Comp 307 A4

Part 1 :

COMP 307 Assignment

Part 1.

1). $P(X,Y) = P(Y,X) = P(X) * P(Y|X)$

$P(X=0, Y=0) = P(X=0) * P(Y=0 | X=0)$

$= 0.3 \times 0.3 = 0.09$

$P(X=0, Y=1) = P(X=0) * P(Y=1 | X=0)$

$= 0.3 \times 0.7 = 0.21$

$P(X=1, Y=0) = P(X=1) * P(Y=0 | X=1)$

$= 0.7 \times 0.8 = 0.56$

$P(X=1, Y=1) = P(X=1) * P(Y=1 | X=1)$

$= 0.7 \times 0.2 = 0.14$

X	Y	P(X, Y)
0	0	0.09
0	1	0.21
1	0	0.56
1	1	0.14

The rule that I was using is the product rule.

1

2) $P(X,Y,Z) = P(X,Y)\,P(Z|X,Y)$

As we know Z is independent from X given Y

Hence $P(Z|X,Y) = P(Z|Y)$

Therefore $P(X,Y,Z) = P(X,Y)\,P(Z|Y)$

$P(X=0, Y=0, Z=0) = P(X=0, Y=0)\,P(Z=0|Y=0)$ ⇐ using the product rule

$= 0.09 \times 0.6 = 0.054$

$P(X=1, Y=0, Z=1) = P(X=1, Y=0)\,P(Z=1|Y=0)$

$= 0.56 \times 0.4 = 0.224$

$P(X=1, Y=1, Z=0) = P(X=1, Y=1)\,P(Z=0|Y=1)$

$= 0.14 \times 0.8 = 0.112$

$P(X=1, Y=1, Z=1) = P(X=1, Y=1)\,P(Z=1|Y=1)$

$= 0.14 \times 0.2 = 0.028$

X	Y	Z	P(X, Y, Z)
0	0	0	0.054
0	0	1	0.036
0	1	0	0.168
0	1	1	0.042
1	0	0	0.336
1	0	1	0.224
1	1	0	0.112
1	1	1	0.028

3).i with the sum rule

$$P(Z=0) = P(X=0, Y=0, Z=0) + P(X=1, Y=0, Z=0) + P(X=0, Y=1, Z=0) + P(X=1, Y=1, Z=0)$$

$$= 0.054 + 0.336 + 0.168 + 0.112 = 0.67$$

$$P(X=0, Z=0) = P(X=0, Y=0, Z=0) + P(X=0, Y=1, Z=0)$$

$$= 0.054 + 0.168$$

$$= 0.222$$

(ii) If X and Z are independent $P(X, Z) = P(X) * P(Z)$

hence take $X=0, Z=0$ as an example.

a: $P(X=0, Z=0) = 0.222$

b: $P(X=0) * P(Z=0) = 0.3 \times 0.67 = 0.201$

as we can see $P(X=0, Z=0) \; != P(X=0) * P(Z=0)$

therefore we can say X and Z are not independent.

4). $$P(X=1, Y=0 \mid Z=1) = \frac{P(X=1, Y=0, Z=1)}{P(Z=1)}$$

$P(Z=1) = 1 - P(Z=0)$ (normalisation rule).

$= 1 - 0.67 = 0.33$

$$P(X=1, Y=0 \mid Z=1) = \frac{0.224}{0.33} \approx 0.6788$$

$$P(X=0 \mid Y=0, Z=0) = \frac{P(X=0, Y=0, Z=0)}{P(Y=0, Z=0)} = \frac{P(X=0, Y=0, Z=0)}{P(X=0, Y=0, Z=0) + P(X=1, Y=0, Z=0)} = \frac{0.054}{(0.054 + 0.336)} = 0.138$$

Part 2:
The code is submitted to the submission link.
Part2 q1:
P(F0 = 0| C = 0) = 0.6442953020134228
P(F0 = 0| C = 1) = 0.35570469798657717
P(F0 = 1| C = 0) = 0.3333333333333333
P(F0 = 1| C = 1) = 0.6666666666666666

P(F1 = 0| C = 0) = 0.4228187919463087
P(F1 = 0| C = 1) = 0.5771812080536913
P(F1 = 1| C = 0) = 0.4117647058823529
P(F1 = 1| C = 1) = 0.5882352941176471

P(F2 = 0| C = 0) = 0.6577181208053692
P(F2 = 0| C = 1) = 0.3422818791946309
P(F2 = 1| C = 0) = 0.5490196078431373
P(F2 = 1| C = 1) = 0.45098039215686275

P(F3 = 0| C = 0) = 0.6040268456375839
P(F3 = 0| C = 1) = 0.3959731543624161
P(F3 = 1| C = 0) = 0.39215686274509803
P(F3 = 1| C = 1) = 0.6078431372549019

P(F4 = 0| C = 0) = 0.6644295302013423
P(F4 = 0| C = 1) = 0.33557046979865773
P(F4 = 1| C = 0) = 0.5098039215686274
P(F4 = 1| C = 1) = 0.49019607843137253

P(F5 = 0| C = 0) = 0.5302013422818792
P(F5 = 0| C = 1) = 0.4697986577181208
P(F5 = 1| C = 0) = 0.6470588235294118
P(F5 = 1| C = 1) = 0.35294117647058826

P(F6 = 0| C = 0) = 0.4966442953020134
P(F6 = 0| C = 1) = 0.5033557046979866
P(F6 = 1| C = 0) = 0.21568627450980393
P(F6 = 1| C = 1) = 0.7843137254901961

P(F7 = 0| C = 0) = 0.6510067114093959

P(F7 = 0| C = 1) = 0.348993288590604
P(F7 = 1| C = 0) = 0.23529411764705882
P(F7 = 1| C = 1) = 0.7647058823529411

P(F8 = 0| C = 0) = 0.7583892617449665
P(F8 = 0| C = 1) = 0.24161073825503357
P(F8 = 1| C = 0) = 0.6666666666666666
P(F8 = 1| C = 1) = 0.3333333333333333

P(F9 = 0| C = 0) = 0.7114093959731543
P(F9 = 0| C = 1) = 0.28859060402684567
P(F9 = 1| C = 0) = 0.3333333333333333
P(F9 = 1| C = 1) = 0.6666666666666666

P(F10 = 0| C = 0) = 0.4161073825503356
P(F10 = 0| C = 1) = 0.5838926174496645
P(F10 = 1| C = 0) = 0.3333333333333333
P(F10 = 1| C = 1) = 0.6666666666666666

P(F11 = 0| C = 0) = 0.6644295302013423
P(F11 = 0| C = 1) = 0.33557046979865773
P(F11 = 1| C = 0) = 0.21568627450980393
P(F11 = 1| C = 1) = 0.7843137254901961

The sample result of the unlabeled data is shown below:

Probability for Spam is 3.020244874387394e-06,Probability for non-spam is 0.0004620049715764379.
Instance : 1 is ['1', '1', '0', '0', '1', '1', '0', '0', '0', '0', '0', '0'] non -spam
Probability for Spam is 5.5140976761978446e-05,Probability for non-spam is 4.0855635930579417e-05.
Instance : 2 is ['0', '0', '1', '1', '0', '0', '1', '1', '1', '0', '0', '1'] spam
Probability for Spam is 0.0001864445537175941,Probability for non-spam is 0.00012776774190121569.
Instance : 3 is ['1', '1', '1', '1', '1', '0', '1', '0', '0', '0', '1', '1'] spam
Probability for Spam is 5.2350911156048155e-06,Probability for non-spam is 0.0006037954762596702.
Instance : 4 is ['0', '1', '0', '0', '0', '0', '1', '0', '1', '0', '0', '0'] non -spam
Probability for Spam is 5.863981931440459e-05,Probability for non-spam is 9.134498979293801e-05.
Instance : 5 is ['1', '1', '1', '0', '1', '1', '0', '1', '0', '0', '1', '1'] non -spam
Probability for Spam is 5.5933366115278225e-05,Probability for non-spam is 4.531325026841299e-05.
Instance : 6 is ['1', '1', '1', '1', '1', '1', '0', '0', '0', '1', '1', '1'] spam
Probability for Spam is 3.43552854461566e-06,Probability for non-spam is 0.000328636441966551.
Instance : 7 is ['0', '0', '0', '0', '1', '1', '0', '1', '0', '0', '0', '0'] non -spam
Probability for Spam is 6.190253957422096e-05,Probability for non-spam is 0.00039404283148337113.
Instance : 8 is ['0', '1', '0', '1', '1', '1', '1', '0', '0', '0', '1', '1'] non -spam
Probability for Spam is 0.0001864445537175941,Probability for non-spam is 3.6936543039323476e-05.
Instance : 9 is ['1', '1', '1', '1', '1', '0', '1', '0', '0', '1', '0', '1'] spam
Probability for Spam is 2.0416855350858785e-05,Probability for non-spam is 0.000688130823577548.

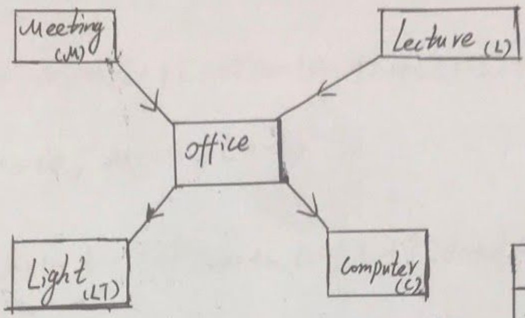
Instance : 10 is ['1', '1', '0', '0', '0', '1', '0', '1', '0', '0', '1', '0'] non -spam

(Instance 0 is non-spam
Instance 1 is spam
Instance 2 is spam
Instance 3 is non-spam
Instance 4 is non-spam
Instance 5 is spam
Instance 6 is non-spam
Instance 7 is non-spam
Instance 8 is spam
Instance 9 is non-spam
)

Part 2 q3:
In naive bayes algorithm we assumes that the attributes are conditionally independent to each other , but in real life most of times attributes do affect each other. Take this dataset as an example, we are using naive bayes based on the assumption, but a email from an invalid reply-to address maybe more likely to contain amounts of text in CAPS which means these attributes are not independent.

Part 3.

M	P(M)
+M	0.7
-M	0.3



L	P(L)
+L	0.6
-L	0.4

C	O	P(C|O)
-C	-O	0.8
-C	+O	0.2
+C	-O	0.2
+C	+O	0.8

LT	O	P(LT|O)
-LT	-O	0.98
-LT	+O	0.5
+LT	-O	0.02
+LT	+O	0.5

O	M	L	P(O|M,L)
-O	-M	-L	0.94
-O	-M	+L	0.2
-O	+M	-L	0.25
-O	+M	+L	0.05
+O	-M	-L	0.06
+O	-M	+L	0.8
+O	+M	-L	0.75
+O	+M	+L	0.95

2) $\frac{2}{2} + \frac{2}{2} + \frac{4}{2} + \frac{4}{2} + \frac{8}{2} = 10$

there are 10 free parameters.

3).

$P(L=+L, M=-m, O=+0, C=+c, LT=-Lt)$

$= P(O=+o \mid M=-m, L=+l) * P(M=-m) * P(L=+l) * P(LT=-lt \mid O=+o) * P(C=+c \mid O=+o)$

$= 0.6 * 0.3 * 0.8 * 0.8 * 0.5$

$= 0.0576$

4). $P(O=+o) = P(O=+o, M, L)$ by the sum rule

$= P(O=+o, M=+m, L=+l) + P(O=+o, M=+m, L=-l) + P(O=+o, M=-m, L=+l)$

$+ P(O=+o, M=-m, L=-l)$

$= P(O=+o \mid M=+m, L=+l) * P(M=+m, L=+l) + P(O=+o \mid M=+m, L=-l) * P(M=+m, L=-l)$

$+ P(O=+o \mid M=-m, L=+l) * P(M=-m, L=+l) + P(O=+o \mid M=-m, L=-l) * P(M=-m, L=-l)$

$= 0.95 * 0.7 * 0.6 + 0.75 * 0.7 * 0.4 + 0.8 * 0.3 * 0.6 + 0.06 * 0.3 * 0.4$

$= 0.7602 \approx 0.76$

The probability Rachel is in officie is 76% of time.

5). $P(C=+c, Lt=+LT \mid O=+o) = P(C=+c \mid O=+o) * P(LT=+lt \mid O=+o)$

$= 0.8 * 0.5$

$= 0.4$.

6). Because it is a common cause effect, there is no effect in between The effect become independent if the common cause (office) is known then LT and C become independent to each other.

5

Part 4.

Q1. i): As X-ray been given then we can say

X (XRay) is the evidence node.

The desired node we want to find the belief of is Pollution

therefore the Query variable is Pollution. (P),

Hidden Variables are Smoker (S), Cancer (C) and Dyspnoea (D)

ii): As the information been given ($P(C=t|P,S)$, $P(P=t)$ $P(S=t)$

and $P(X=t|C)$) we want to find the relationship between

X and P.

Hence we can join S and P together then we can join

C into the table and eliminate S then join X, then

eliminate C hence only P and X left.

$$(P,S \text{ joint} \Rightarrow P(P,S) \Rightarrow P(C,S,P) = P(C=t|P,S) * P(S,P) \xRightarrow{\text{eliminate } S} P(C,P))$$

$$\cancel{\Rightarrow \text{join } X \Rightarrow P(X|C,P) \xRightarrow{\text{eliminate } C} P(X|P)}).$$

$$\Rightarrow \text{join } X \Rightarrow P(X,C,P) \xRightarrow{\text{eliminate ``C''}} P(X,P)).$$

iii)

$$P(P=t|X=t) = \alpha P(P=t, X=t, S, C) = \alpha P(P=t) * P(S) * P(C|P=t,S) * P(X=t|C)$$

$$= \alpha (0.9 * 0.3 * 0.05 * 0.9 + 0.9 * 0.7 * 0.02 * 0.9 + 0.9 * 0.3 * 0.95 * 0.2 + 0.9 * 0.7 * 0.98 * 0.2)$$

$$= \alpha * 0.19827.$$

$$P(P=f \mid X=t) = \alpha \; P(P=f, X=t, S, C) = \alpha P(P=t) * P(S) * P(C \mid P=f, S) * P(X=t \mid C)$$

$$= \alpha \sum_{S,C} P(P=f) * P(S) * P(C \vee P=f, S) * P(X=t \vee C)$$

$$= \alpha (0.1 * 0.3 * 0.03 * 0.9 + 0.1 * 0.7 * 0.001 * 0.9 + 0.1 * 0.3 * 0.97 * 0.2 + 0.1 * 0.7 * 0.999 * 0.2)$$

$$= \alpha \, 0.020678$$

$$P(P=t \mid X=t) = \frac{0.19827}{(0.19827 + 0.020678)} \approx 0.9056 = 90.56\%$$

2. X is independent with D with given C because C is common cause of X and D.

X is independent with (P and S) ~~with~~ given C, because P, S are indirect cause for X

D is independent with (P and S) given C, because P, S are indirect cause for X.

7