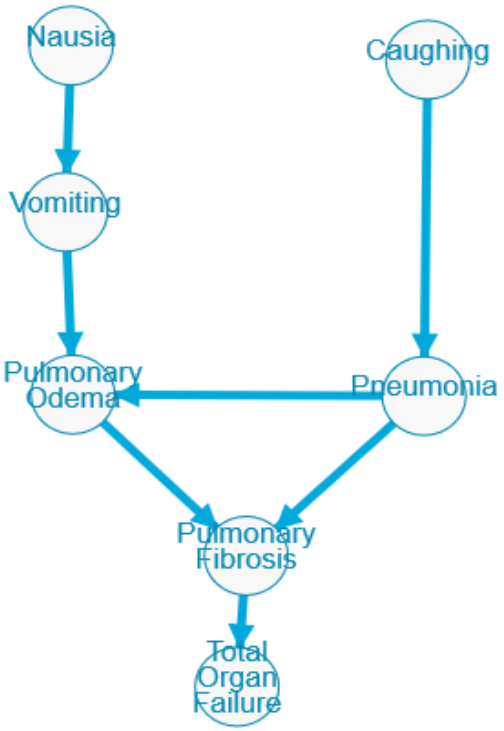
Question 1

A. Using the knowledge base below show the corresponding Bayes Net.



B. Using the ideas of conditional independence discussed in Ch. 14 of R&N, identify which variables in this network are conditionally independent of the others and under what conditions. Where appropriate, it is sucient to make shorthand statements such as X is conditionally independent of the rest of the network given Y, to save enumerating all the nodes in the network.

Answer:

Nausea is the first node in the network and does not depend on any other symptoms. So, it can be written as P(N). Vomiting is dependent on Nausea. So, it can be reduced to P(V|N). Coughing is also not dependant on any other symptoms. So, it can be written as P(C). Similarly, Pneumonia is dependent on Coughing. So, it can be written as P(Pn|C). Pulmonary Oedema is independent of Nausea and Coughing when Vomiting and Pneumonia are given. It can be written as P(PO|V, Pn).

Pulmonary Fibrosis is independent of Nausea, Vomiting and Coughing when Pulmonary Oedema and Pneumonia are given. It can thus be written as P(PF|PO,Pn).

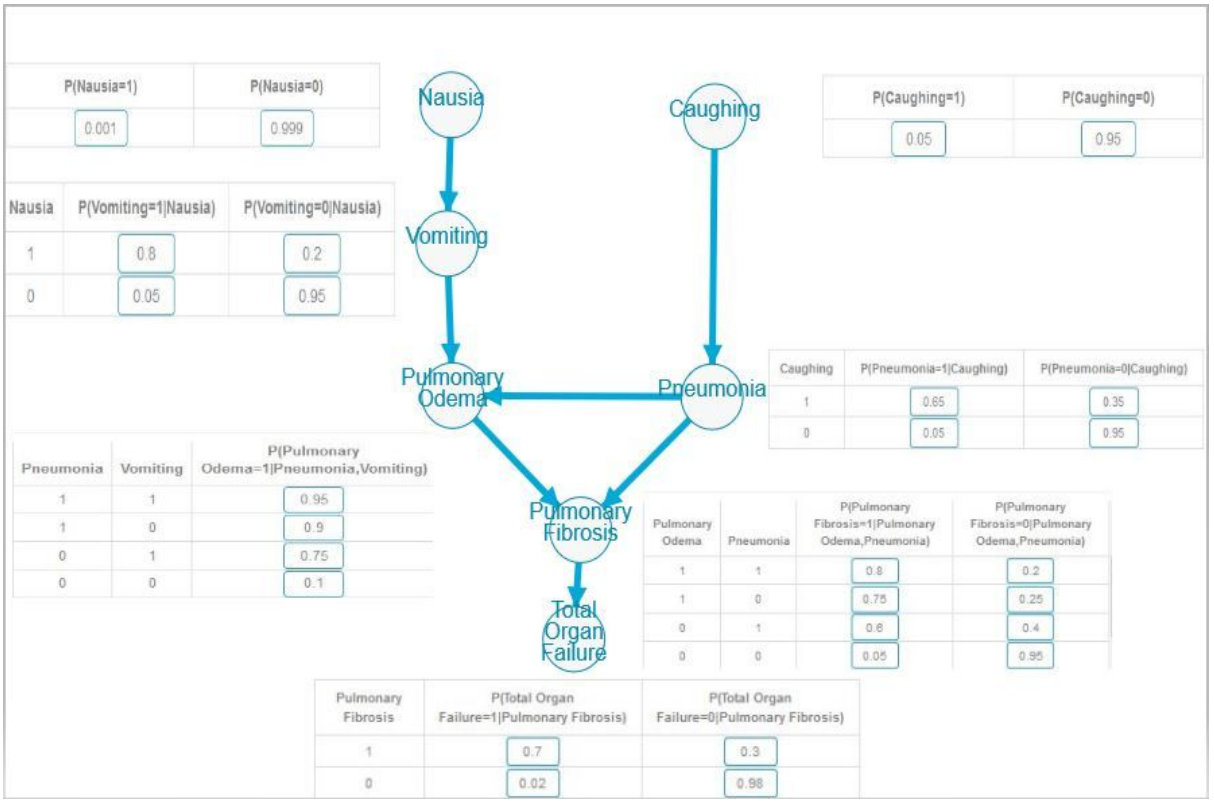
Total Organ Failure is independent of Pneumonia, Pulmonary Oedema, Vomiting, Coughing and Nausea when Pulmonary Fibrosis is known. It can thus be written as P(TOF|PF).

Therefore, the joint distribution P(Nausea, Vomiting, Coughing, Pneumonia, Pulmonary Oedema, Pulmonary Fibrosis, Total Organ Failure) is the probabilities of the following

P(Nausea)
P(Coughing)
P(Vomiting|Nausea)
P(Pneumonia|Coughing)
P(Pulmonary Oedema|Vomiting, Pneumonia)
P(Pulmonary Fibrosis|Pulmonary Oedema, Pneumonia)
P(Total Organ Failure|Pulmonary Fibrosis)

C. Use the prior and conditional probabilities provided to construct probability tables for each node of this network.

Answer: Probability table is given below



Then use exact inference to answer the following questions.
What is the probability that:

a. The symptoms include Nausea and Vomiting.
Answer:

$$P(\text{Vomiting} = T, \text{Nausea} = T) = P(\text{Vomiting} = T|\text{Nausea} = T) * P(\text{Nausea} = T)$$
$$= 0.8 * 0.07$$
$$= 0.056$$

Thus, the probability of the symptoms that include Nausea and Vomiting is **0.056**.

b. The symptom is Coughing given that it is Total Organ Failure.

Answer: TABLE CALCULATIONS

Total Organ Failure	**Pneumonia**	**Pulmonary Oedema**	**Calc**	**Answer**
T	T	T	0.7*0.8 + 0.02*0.2	0.564
T	T	F	0.7*0.6 + 0.02*0.4	0.428
T	F	T	0.7*0.75 + 0.02*0.25	0.53
T	F	F	0.7*0.05 + 0.02*0.95	0.054
F	T	T	0.3*0.8 + 0.98*0.2	0.436
F	T	F	0.3*0.6 + 0.98*0.4	0.572
F	F	T	0.3*0.75 + 0.98*0.25	0.47
F	F	F	0.3*0.05 + 0.98*0.95	0.946
Total Organ Failure	**Pneumonia**	**Vomiting**	**Calc**	**Answer**
T	T	T	0.564*0.95 + 0.428*0.05	0.5572
T	T	F	0.564*0.75 + 0.428*0.25	0.53
T	F	T	0.53*0.9 + 0.054*0.1	0.4824
T	F	F	0.53*0.1 + 0.054*0.9	0.1016
F	T	T	0.436*0.95 + 0.572*0.05	0.4428
F	T	F	0.436*0.75 + 0.572*0.25	0.47
F	F	T	0.47*0.9 + 0.946*0.1	0.5176
F	F	F	0.47*0.1 + 0.946*0.9	0.8984
Total Organ Failure	**Coughing**	**Vomiting**	**Calc**	**Answer**
T	T	T	0.5572*0.65 + 0.4824*0.35	0.531
T	T	F	0.53*0.65 + 0.1016*0.35	0.38
T	F	T	0.5572*0.05 + 0.4824*0.95	0.4861
T	F	F	0.53*0.05 + 0.1016*0.95	0.123
F	T	T	0.4428*0.65 + 0.5176 * 0.35	0.4698
F	T	F	0.47*0.65 + 0.8984 * 0.35	0.6199
F	F	T	0.4428*0.05 + 0.5176*0.95	0.5138
F	F	F	0.47*0.05 + 0.8984*0.95	0.8769
Total Organ Failure	**Coughing**	**Nausea**	**Calc**	**Answer**
T	T	T	0.531*0.8 + 0.38*0.2	0.5008
T	T	F	0.531*0.05 + 0.38*0.95	0.3875
T	F	T	0.486*0.8 + 0.123*0.2	0.4134
T	F	F	0.486*0.05 + 0.123*0.95	0.1411
F	T	T	0.4689*0.8 + 0.6199*0.2	0.4991
F	T	F	0.4689*0.05 + 0.6199*0.95	0.6123
F	F	T	0.5138*0.8 + 0.8769*0.2	0.5894
F	F	F	0.5138*0.05 + 0.8769*0.95	0.8615

	Total Organ Failure	Coughing	Calc	Answer
	T	T	0.5008*0.07 + 0.3875*0.93	0.3954
	T	F	0.4131*0.07 + 0.1411*0.93	0.1601
	F	T	0.4991*0.07 + 0.6123*0.93	0.6043
	F	F	0.5864*0.07 + 0.8615*0.93	0.8422

Explanation:
In order to calculate the probability of symptom given its a total organ failure, we first check what Total Organ failure is dependant on. It can be seen that TOF is dependant on Pulmonary Fibrosis.
But, we can see from the bayes architecture that Pulmonary Fibrosis is dependant on both Pulmonary Oedema and Pneumonia.
However, Pulmonary Oedema is dependant on both Pneumonia and Vomiting. We can also notice that Pneumonia is dependant on Coughing and Vomiting is dependant on Nausea.

This combining all the above probabilities, we get the following:
P(TOF = T, C=T)*P(C=T) = 0.3954*0.05 = 0.01977
P(TOF = T, C=F)*P(C=F) = 0.1601*0.95 = 0.1521
P(C|TF) can be determined by the following equation.
P(C=T,TOF=T)/P(TOF =T)
P(C|TF) = P(C, TF)/P(TF)
= 0.01977/0.01977 + 0.1521
= 0.115
Thus, the probability of Coughing being the symptom when Total Organ Failure occurs is **0.115.**

c. The symptom is Pulmonary Oedema.
Answer:
P(Pulmonary Oedema = T)
= P(Pulmonary Oedema = T|Vomiting, Pneumonia)
= P(Pulmonary Oedema = T|Vomiting = T, Pneumonia = T) *
P(Vomiting = T|Nausea) P(Pneumonia = T|Coughing) +
P(Pulmonary Oedema = T|Vomiting = T, Pneumonia = F)*
P(Vomiting = T|Nausea) P(Pneumonia = F|Coughing)+
P(Pulmonary Oedema = T|Vomiting = F, Pneumonia = T)*
P(Vomiting = F|Nausea) P(Pneumonia = T|Coughing) +
P(Pulmonary Oedema = T|Vomiting = F, Pneumonia = F)*
P(Vomiting = F|Nausea) P(Pneumonia = F|Coughing)

P(Vomiting=T|Nausea) =P(Vomiting|Nausea=T)P(Nausea=T)+
P(Vomiting|Nausea=F) P(Nausea = F)
= 0.8*0.07 + 0.05*0.93
= 0.1025

P(Vomiting=F|Nausea) = 1 – 0.1025 = **0.8975**
P(Pneumonia=T|Coughing) = P(Pneumonia|Coughing=T) P(Coughing=T)+
P(Pneumonia|Coughing = F) P(Coughing = F)
= 0.65*0.05 + 0.05*0.95
= **0.08**
P(Pneumonia=F|Coughing) = 1 – 0.08 = **0.92**

Now substituting the calculated values, we get
P(Pulmonary Oedema = T) =(0.95*0.1025*0.08) + (0.9*0.1025*0.92) +
(0.75*0.8975*0.08) + (0.1*0.8975*0.92)
= 0.2291

Thus, the probability of symptom being Pulmonary Oedema is **0.2291**

Question 2

Your task in this assignment is to Implement a Naive Bayes classifier and to evaluate it on the available data.

Answer:

The attached NaiveBayes.py script runs and generates the probability and results files.
Given below are the confusion matrix for the two data sets.

Confusion Matrix for Spec Heart Dataset

Pred\Expected	**Class = 1**	**Class = 0**
Class = 1	40	4
Class = 0	3	7

Confusion Matrix for Mushroom Dataset

Pred\Expected	**Class = 1**	**Class = 0**
Class = 1	431	2
Class = 0	0	696

Question 3 NLP

Answer:

Generated sequences for the author "Jane Austen"

1. mildness bear attacks tolerable tranquillity mr collins devoted morning driving gig showing country went away leaving lovely wadham arms wish
2. ascertaining upper window wore blue coat rode black horse invitation dinner soon afterwards party fine ladies issuing well known commodious
3. royal last saturday perfectly approve going regret go far see much proposed least see leisure comfort built obliged give six
4. reverting first saying astonished intimacy mr darcy sends love henry tells soon bill miss chaplin 14l pay account bill shall
5. sheet believe affectionately cass eliz austen said nothing seen continued mrs gardiner well enough return manydown fancy days fine thing
6. revenue seems happy beauty enough figure london including everybody sixty six considerably eliza expected quite enough fill cheapside cried bingley
7. replies seemed happiest memories world nothing past recollected pain lydia led voluntarily subjects sisters would alluded world think three months
8. affection family told meant london next day aunt jane see one rational days poor amos hopes skewers well left house
9. specific permission charge anything copies ebook complying rules easy may use ebook nearly purpose creation derivative works reports performances research
10. west made mind like novels really miss edgeworth egerton increase interest wish could say moment deliberation lady catherine respectable sensible

Qualitative analysis

It can be seen in sentence 8 that the model is able to correctly determine the word pairs while framing a sentence. Here, "aunt jane", "next day" are correctly mapped.

Furthermore, in sentence 2, "wore blue coat rode black horse" makes some sense as the set of words were generated depending on previously generated words. Thus, for a sequence, when the stop words are appended to the right parts of the generated sentences, we can say the markov chain has the ability to determine and generate cohesive sentences.

EXTRA CREDIT

Analysis

Inorder to see how each author chooses a set of words while writing a novel, I decided to take the works of two popular novels and check their similarity. I also decided to test out how similar inter author books turn out to be. These are my following deductions:

While testing on books from two different authors - Jane Austen and Charles Dickens

Jane Austen's book

"maria alarm every moment hard cases believed impossible true said mr bennet replied returned mrs long thing two nieces selfish"

Prob1: 1.6358147018126702**e-11**

Prob2: 6.666666666666671**e-73**

Charles Dicken's book

"peevish fine black man carry cards letters golden salver copper coloured woman linen bright handkerchief round head directions well well"

Prob1: 1.0000000000000007**e-76**

Prob2: 7.809559212858905**e-07**

When testing on the same authors books separately, I got the following:-

Book1

"newport june 1892 illustration letters jane austen ebook use anyone anywhere cost almost restrictions whatsoever may copy give away time"

Prob1: 1.156925355176084**e-06**

Prob2: 1.6666666666666652**e-25**

Book2

"promotion distribution project gutenberg literary archive foundation non profit could ideas flow rapidly time express means letters sometimes represent jenny"

Prob1: 7.480120830879859**e-28**

Prob2: 7.174097467689962**e-09**

Here, we can see that the power of probabilities are **e-25** and **e-06** while taking an alpha of **e-04** for books written by the same author. This shows that the authors choice of words are similar in the books she writes

However, while comparing the writings of Jane Austen and Charles Dickens, we can see that there is a massive difference in the power of the probabilities. In the example given above, we can see the powers to be **e-11** and **e-73** for the first sentence and **e-76** and **e-07** for the second sentence. This shows that the choice of words by the two authors are completely different.