**ECS7005P - Risk and Decision Making for Big Data and AI 2023/2024**

# Question 1

Dr Who works at MCU Clinic, where he attends to patients with STDs and UTIs. He will diagnose a patient with an STD correctly 70% of the time and will mistake it for UTI only 30% of the time. He will diagnose a patient with a UTI correctly 80% of the time and mistake it for an STD 20% of the time. A survey done by the Clinic shows that 65% of all patients that visit the Clinic have STD and 35% have UTI (not necessarily diagnosed with the correct disease). Dr Who attends to a new patient and diagnoses him with an STD. What is the probability that the new patient has a UTI?

**[20 marks]**

# Answer 1

To tackle this problem, we would use Bayes’ theorem; the formula for which is:

The steps are outlined below:

1. **First, we define the events A and B:**

* = patient has a UTI.
* = patient is diagnosed with an STD.

1. **What are our known probabilities, let us note them down**:

* is 35% - prior probability that a patient has a UTI.
* is 20% - probability of receiving an STD diagnosis if the patient already has a UTI.
* is 70% - probability of receiving an STD diagnosis if the patient already has one.
* is 65% - The prior probability of an STD in a patient.

We are trying to find the probability that the new patient has a UTI given Dr Who has diagnosed them with an STD. This can be represented as:

We can plug our known probabilities into our Bayes’ Theorem equation to find the answer. However, we must first deduce using the total probability rule.

The total probability rule is as follows: .

Calculations:

Given that the new patient has tested positive for an STD, the likelihood that they have a UTI is, therefore, roughly 0.1333, or 13.33%.

# Question 2

Create a Bayesian Network to model question 1 and answer the same question: “Dr Who attends to a new patient and diagnoses him with an STD. What is the probability that the new patient has a UTI?”. You can use any software you prefer. (If you are using Agena.ai or GeNIe please provide a picture of the model structure and a picture of the NPTs. If you are using R or Python, please provide the code and the outputs.)

**[15 marks]**

# Answer 2

I will be using Python to model question 1 and answer; What is the probability that the new patient has a UTI?

We will break down the code into steps. I have used the <https://pgmpy.org/exact_infer/ve.html> website as a reference point for my code.

Step 1: Import necessary classes.

First step is for us to import all the necessary classes which we will use to create our model.

A close-up of a text

Description automatically generated

Figure 1 – Required classes for our model.

Figure 1 depicts the classes we will be using for this model. Let’s go through them one by one.



According to the pgmpy website ‘pgmpy is a pure python implementation for Bayesian Networks with a focus on modularity and extensibility. Implementations of various algorithms for Structure Learning, Parameter Estimation, Approximate (Sampling Based) and Exact inference, and Causal Inference are available’. (pgmpy, 2023)

From pgmpy we import the BayesianNetwork class. This class is used to define the structure of our model, this includes adding nodes and edges to represent the random variables and their dependencies.



The line of code is importing the TabularCPD class from the pgmpy library. This class stands for "Tabular Conditional Probability Distribution." We will use the Tabular CPD to represent a conditional probability distribution in a table format. Each CPD in a Bayesian Network determines the probability of a variable's state given the states of its parents. It is tabular because the probabilities are organised in a table, with rows representing the variable's probability and columns representing the parent state combinations.



This line of code is used to import the VaribleElimination class. This class is used to compute the marginal distribution of a set of variables. Simply said, it allows you to answer network queries like "What is the probability of 'A' given 'B' and 'C'?"

Step 2: Define structure.



Figure 2 - Structure of our Bayesian network model

We create a Bayesian network object and assign it to the variable ‘clinic’. Inside the Bayesian network object, we have a list containing one tuple. This tuple represents a directed edge in the network. The model implies that 'Disease' has a direct influence on 'TestResult'. This represents the causal link in which the outcome of a medical test is determined by the actual sickness present.

This code lays the groundwork for modelling the relationship between diseases and test results in a clinic, with conditional probability distributions added later to quantify the relationships implied by the network's edges.

Step 3: Define conditional probability distribution for the nodes of the model.

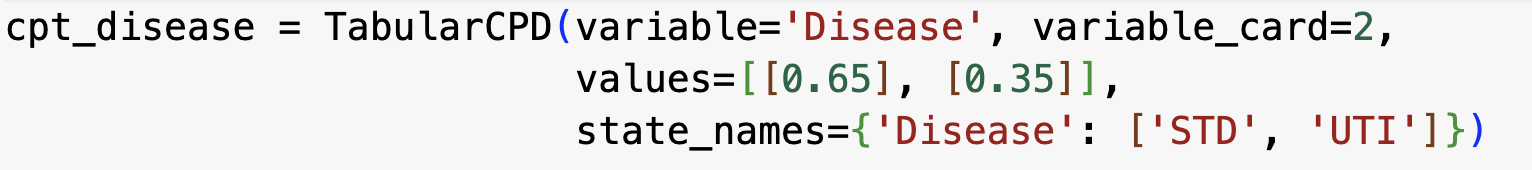


Figure 3 – Creating a TabularCPD object for disease node.

In this step, we create a **TabularCPD** object and assign it to the variable **cpt\_disease** to represent the conditional probability distribution of the 'Disease' node in our Bayesian network. This node is a categorical variable with two possible outcomes or states representing the potential diagnoses: 'STD' and 'UTI'.

The **TabularCPD** function is essential for defining a discrete probability distribution in tabular format. We define the variable of interest, 'Disease', and declare its cardinality (**variable\_card**) as 2, indicating the number of possible states the variable can be in.

The **values** argument accepts an array of probabilities representing each stage of 'Disease'. We assign a 65% probability to the state 'STD' and a 35% probability to 'UTI' based on the empirical facts presented in question 1.

Finally, the **state\_names** input is a dictionary with precise labels for each state of the 'Disease' variable. This improves the clarity of our model by mapping the integer-encoded states to their real-world counterparts: ['STD', 'UTI'].

This CPD is essential because it encodes our preconceived notions about the prevalence of each condition before examining the test results. It serves as the foundation for Bayesian inference, which updates our beliefs in the face of new data, such as a positive or negative test result.

A screenshot of a computer

Description automatically generated

Figure 4 - Creating a TabularCPD object for TestResult node.

Using the **cpt\_test** variable, we establish a TabularCPD instance for the **'TestResult'** node as we did for the 'Disease'. This defines the conditional probability distribution for the test results in the presence of an illness.

The TabularCPD function here specifies 'TestResult' as the variable, with two outcomes—**'Positive'** and **'Negative'**—represented by **variable\_card=2**. The significant difference in this CPD is the introduction of the evidence and **evidence\_card** parameters, which rely on the 'Disease' node.

The values parameter is a matrix representing the likelihood of a positive or negative test result given the current disease status. The matrix is organised so that each column corresponds to a state of the 'Disease' node (first 'STD', then 'UTI'), and each row corresponds to a state of the 'TestResult' node (first 'Positive', then 'Negative').

* For an 'STD' illness state, the chances of receiving a 'Positive' test result are 70%, and a 'Negative' one is 30%.
* For a 'UTI' disease state, the test results probability is 20% 'positive' and 80% 'negative'.

By defining **evidence=['Disease']**, we demonstrate that test result probabilities are conditional on the presence of the disease, which is consistent with how medical tests work in practice—they have varying frequencies of true positives and false positives depending on the actual condition.

**evidence\_card=[2]** indicates that the 'Disease' node, which contains the evidence, has two states, and this cardinality must correspond to the number of columns in the values matrix.

Finally, **state\_names** assign relevant labels to the 'TestResult' and 'Disease' states, making the CPD more understandable.

This CPD enables us to represent the diagnostic test's performance precisely, indicating its ability to detect diseases (sensitivity) and confirm their absence (specificity), a critical feature of probabilistic diagnostic reasoning in the Bayesian paradigm.

A close up of words

Description automatically generated

Figure 5 – Add the CPD to our Bayesian network.

This code snippet adds the previously created Conditional Probability Distributions (CPDs) to the Bayesian Network and checks its validity.

**clinic.add\_cpds(cpt\_disease, cpt\_test)**: This method adds the CPDs to the Bayesian Network instance named clinic. cpt\_disease is the CPD that specifies the probabilities for the 'Disease' node, and cpt\_test is the CPD for the 'TestResult' node that also considers the 'Disease' node as its parent, reflecting the conditional probabilities of test results given the disease.

A screenshot of a computer code

Description automatically generated

This code snippet is where the model is put into action to perform probabilistic inference.

**inference = VariableElimination(clinic):** This line produces an inference object called **inference** using the **VariableElimination** algorithm. The object is initialised with the Bayesian Network model clinic, including the network structure and CPDs we have created above. VariableElimination is a widely used exact inference procedure in probabilistic graphical models. It systematically eliminates variables by summing them out of the joint distribution.

**query\_result = inference.query(variables=['Disease'], evidence={'TestResult': 'Positive'}):**Using the inference object's query method, you ask the network for the probability distribution of the **'Disease'** node based on some evidence about the **'TestResult'** node. The evidence shows that the test result is **'Positive'**, which is what you would anticipate from an STD test. The query will return the **'Disease'** posterior probability distribution, which considers the prior probabilities and the chances of detecting the evidence (a positive test result).

**probability\_UTI\_given\_positive = query\_result.values[1]:** This line calculates the likelihood of having a UTI using the query result, assuming that **'STD'** is indexed at **0** and **'UTI'** is indexed at **1** in the 'Disease' node. The indices depend on the order you specified in the **state\_names** argument when you created the CPDs. The values attribute of the result contains the probability for each state of the **'Disease'** node based on the evidence.

The result is 0.1333 which is 13.33% and this matches the answer we have deduced in question 1.

# Question 3 and Answers

1. Are the below interpretations correct or wrong and why?
   1. Consider the relationship between outdoor temperature and ice cream sales. A negative correlation coefficient between these variables would indicate that that colder temperatures lead to lower demand for ice cream.

This interpretation is **wrong**. A negative correlation indicates that while variable A increases, variable B decreases, and vice versa. A negative correlation coefficient between outdoor temperature and ice cream would suggest that colder temperature relates to increased ice cream sales and not reduced ice cream sales.

* 1. Suppose there's a positive correlation between the number of umbrellas sold and the incidence of flu cases in a city. This correlation suggests that buying more umbrellas can causes more people to get the flu.

Correlation does not imply causation. This interpretation is **wrong**. Correlation indicates the relationship between two variable it does not imply any cause or effect. An explanation for this relationship witnessed could be a cofounding variable like weather.

A cofounding variable is one that affect both independent and dependent variable causing a false association. Rainy or chilly weather may encourage people to buy umbrellas, but it may also produce conditions conducive to the development of the flu.

* 1. You set a significance level (alpha) of 0.05 before conducting your hypothesis test. After analysing the data, you obtain a p-value of 0.06. An interpretation would be that the observed result is statistically significant at the 0.05 level, and you reject the null hypothesis accordingly.

This interpretation is **wrong**. An appropriate interpretation is that the observed result is not statistically significant at the 0.05 level. As a result, you will fail to reject the null hypothesis.

The significance level (alpha) is a threshold to determine if the observed result is statistically significant or not. A statistically significant observed result is one where the p-value is less than or equal to the alpha value set. If the p-value is greater than the alpha value set, the observed result is not statistically significant.

A p-value of 0.06 indicates a 6% chance of detecting your data or something more extreme if the null hypothesis is correct. Because this likelihood exceeds your selected threshold of 5% (0.05), you lack adequate evidence to infer that the effect observed in your data could not have occurred by chance alone. Thus, the null hypothesis remains valid.

* 1. Suppose you conduct a study on the relationship between exercise and weight loss and obtain a p-value of 0.25 (significance level is 0.05). This non-significant result indicated that there is no relationship between exercise and weight loss.

The interpretation is **incorrect**. The study's p-value of 0.25 implies that there is insufficient evidence to determine a statistically significant association between exercise and weight loss.

The absence of a statistically significant association does not imply that one does not exist. There are many factors which could contribute towards a study’s failure to detect a relationship, this includes a small sample size, insufficient variability in the data, measurement error etc.

A p-value greater than the significance level implies that the results are not convincing enough to reject the null hypothesis of no influence in the context of a statistical test; it does not prove the null hypothesis correct.

* 1. Suppose a study calculates a 95% confidence interval for the average weight loss due to a new diet to be [5 kg, 7 kg]. This means that if we were to repeat the study many times, approximately 95% of the resulting confidence intervals would contain the true average weight loss.

This interpretation is **correct**. This interpretation is correct. A 95% confidence interval indicates that if the same population is sampled 100 times and we calculate a confidence interval for each sample, around 95 of 100 confidence intervals will contain the true value – the genuine average weight reduction owing to the new diet.

1. After a recession period, the Annualised Growth Rate is found to follow a normal distribution with mean -3 and variance 10. What is the probability that the growth rate will be: (You can use any software you prefer.)
   1. between 1 and 3
      1. The probability that the growth rate will be between 1 and 3 is approximately 7.41%.
   2. greater than -5
      1. The probability that the growth rate will be greater than -5 is approximately 73.65%.

A screenshot of a computer code

Description automatically generated

Does not ask to give explanation so just screen shot given.

[20 marks]

# Question 4

A study into the effectiveness of two different treatment procedures (A and B) has been reported on patients suffering with a particular Disease which can be either chronic or acute. The results were:

1. Which treatment has the better success rate overall? [5 marks]
2. Which treatment is better for patients with the acute form of the Disease? [5 marks]
3. Which treatment is better for patients with the chronic form of the Disease? [5 marks]
4. State what is Simpson Paradox and why this is an example of it?   
   [5 marks]

# Answers

1. To determine the better success rate overall, we assess the total success rate for each treatment:
   1. For Treatment A: (16+30/16+4+30+50) × 100 = 46%
   2. For Treatment B: (62+6/62+18+6+14) × 100 = 68%

Treatment B had a better success rate overall.

1. To determine which treatment is better for patients with the acute form of the disease, we assess the acute success rate for each treatment:
   1. For Treatment A: (16/16+4) × 100 = 80%
   2. For Treatment B: (62/62+18) × 100 = 77.5%

Treatment A had a better success rate for acute patients; therefore, we assume that it is the better treatment for patients with the acute form of the disease.

1. To determine which treatment is better for patients with the chronic form of the disease, we assess the chronic success rate for each treatment:
   1. For Treatment A: (30/30+50) × 100 = 37.5%
   2. For Treatment B: (6/20) × 100 = 30%

Treatment A had a better success rate for chronic patients; therefore, we assume that it is the better treatment for patients with the chronic form of the disease.

1. Regarding Simpson's Paradox, this scenario is a classic example. When the data are analysed independently, Treatment A appears to be more effective for both the acute and chronic versions of the condition; yet, when the data are combined, Treatment B appears to have a greater overall success rate. This demonstrates how combined data can produce a different impression than separate group analysis, emphasising the significance of stratification in statistical analysis. The paradox demonstrates how aggregate data can sometimes lead to the opposite result. The paradox is usually caused by an unaccounted-for variable or confounding factor that significantly impacts the outcome. This variable can influence the groups differently, resulting in the observed paradox.

# Question 5

Sandra attends a circus and is presented with the following rules of Shell Game.

* A bet cost £5.
* You can only pick 1 shell from 3 shells.
* If the ball is under your chosen shell, you win £20.

1. What is the total expected utility of Sandra decision to play the Shell Game? [5 marks]
2. If Sandra values the excitement of playing the game as having a utility of £7. What is the total expected utility of Sandra decision to play the Shell Game? [5 marks]
3. Should Sandra play the Shell Game (if considering the excitement of playing)? [5 marks]
   1. Total expected utility of not playing = 0
4. Use any software to create an influence diagram based on the information above (without considering the excitement of playing). Should Sandra play the Shell Game? [10 marks]

# Answer

1. The total expected utility in this case can be calculated looking at the monetary gain or loss.   
     
   The probability of a win (choosing a shell with a ball underneath) is 1/3 and a loss (choosing a shell which does not have a ball underneath) is 2/3.   
     
   Net gain for a win is £15 as you win £20 pounds, but you must pay £5 to play.   
     
   The expected value is EV = (Probability of Winning × Winnings) + (Probability of Losing × Loss)   
     
   EV = (1/3 × £15) + (2/3 × −£5)   
   EV = 1.666666… or 1.67 rounded to 2dp.

As a result, under the assumptions we make, Sandra's decision to play the Shell Game has a total expected utility equal to the game's expected value of £1.67. This means Sandra receives an average utility value of £1.67 for each game she plays.

1. To determine the expected utility of Sandra's decision to play the Shell Game, we must account for both the expected monetary value and the enjoyment she obtains from the game.

Based on our prior calculations, we determined that the game's projected monetary value is £1.67. If Sandra values the excitement of playing the game at £7, add that amount to the predicted monetary value to calculate the total expected utility.

As a result, the total expected utility is the sum of the predicted monetary value and the utility of excitement:

Total Expected Utility = Utility from Excitement + Expected Monetary Value

= £7 + £1.67 = £8.67.

1. The preceding computation indicates that the overall expected utility of playing the game, including monetary expectation and thrill value, is around £8.67.

The whole expected utility of not playing is stated as zero. This is because Sandra neither gains nor loses money, and she does not feel the excitement of the game.

* Total estimated utility of playing: around £8.67.
* The total estimated utility of not playing is £0.

Sandra should play the Shell Game based on her excitement value and the game's projected utility (£8.67), which exceeds the utility of not participating (0). This decision is based on the increased value she places on the excitement of the game, making the total experience of playing more desirable to her than not participating.

1. Diagram is below:

A diagram of a game

Description automatically generated

A screenshot of a computer

Description automatically generatedEven without considering the excitement of playing Sandra should play as she still has a positive utility of £1.67 each time she plays. The utility of not playing is 0.