P.1. a) Count the No. of class C_1 and C_2 when choosing attribute X, Y, and Z respectively:

X	C_1	C_2
0	60	60
1	40	40

Y	C_1	C_2
0	40	60
1	60	40

Z	C_1	C_2
0	30	70
1	70	30

use the error rate defined,

for X: $\frac{120}{200} \times (1-\frac{1}{2}) + \frac{80}{200} \times (1-\frac{1}{2}) = 0.5$

for Y: $\frac{1}{2} \times \frac{40}{100} + \frac{1}{2} \times \frac{40}{100} = 0.4$

for Z: $\frac{1}{2} \times \frac{30}{100} + \frac{1}{2} \times \frac{30}{100} = 0.3$

there fore, for the first splitting attribute, we choose Z.

for the second split node,

when $Z=0$,

X	C_1	C_2
0	15	45
1	15	25

error rate: $\frac{15+15}{60+40} = 0.3$

Y	C_1	C_2
0	15	45
1	15	25

error rate: $\frac{15+15}{60+40} = 0.3$

when $Z=1$:

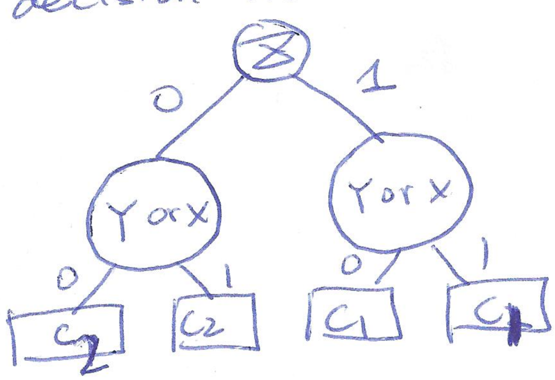
X	C_1	C_2
0	45	15
1	25	15

error rate: $\frac{15+15}{60+40} = 0.3$

Y	C_1	C_2
0	25	15
1	45	15

$\frac{15+15}{40+60} = 0.3$

thus, the decision tree would be:



b)

when $X=0$,

Y	C1	C2
0	5	~~45~~ 55
1	55	~~15~~ 5

error rate: $\frac{10}{120} = \frac{1}{12}$

Z	C1	C2
0	15	45
1	45	15

error rate: $\frac{30}{120} = \frac{1}{4}$

we use Y as the split attribute.

when $X=1$

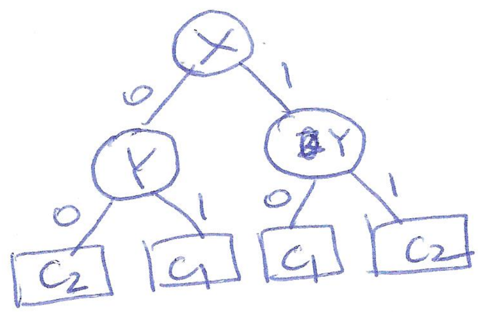
Y	C1	C2
0	35	5
1	5	35

error rate: $\frac{10}{80} = \frac{1}{8}$

Z	C1	C2
0	15	25
1	25	15

error rate: $\frac{30}{80} = \frac{3}{8}$

we use ~~Z~~ Y as the split attribute.



c) part a) use greedy heuristic, however, this method produces higher error rate than the part b) method, thus, we conclude that the greedy heuristic doesn't make sure the best solution.

P. 2 a) $S = \{s_1, s_2 \cdots s_n\}$ by sampling n times from $\{1:n\}$ with repacement, for any index m from $\{1:n\}$

$p(s_i = m) = \frac{1}{n}$.

$p(s_i \neq m) = 1 - \frac{1}{n}$.

since with replacement, so after n time's sampling

$p(m \notin S) = (1 - \frac{1}{n})^n$.

so $p(m \in S) = 1 - (1 - \frac{1}{n})^n$, while when n is large

$\lim_{n \to \infty} 1 - (1 - \frac{1}{n})^n = 1 - \frac{1}{e} = 0.632$

b). the accuray will be 50%, since the training and testing data are generated ramdomly

c). Same reason at b). the accuracy is 50%

d). accuracy:

$$\frac{1}{b} \sum_{i=1}^{b} (0.632 \times \varepsilon_i + 0.368 \times acc_s)$$

b: bootstrap samples number

ε_i: 50%

acc_s: 100%

so, overall accuracy: 0.684

this is over-estimated.

P. 3 a) conditional P:

$P(A|+)$: $\begin{cases} P(A=1|+) = \frac{3}{5} = 0.6 \\ P(A=0|+) = \frac{2}{5} = 0.4 \end{cases}$

$P(B|+)$: $\begin{cases} P(B=0|+) = \frac{4}{5} = 0.8 \\ P(B=1|+) = \frac{1}{5} = 0.2 \end{cases}$

$P(C|+)$: $\begin{cases} P(C=0|+) = \frac{1}{5} = 0.2 \\ P(C=1|+) = \frac{4}{5} = 0.8 \end{cases}$

$P(A|-)$: $\begin{cases} P(A=0|-) = \frac{3}{5} = 0.6 \\ P(A=1|-) = 0.4 \end{cases}$

$P(B|-)$: $\begin{cases} P(B=0|-) = 0.6 \\ P(B=1|-) = 0.4 \end{cases}$

$P(C|-)$: $\begin{cases} P(C=0|-) = 0 \\ P(C=1|-) = 1 \end{cases}$

b):

$$P(+|A=0,B=1,C=0) = \frac{P(A=0|+)\,P(B=1|+)\cdot P(C=0|+)\cdot P(+)}{P(A=0,B=1,C=0)}$$

$$= \frac{0.4\times 0.2\times 0.2\times 0.5}{\alpha}$$

$$= \frac{1}{\alpha}\times 0.08\times 0.1$$

$$= \frac{0.008}{\alpha}$$

$$P(-|A=0,B=1,C=0) = \frac{P(A=0|-)\,P(B=1|-)\cdot P(C=0|-)\cdot P(-)}{P(A=0,B=1,C=0)}$$

$$= \frac{0.6\times 0.6\times 0\times 0.5}{\alpha}$$

$$= 0.$$

$\Rightarrow$ the label is "+".

c).

$P(A=0|+) = \frac{2+2}{5+4} = \frac{4}{9}$

$P(A=0|-) = \frac{3+2}{5+4} = \frac{5}{9}$

$P(B=1|+) = \frac{1+2}{5+4} = \frac{3}{9}$

$P(B=1|-) = \frac{2+2}{5+4} = \frac{4}{9}$

$P(C=0|+) = \frac{1+2}{5+4} = \frac{3}{9}$

$P(C=0|-) = \frac{0+2}{5+4} = \frac{2}{9}$

d) $P(+|A=0,B=1,C=0)$

$= \frac{1}{\alpha}\times 0.5\times\frac{4}{9}\times\frac{3}{9}\times\frac{3}{9}$

$P(-|A=0,B=1,C=0)$

$= \frac{1}{\alpha}\times 0.5\times\frac{5}{9}\times\frac{4}{9}\times\frac{2}{9}$

compare, so

the label is: "−".