Quiz 4 sample solutions

Note Title 10/25/2007

Qu 1

prior:

t	W	M	F
$P(T=t)$	0.7	0.2	0.1

unnormalized posterior (prior × likelihood):

$P(T|R)$

T	R = H	R = L	R = N
W	0.56	0.07	0.07
M	0.04	0.06	0.1
F	0.03	0.06	0.01
total:	0.63	0.19	0.18

normalized posterior (columns sum to 1.0):

$P(T|R)$

T \ R	H	L	N
W	0.89	0.37	0.39
M	0.06	0.32	0.51
F	0.05	0.32	0.06

Part (a): for 0-1 loss function, just choose the most likely outcome ie. max in each column of the posterior table above, so rule is:

$$H \mapsto W, \quad L \mapsto W, \quad N \mapsto M$$

Part (b): Need to explicitly calculate expected costs, using

$$\text{cost}(\text{estimate } T \mid R) = \sum_{\text{actual } T} \underbrace{\text{cost}(\text{actual}, \text{estimate}) \times P(\text{actual } T \mid R)}_{\text{individual entries in following table}}$$

sum is bottom row of table

Expected costs

observed R:		H			L			N		
estimated T:		W	M	F	W	M	F	W	M	F
actual T	W	0	0.89	0.89	0	0.37	0.37	0	0.39	0.39
	M	0.32	0	0.25	1.58	0	1.26	2.78	0	2.22
	F	0.14	0.05	0	0.95	0.32	0	0.17	0.06	0
total expected cost:		0.46	0.94	1.14	2.53	0.68	1.63	2.94	0.44	2.61

Optimal Bayes rule chooses lowest expected cost given observation, ie.

$H \longmapsto W$, $L \longmapsto M$, $N \longmapsto M$.

Question 2

In a Bayesian network,

$$P(\text{node} \mid \text{parents}) = P(\text{node} \mid \text{parents and any subset of ancestors})$$

Therefore, we circle the following:

$P(U|G,L,A) = P(U|G,L,A,C,D)$

$P(A|C,D,G) = P(A|C,D,G,W)$

$P(G|D,W) = P(G|D)$

$P(U|G,L,A) = P(U|G,L,A,D)$

$P(U|A,G,C,D,W) = P(U|A,G)$

Question 3

Sorted values are

34 35 37 42 58 65 94 94

(a) Range is 34 → 94 i.e. 60 wide.
So each bin should be $60 \div 4 = 15$ wide.
i.e. boundaries at 34, 49, 64, 79, 94.
But it's good to allow values < 34 and > 94,
so extend leftmost and rightmost bins to $-\infty, \infty$ resp.
Final bins are:

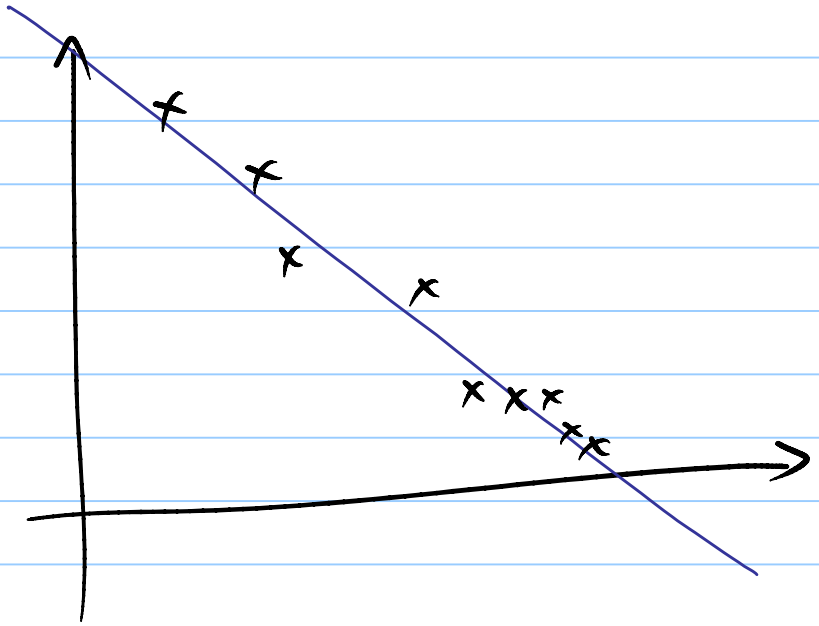
$$[-\infty, 49), [49, 64), [64, 79), [79, \infty)$$

(b) 4 bins, 8 points $\Rightarrow$ 2 points per bin.
Choose bin splits ½ way between points, and
as above extend to $\pm\infty$; obtaining

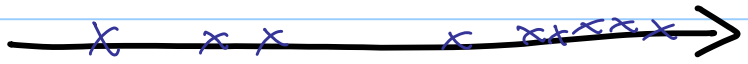
$$[-\infty, 36), [36, 50), [50, 79.5), [79.5, \infty)$$

Question 4

(i) Direction of highest variance is marked in blue:



Projecting onto the marked line results in new dataset:



(ii) Spread in direction of 1st principal cmpt is about 5 times (i.e., standard deviation) greater than spread in the perpendicular direction.
Variance is proportional to $(\text{std dev})^2$, so

ratio of variance is $5^2 : 1$ ie. $25 : 1$.

ie. first component has $\frac{25}{26} \times 100\%$ of the variance

ie. about 96%.

Question 5

i) C → BC → terminates

ii) ABC → AB → terminates.