Q1.

Use Rainforest algorithm.

Assume C class labels.
The most memory required will be for AVC-set for the root of the tree.

Scan the database once and construct the AVC-list for each of the 50 attributes.

Size of each AVC-list = 100 C
Total size of AVC-lists = 100 C × 50 = 5000C

↓
Can fit in 512 MB memory for reasonable C.

For the computation of other AVC-sets

↳ They will be smaller
↳ Less attributes.

We can compute the AVC-set for nodes at the same level of the tree in parallel.

⇓
To reduce the no. of scans.

With such small AVC-sets per node,
we can probably fit the level in memory.

Q2.

Instead of removing an entire subtree
pruning rules allows fine-grained control by removing only specific conditions within rules.

Some parts of a subtree can still contribute useful rules.

Q3.

		R0	$\phi \to +$	
100 +		R1	A → +	4 + 1 −
400 −		R2	B → +	30 + 10 −
		R3	C → +	100 + 90 −

(I) Foil info gain.

$$FoilGain = P_1 \left[\log_2 \frac{P_1}{P_1+n_1} - \log_2 \frac{P_0}{P_0+n_0} \right]$$

R0 ⇒ $\{\phi\} \to +$

$P_0 = 100$

$n_0 = 400$

R1. $A \to +$

$P_1 = 4$

$n_1 = 1$

$$\text{Foil Gain (R1)} = P_1 \left(\log \frac{P_1}{P_1+n_1} - \log \frac{P_0}{P_0+n_0} \right)$$

$$= 4 \left(\log \frac{4}{5} - \log \frac{100}{500} \right)$$

$$= 4(-0.3219 + 2.3219)$$

$$= 8$$

R2. $P_1 = 30$

$n_1 = 10$

$$\text{Foil Gain (R2)} = 30 \left(\log \frac{30}{40} - \log \frac{100}{500} \right)$$

$$= 30(1.9069)$$

$$= 57.207$$

R3. $p_1 = 100$
$n_1 = 90$

$$\text{Foil Gain}(R3) = 100\left(\log\frac{100}{190} - \log\frac{100}{500}\right)$$

$$= 139.59$$

Best Candidate = R3
Worst " = R1

II Likelihood ratio statistic

R1.

$$ef(+) = 5 \times \frac{100}{500} = 1$$

$$ef(-) = 5 \times \frac{400}{500} = 4$$

$$\text{Likelihood ratio} = 2\left[\sum f_i \log_2 \frac{f_i}{ef_i}\right]$$

$$= 2\left[f_{(+)} \log_2 \frac{f_{(+)}}{ef_{(+)}} + f_{(-)} \log_2 \frac{f_{(-)}}{ef_{(-)}}\right]$$

$$= 2\left[4 \cdot \log_2 \frac{4}{1} + 1 \cdot \log_2 \frac{1}{4}\right]$$

$$= 2\left[8 + (-2)\right]$$

$$= 12$$

R2.

$$ef(+) = 40 \times \frac{100}{500} = 8$$

$$ef(-) = 40 \times \frac{400}{500} = 32$$

$$LR(R2) = 2\left[30 \cdot \log_2 \frac{30}{8} + 10 \cdot \log_2 \frac{10}{32}\right]$$

$$= 2\left[57.207 - 16.781\right] = 80.852$$

R3.

$$ef(+) = 190 \times \frac{100}{500} = 38$$

$$ef(-) = 190 \times \frac{400}{500} = 152$$

$$LR(R3) = 2\left[100 \log_2 \frac{100}{38} + 90 \log_2 \frac{90}{152}\right]$$

$$= 2\left[139.593 - 68.047\right]$$

$$= 143.092$$

Best ⇒ R3
Worst ⇒ R1

III Laplace Measure $= \frac{f_+ + 1}{n + K}$ K ⇒ No. of classes

$$L(R1) = \frac{4+1}{5+2} = 0.7143$$

$$L(R2) = \frac{30+1}{40+2} = 0.7381$$

$$L(R3) = \frac{100+1}{190+2} = 0.526$$

Best = R2
Worst = R3

IV m-estimate measure $k=2$ $p_+ = 0.2$ $me = \frac{f_+ + kp_+}{n+k}$

$$me(R1) = \frac{4 + 2(0.2)}{5+2} = 0.6286$$

$$me(R2) = \frac{30 + 0.4}{40+2} = 0.7238$$

$$me(R3) = \frac{100 + 0.4}{190 + 2} = 0.5229$$

Best = R2
Worst = R3

V Rule accuracy $= \frac{f_+}{n}$

$$RA(R1) = \frac{4}{5} = 0.80$$

$$RA(R2) = \frac{30}{40} = 0.75$$

$$RA(R3) = \frac{100}{190} = 0.5263$$

Best = R1
Worst = R3

Q4.

(a) $P(S|UG) = 0.15$ $\quad P(G) = 0.2$

$P(UG) = 0.8$

$P(S|G) = 0.23$

$$P(G|S) = \frac{P(S|G) \cdot P(G)}{P(S)} = \frac{0.23 \times 0.2}{0.15 \times 0.8 + 0.23 \times 0.2}$$

$$= 0.277$$

(b) $P(G) < P(UG)$

(c) $P(G|S) = 0.277$

$$P(UG|S) = \frac{P(S|UG) \cdot P(UG)}{P(S)}$$

$$= \frac{0.15 \times 0.8}{0.166} = 0.723$$

(d) $P(D|UG) = 0.10$

$P(D|G) = 0.30$

$P(D) = 0.10 \times 0.8 + 0.30 \times 0.20 = 0.14$

$P(S) = 0.166$

$P(D \cap S|G) = P(D|G) \cdot P(S|G)$

$= 0.30 \times 0.23$

$= 0.069$

$P(D \cap S|UG) = P(D|UG) \cdot P(S|UG)$

$= 0.10 \times 0.16$

$= 0.015$

$P(UG|D \cap S) = \frac{P(D \cap S|UG) \cdot P(UG)}{P(D \cap S)}$

$= \frac{0.015 \times 0.8}{P(DS)} = \frac{0.012}{P(DS)}$

$P(G|DS) = \frac{P(DS|G) \cdot P(G)}{P(DS)}$

$= \frac{0.069 \times 0.2}{P(DS)} = \frac{0.0138}{P(DS)}$

$P(G|DS) > P(UG|D \cap S)$

Q5. (a)

$P(A=1|+) = \frac{3}{5}$ $\quad P(A=1|-) = \frac{2}{5}$

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

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

$P(B=0|+) = \frac{4}{5}$ $\quad P(B=0|-) = \frac{3}{5}$

$P(C=1|+) = \frac{4}{5}$ $\quad P(C=1|-) = \frac{5}{5} = 1$

$P(C=0|+) = \frac{1}{5}$ $\quad P(C=0|-) = \frac{0}{5} = 0$

(b) Test = (A=0, B=1, C=0)

$$P(C_i|X) = \frac{P(X|C_i) \cdot P(C_i)}{P(X)}$$

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

$$= \frac{2/5 \times 1/5 \times 1/5 \times 5/10}{k}$$

$$= \frac{0.008}{k}$$

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

$$= \frac{3}{5} \cdot \frac{2}{5} \cdot 0 \cdot \frac{5}{10}$$

$$= 0$$

Prediction ⇒ + class

(c) $p = 0.5$

$m = 4$

m-estimate

$$P(A|B) = \frac{n_c + mp}{n + m}$$

n_c → # of times A ∧ B happened

n → # of times B happened.

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

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

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

$$P(B=0|+) = \frac{4+2}{9} = \frac{6}{9}$$

$$P(C=1|+) = \frac{4+2}{9} = \frac{6}{9}$$

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

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

$$P(A=0|-) = \frac{5}{9}$$

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

$$P(B=0|-) = \frac{5}{9}$$

$$P(C=1|-) = \frac{5+2}{9} = \frac{7}{9}$$

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

(d)

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

$$= \frac{\frac{4}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{5}{10}}{k}$$

$$= \frac{0.0247}{k}$$

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

$$= \frac{\frac{5}{9} \cdot \frac{4}{9} \cdot \frac{2}{9} \cdot \frac{5}{10}}{k}$$

$$= \frac{0.0274}{k}$$

Class : (-)