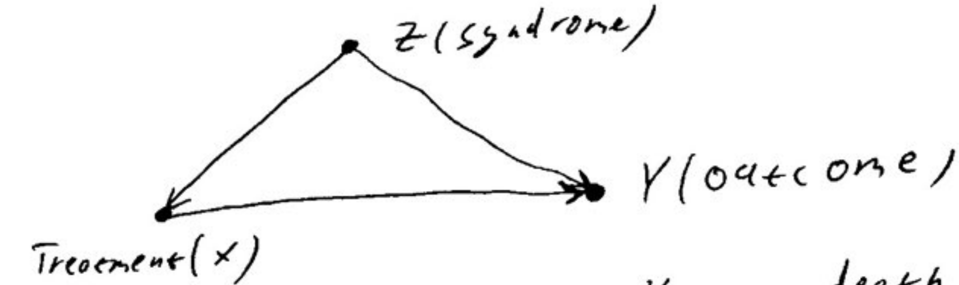
Exercis 6

Ulzii-Utas .N

E1



$Z = z_1$ - presence of syndrome $\quad Y = y_1$ - death

$Z = z_0$ - absence of syndrome $\quad Y = y_0$ - survival

$X = x_1$ - take drug $\quad P(Z=z_1) = r$

$X = x_0$ - don't take drug $\quad P(Z=z_0) = 1-r$

$$\left.\begin{array}{l} P(Y=y_1 \mid Z=z_0, X=x_1) = P_2 \\ P(Y=y_1 \mid Z=z_0, X=x_0) = P_1 \\ P(Y=y_1 \mid Z=z_1, X=x_0) = P_3 \\ P(Y=y_1 \mid Z=z_1, X=x_1) = P_4 \end{array}\right\} \text{for case } Y=y_0 \text{ is to minus 1 } \underline{\underline{1-P(\dots)}}$$

$$\left.\begin{array}{l} q_1 = P(X=x_1 \mid Z=z_0) \\ q_2 = P(X=x_1 \mid Z=z_1) \end{array}\right\} \text{for case } X=x_0 \text{ is to minus } \underline{\underline{1-P(\dots)}}$$

$$P(X,Y,Z) = P(Z)\cdot P(X \mid \underbrace{Z}_{\text{parent}}) \cdot P(Y \mid \underbrace{X, Z}_{\text{parent}})$$

Lets find for all different values of X, Y, Z.

$$P(X=x_0, Y=y_0, Z=z_0) = P(Z=z_0) \cdot P(X=x_0 \mid Z=z_0) \cdot P(Y=y_0 \mid X=x_0, Z=z_0)$$

$$= (1-r)(1-q_1)(1-P_1)$$

$$P(X=x_0, Y=y_1, Z=z_0) = (1-r)(1-q_1)\,p_1$$

$$P(X=x_1, Y=y_0, Z=z_0) = P(Z=z_0)\,P(X=x_1 \mid Z=z_0)\,P(Y=y_0 \mid X=x_1, Z=z_0)$$

$$= (1-r)\,q_1\,(1-p_2)$$

$$P(X=x_1, Y=y_1, Z=z_0) = (1-r)\,q_1\,p_2$$

$$P(X=x_0, Y=y_0, Z=z_1) = r\,(1-q_2)(1-p_3)$$

$$P(X=x_0, Y=y_1, Z=z_1) = r\,(1-q_2)\,p_3$$

$$P(X=x_1, Y=y_0, Z=z_1) = r\,q_2\,(1-p_4)$$

$$P(X=x_1, Y=y_1, Z=z_1) = r\,q_2\,p_4$$

$$P(X,Y) = \sum_Z P(X \mid Z)\,P(Y \mid X, Z)$$

$$P(X=x_0, Y=y_0) = P(X=x_0 \mid Z=z_0)\,P(Y=y_0 \mid X=x_0\ Z=z_0) + P(X=x_0 \mid Z=z_1)\,P(Y=y_0 \mid X=x_0\ Z=z_1)$$

$$= (1-q_1)(1-p_1) + (1-q_2)(1-p_3)$$

$$P(X=x_0, Y=y_1) = (1-q_1)\,p_1 + (1-q_2)\,p_3$$

$$P(X=x_1, Y=y_0) = q_1\,(1-p_2) + q_2\,(1-p_4)$$

$$P(X=x_1, Y=y_1) = q_1\,p_2 + q_2\,p_4$$

$$P(X,Z) = P(Z)\,P(X \mid Z)$$

$$P(X=x_0, Z=z_0) = (1-r)(1-q_1) \qquad P(X=x_1, Z=z_0) = (1-r)\,q_1$$

$$P(X=x_0, Z=z_1) = r\,(1-q_2) \qquad P(X=x_1, Z=z_1) = r\,q_2$$

$$P(Y, Z) = \sum_{X} P(Z)\, P(Y \mid X, Z)$$

$$P(Y=y_0, Z=z_0) = P(Z=z_0)\, P(Y=y_0 \mid X=x_0, Z=z_0) + P(Z=z_0)\, P(Y=y_0 \mid X=x_1, Z=z_0)$$
$$= (1-r)(1-p_1) + (1-r)(1-p_2)$$

$$P(Y=y_0, Z=z_1) = r(1-p_3) + r(1-p_4)$$
$$P(Y=y_1, Z=z_0) = (1-r)\, p_1 + (1-r)\, p_2$$
$$P(Y=y_1, Z=z_1) = r\, p_3 + r\, p_4$$

<u>E2</u> $P(y_1 \mid x_1) - P(y_1 \mid x_0)$

$$P(y \mid x) = \frac{P(x, y)}{P(x)} = \frac{P(x, y)}{P(z)\, P(x \mid z)}$$

(1) $Z = z_1$

$$P(y \mid x) = \frac{P(x \mid z=z_1)\, P(y \mid x, z=z_1)}{P(z=z_1)\, P(x \mid z=z_1)}$$
$$= \frac{P(y \mid x, z=z_1)}{P(z=z_1)}$$

$$P(y_1 \mid x_1) = \frac{P(y_1 \mid x_1, z_1)}{P(z_1)} = \frac{p_4}{r}$$

$$P(y_1 \mid x_0) = \frac{P(y_1 \mid x_0, z_1)}{P(z_1)} = \frac{p_3}{r}$$

$$P(y_1 \mid x_1) - P(y_1 \mid x_0) = \frac{p_4 - p_3}{r}$$

$$P(Y, Z) = \sum_X P(Z)\, P(Y \mid X, Z)$$

$$P(Y=y_0, Z=z_0) = P(Z=z_0)\, P(Y=y_0 \mid X=x_0, Z=z_0) + P(Z=z_0)\, P(Y=y_0 \mid X=x_1, Z=z_0)$$

$$= (1-r)(1-P_1) + (1-r)(1-P_2)$$

$$P(Y=y_0, Z=z_1) = r(1-P_3) + r(1-P_4)$$

$$P(Y=y_1, Z=z_0) = (1-r)\, P_1 + (1-r)\, P_2$$

$$P(Y=y_1, Z=z_1) = r P_3 + r P_4$$

<u>E2</u> $P(y_1 \mid x_1) - P(y_1 \mid x_0)$

$$P(y \mid x) = \frac{P(x, y)}{P(x)} = \frac{P(x, y)}{P(z)\, P(x \mid z)}$$

(1) $Z = z_1$

$$P(y \mid x) = \frac{P(x \mid z=z_1)\, P(y \mid x, z=z_1)}{P(z=z_1)\, P(x \mid z=z_1)}$$

$$= \frac{P(y \mid x, z=z_1)}{P(z=z_1)}$$

$$P(y_1 \mid x_1) = \frac{P(y_1 \mid x_1, z_1)}{P(z_1)} = \frac{P_4}{r}$$

$$P(y_1 \mid x_0) = \frac{P(y_1 \mid x_0, z_1)}{P(z_1)} = \frac{P_3}{r}$$

$$P(y_1 \mid x_1) - P(y_1 \mid x_0) = \frac{P_4 - P_3}{r}$$

② $z = z_0$

$$P(y|x) = \frac{P(\cancel{x}|z_0)\,P(y|x, z_0)}{P(z_0)\,P(x|z_0)} = \frac{P(y|x, z_0)}{P(z_0)}$$

$$P(y_1|x_1) = \frac{P_2}{1-r} \qquad P(y_1|x_0) = \frac{P_1}{1-r}$$

$$P(y_1|x_1) - P(y_1|x_0) = \frac{P_2 - P_1}{1-r}$$

③ $z = (z_0\ z_1)$

$$P(y|x) = \sum_z \frac{P(y|x, z)}{P(z)}$$

$$P(y_1|x_1) = \frac{P_2}{1-r} + \frac{P_4}{r}$$

$$P(y_1|x_0) = \frac{P_1}{1-r} + \frac{P_3}{r}$$

$$P(y_1|x_1) - P(y_1|x_0) = \frac{P_2 - P_1}{1-r} + \frac{P_4 - P_3}{r}$$

$$= \frac{(P_2 + P_3 - P_1 - P_4)r + P_4 - P_3}{(1-r)r}$$

E3 ? not sure.

E4

$$P(y \mid do(x)) = \sum_z P(Y=y \mid X=x, Z=z)\, P(Z=z)$$

$$P(y_0 \mid do(x_0)) = P(y_0 \mid x_0, z_0)\, P(z_0) + P(y_0 \mid x_0 z_1)\, P(z_1)$$
$$= (1-P_1)(1-r) + (1-P_3)\, r$$

$$P(y_1 \mid do(x_0)) = P_1 (1-r) + P_3\, r$$

$$P(y_0 \mid do(x_1)) = P(y_0 \mid x_1, z_0)\, P(z_0) + P(y_0 \mid x_1 z_1)\, P(z_1)$$
$$= (1-P_2)(1-r) + (1-P_4)\, r$$

$$P(y_1 \mid do(x_1)) = P_2 (1-r) + P_4\, r$$

E5

$$ACE = P(y_1 \mid do(x_1)) - P(y_1 \mid do(x_0)) =$$
$$= P_2 - P_2 r + P_4 r - P_1 + P_1 r - P_3 r$$
$$= P_2 - P_1 + r(P_1 + P_4 - P_2 - P_3)$$

$$RD = \frac{(P_2 + P_3 - P_1 - P_4)\, r + P_4 - P_3}{(1-r)\, r}$$

RD normalized with $(1-r)r$ and P_n values looks like negatively inversed.

E7

A split (800)

C_1 C_2

$(\underset{n}{300}, \underset{m}{100})$ $(\underset{n}{100}, \underset{m}{300})$

True $300 \rightarrow C_1$
False $100 \rightarrow C_2$

False $100 \rightarrow C_1$
True $300 \rightarrow C_2$.

All 800 data is divided into 400, 400
left branch is for predicting C_1 and
has 300 points from C_1 and 100 points from C_2.
right branch is also same.

So error rate $= \dfrac{FP + FN}{TP + TN + FP + FN} =$

$$= \frac{100 + 100}{800} = \frac{1}{4} = 0,25$$

all False prediction divided by all values
is error rate.

B split

C_1 C_2

(200, 400) (200, 0)

True → 200
False → 400

False → 200
True → 0.

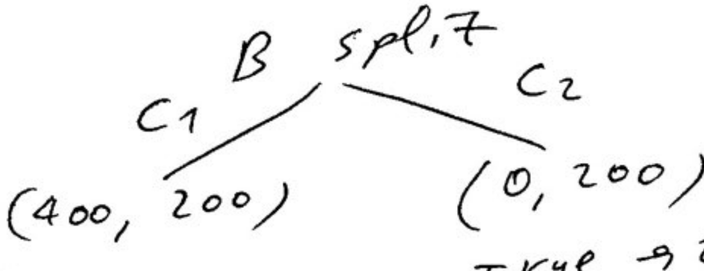
Here, I am assuming the question is
wrong. There is no way error rate is equal to A.
We have 600 points wrongly predicted in B

where A predicted only 200 wrong.

I am going to flip the numbers

(200, 400) → (400, 200)

(200, 0) → (0, 200)



True → 400
False → 200

True → 200
False → 0.

$$\text{error rate} = \frac{FP+FN}{TP+TN+FP+FN} = \frac{200}{800} = 0{,}25$$

so both error rate is same 0,25.

E 8. split A (using the table from lecture)

	C_1	C_2	
pred C_1	300 (n_{11})	100 (n_{12})	400 (n_{+1})
pred C_2	100 (n_{21})	300 (n_{22})	400 (n_{+2})
	400 (n_{1+})	400 (n_{2+})	n

$$Gini(\text{pred } C_1) = 2\ \frac{n_{11}}{n_{+1}}\ \frac{n_{12}}{n_{+1}} = 2 \cdot \frac{300 \cdot 100}{160000} = \frac{3}{8}$$

$$Gini(\text{pred } C_2) = 2 \cdot \frac{n_{21}}{n_{+2}}\ \frac{n_{22}}{n_{+2}} = 2 \cdot \frac{100 \cdot 300}{160000} = \frac{3}{8}$$

$$Gini = \frac{400}{800} \cdot \frac{3}{8} + \frac{400}{800} \cdot \frac{3}{8} = \frac{3}{8}$$

split B

	C_1	C_2	
pred(C_1)	400(n_{11})	200(n_{12})	600 (n_{+1})
pred C_2	0(n_{21})	200(n_{22})	200 (n_{+2})
	400(n_{1+})	400 (n_{2+})	n

$$Gini(pred\,c_1) = 2 \cdot \frac{n_{11}}{n_{+1}} \frac{n_{12}}{n_{+1}} = 2 \cdot \frac{\cancel{400} \cdot 200}{3\cancel{6}_9\,\cancel{0000}} = \frac{4}{9}$$

$$Gini(pred\,c_2) = 2 \frac{n_{21}}{n_{+2}} \frac{\overset{=0}{n_{22}}}{n_{+2}} = 0$$

$$Gini = \frac{n_{1+}}{n} \cdot Gini(pred\,c_1) = \frac{\cancel{400}}{\cancel{800}\,\cancel{2}} \cdot \frac{\cancel{4}^2}{9} = \frac{2}{9}$$

$$Gini(A) = \frac{3}{8} > Gini(B) = \frac{2}{9}$$

since Gini(A) has more Gini, we are more uncertain about split A and more certain about split B.

$$entropy = -\sum p_c \log p_c = -p_1 \log p_1 - p_2 \log p_2$$

A split $entropy(pred\,c_1) = -\frac{n_{11}}{n_{+1}} \log \frac{n_{11}}{n_{+1}} - \frac{n_{12}}{n_{+1}} \log \frac{n_{12}}{n_{+1}}$

$$= -\frac{300}{400} \log \frac{3}{4} - \frac{100}{400} \log \frac{1}{4} = 0,31127 + 0,5 = 0,81127$$

$$entropy(pred\,c_2) = -\frac{n_{21}}{n_{+2}} \log \frac{n_{21}}{n_{+2}} - \frac{n_{22}}{n_{+2}} \log \frac{n_{22}}{n_{+2}}$$

$$= -\frac{100}{400} \log \frac{1}{4} - \frac{300}{400} \log \frac{3}{4} = 0,81127.$$

$$entropy = \frac{n_{1+}}{n}\, entropy(pred\,c_1) + \frac{n_{2+}}{n}\, entropy(c_2)$$

$$= \frac{1}{2}(0,81127 + 0,81127) = 0,81127.$$

B split $entropy(pred\,c_1) = -\frac{4}{6} \log \frac{4}{6} - \frac{2}{6} \log \frac{2}{6} =$

$$= 0,38997 + 0,52832 = 0,91829$$

$$entropy(pred\,c_2) = 0$$

$$entropy = \frac{1}{2} \cdot 0,91829 = 0,45914.$$

$$entropy\ A = 0,81127 > entropy\ B = 0,45914$$

E9 $\quad f_s(x_s) = E_{x_c}[f(x_s, x_c)]$

$$f(x) = f_1(x_s) + f_2(x_c)$$

$$f_s(x_s) = E_{x_c}[f(x_s, x_c)] = E_{x_c}[f_1(x_s, x_c) + f_2(x_c, x_c)]$$

$$= \underbrace{E_{x_c}[f_1(x_s + x_c)]}_{f_1(x_s)} + E_{x_c}[f_2(x_c, x_c)]$$

$$= f_1(x_s) + \underbrace{E_{x_c}[f_2(x_c, x_c)]}$$

doesn't depend on X_s
so we can assume it is constant

$$= f_1(x_s) + C$$

E 10 $\quad f(x) = f_1(x_s) \cdot f_2(x_c)$

$$f_s(x_s) = E_{x_c}[f(x_s, x_c)] = E_{x_c}[f_1(x_s, x_c) f_2(x_c, x_c)]$$

$$= \underbrace{E_{x_c}[f_1(x_s, x_c)]}_{f_1(x_s)} E_{x_c}[f_2(x_c, x_c)]$$

$$= f_1(x_s) \; \underbrace{E_{x_c}[f_2(x_c, x_c)]}$$

doesn't depend on x_s
so assume constant

$$= f_1(x_s) \cdot C.$$