Artificial Intelligence 2

Solutions

Main Summer Examinations 2024

Note

Answer ALL questions. Each question will be marked out of 20. The paper will be marked out of 60, which will be rescaled to a mark out of 100.

Question 1

(a) Explain the following concepts in logistic regression:

(i) Dependent (response) variable **[3 marks]**

(ii) Odds **[3 marks]**

(b) Sudden arrhythmic death syndrome, or SADS, is when someone dies suddenly following a cardiac arrest with no obvious cause. Several potential risk factors, such as age, blood pressure and cigarette smoking are used as independent variables for investigation. The dependent variable (response) is 2-year incidence of SADS in females without prior coronary heart disease. We use the data to fit a logistic regression and the model coefficients are shown in Table 1, with the model interception of -15.3 (not shown in the Table)

θ_i	Risk Factor	Model Coefficient θ
θ_1	Blood Pressure (mm Hg)	0.0019
θ_2	Weight (% of study mean)	-0.0060
θ_3	Cholesterol (mg/100 mL)	0.0056
θ_4	Glucose (mg/100 mL)	0.0066
θ_5	Smoking (cigarettes/day)	0.0069
θ_6	Hematocrit (%)	0.111
θ_7	Vital capacity (centiliters)	-0.0098
θ_8	Age (years)	0.0686

Table 1: Logistic Regression model coefficients

(i) Based on the values in Table 1, give brief interpretations of how age (θ_8) and vital capacity (θ_7) affect the estimated risk of women SADS, respectively. **[6 marks]**

(ii) Using logistic regression, predict the probability of SADS for a 50 year old woman with systolic blood pressure of 120 mmHg, a relative weight of 100% a cholesterol level of 250 mg/100mL, a glucose level of 100 mg/100mL, a hematocrit of 40%, and a vital capacity of 450 centiliters who smokes 10 cigarettes per day. **[8 marks]**

Model answer / LOs / Creativity:

(a) Explain the following concepts in logistic regression:

(i) An dependent (response) variable: In a logistic regression the dependent, or known as response variable Y, is a binary random variable that indicates whether or not the sample have a particular outcome, say cancer (1) or not (0).

(ii) Odds: The odds are the ratio of the probability that an outcome occurs to the probability that the outcome does not occur.

(b) Sudden arrhythmic death syndrome

(i) The age coefficient $\theta_8 = 0.0686$ means that after holding all the other factors fixed (weight, smoking status, cholesterol levels, etc.) for every extra year of age the **log odds** of a woman's risk of sudden death goes up by .0686. For the vital capacity coefficient, $\theta_7 = -0.0098$ means that all else equal, for every extra centiliter of vital capacity, a woman's **log odds** of sudden death goes down by .0098. The answer should include the quantitative change of log odds or probabilities of sudden death, otherwise, deduct 3 marks.

(ii) Plugging in the given values will give us the log odds of SADS for such a woman:

$$\begin{aligned}\log(odds) &= -15.3 + .0019(120) - .006(100) + .0056(250) + .0066(100) \\ &+ .0069(10) + .111(40) - .0098(450) + .0686(50) \\ &= -10.083\end{aligned}$$

The actual probability :

$$P(Y = 1) = \frac{1}{1 + \exp(10.083)} = 0.00004178$$

Learning outcomes: 1). Conceptually understand frameworks for consistent treatment of uncertainty in AI; and 2) Understand principles of inference and fitting to data in AI models under uncertainty. The creative part is (b).

Question 2

You are a General Practitioner (GP) who specialised in three ailments, Flu , Cold and Allergy, denoted as F, C and A, respectively. Denote possible symptoms as V (fever), H (cough), R (runny nose) and I (itchy eyes). Each ailment has the following symptoms:

- Flu: V, H and R
- Cold: H and R
- Allergy: H, R and I

We assume all the ailments and symptoms are binary random variables.

(a) Draw a Bayesian network to diagnose these three ailments based on patients' symptoms. **[6 marks]**

(b) Write down the joint probability distribution represented by this Bayesian network. **[4 marks]**

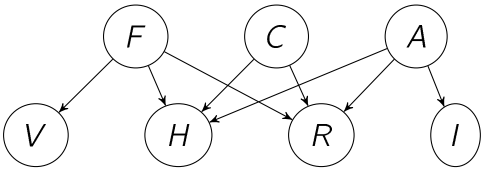
(c) How many parameters are required to describe this joint probability distribution? Show your working. **[4 marks]**

(d) Suppose a patient complains that he/she has two symptoms, cough ($H = 1$) and runny nose ($R = 1$). What diagnosis can you make just based on this information? What new physiological and background information you could collect for a more accurate diagnosis? Explain why the new information can help. **[6 marks]**

Model answer / LOs / Creativity:

Learning outcomes: The creative part is (b).

(a) The following draw is acceptable:



(b) The joint probability distribution represented by the Bayesian network:

$$\begin{aligned}&P(F, C, A, V, H, R, I)\\ &= P(F)P(C)P(A)P(V|F)P(H|F, C, A)P(R|F, C, A)P(I|A)\end{aligned} \tag{1}$$

(c) Total number of parameter is 23, but the student should show the following working:

Conditional Probability	number of parameters
$P(F)$	1
$P(C)$	1
$P(A)$	1
$P(V\|F)$	2
$P(H\|F, C, A)$	8
$P(R\|F, C, A)$	8
$P(I\|A)$	2

(d) Coughing and has a runny nose means $H = 1$ and $R = 1$, which cannot use explaining away to help us to pin point any disease since all competing causes are possible. However, in this situation, we can make diagnosis based on the the probability of those ailments, e.g., $P(F)$, $P(C)$, and $P(A)$. (2 marks) We can collect the temperature, and if the patient has fever, then he has Flu. This is called explaining away (2 marks). But if the patient has no fever, then we can use the patient's personal health history, e.g., whether the patient has allergy before, which will increase the probability of allergy. Also, if we know it's flu season then we can increase the probability of flu. If the pollen level is high, then we can increase the probability of allergy. The student can get 2 marks if the discussion is reasonable. (2 marks)

Learning outcomes: 1). Conceptually understand frameworks for consistent treatment of uncertainty in AI; and 2) Understand principles of inference and fitting to data in AI models under uncertainty.

Question 3 Constraint Satisfaction Problems (CSPs)

(a) Please briefly explain the reason of not using breadth first search (BFS) to solve CSPs. **[6 marks]**

(b) Suppose we have 3 professors who need to deliver 6 modules (M_1-M_6) in computer science. Each module needs to be delivered during $9:00-10:00$ in the morning of one or several days.

- M_1: Artificial Intelligence needs to be delivered on both Monday and Thursday;
- M_2: JAVA needs to be delivered on three days, Tuesday, Wednesday and Thursday;
- M_3: Evolutionary Computation needs to be delivered on Wednesday;
- M_4: Data Structure needs to be delivered on both Wednesday and Friday;
- M_5: Software Engineering needs to be delivered on both Tuesday and Friday;
- M_6: Network needs to be delivered on Thursday.

Note each professor can deliver some modules. Here are the details

- Professor A can deliver M_1, M_2, M_3, M_4;
- Professor B can deliver all the modules (i.e., $M_1, M_2, M_3, M_4, M_5, M_6$);
- Professor C can deliver M_2, M_5, M_6.

Note that a professor can only deliver at most one module in one day (e.g, Professor A cannot deliver both M_1 and M_2 since both modules need to be delivered on Thursday). Each module can only be delivered by a professor. Your task is to assign professors to modules so that each module can be successfully delivered.

(i) Formulate this problem as a CSP problem. Specifically, you need to give variables, domains and constraints. We use variable M_i to represent the ith module ($i \in \{1, \ldots, 6\}$) and Professor A delivering M_i can be expressed as $M_i = A$. Domain of each variable (denoted by $d(M_i)$) is a subset of $\{A, B, C\}$. Constraints should be specified formally (i.e., using the form of $M_i \neq M_j$). **[6 marks]**

(ii) Draw the constraint graph associated with your CSP. **[3 marks]**

(iii) Use **backtracking search with forward checking and ordering** to find a solution of this problem. Let us assume that tie of variables is broken numerically (i.e., in the order of $M_1, M_2, \ldots, M_6$). The value of a variable should be considered alphabetically (i.e., in the order of A, B, C). Give the order of states to be visited. **[5 marks]**

Model answer / LOs / Creativity:

(a) The reason that breadth first search cannot be used to solve CSPs is that a solution of a CSP is always in the bottom layer of the search tree, and BFS needs to traverse all the nodes of the tree (except some in the last level) to find it.

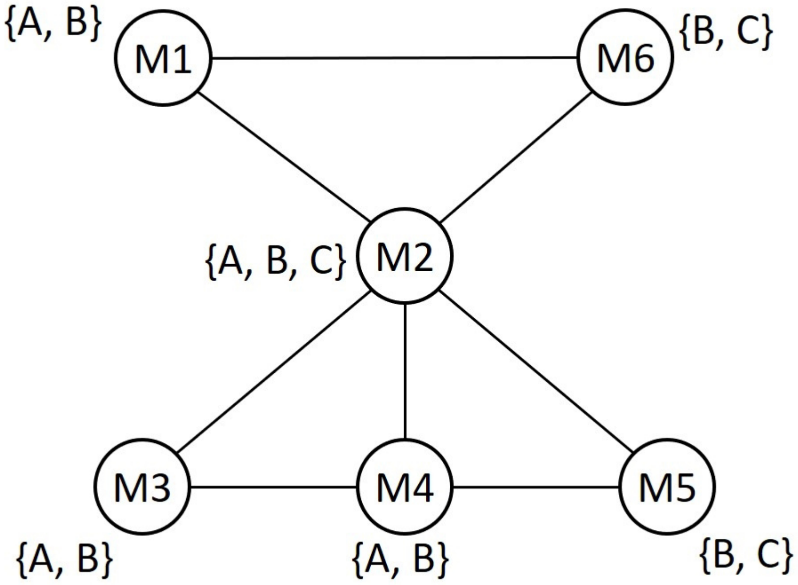
(b) (i) The variables are $M_1, M_2, M_3, M_4, M_5, M_6$.
The domains are as follows, we use $d(M)$ to denote the domain of M

$$d(M_1) = \{A, B\}, \quad d(M_2) = \{A, B, C\}, \quad d(M_3) = \{A, B\}$$
$$d(M_4) = \{A, B\}, \quad d(M_5) = \{B, C\}, \quad d(M_6) = \{B, C\}.$$

The constraints are

$$M_1 \neq M_2 \quad M_1 \neq M_6 \quad M_2 \neq M_3 \quad M_2 \neq M_4$$
$$M_2 \neq M_5 \quad M_2 \neq M_6 \quad M_3 \neq M_4 \quad M_4 \neq M_5$$

(ii) The constraint graph is



(iii) The order of variables to be visited is

$$M_1 = A, M_2 = B, M_3 = A, M_2 = C, M_5 = B, M_4 = A, M_3 = B, M_6 = B$$

The final solution is

$$(M_1, M_2, M_3, M_4, M_5, M_6) = (A, C, B, A, B, B)$$

Learning outcomes: Formulating a practical problem as CSP and then solving it. The creative parts are (b)-(ii) and (b)-(iii).