Note: The names of the TA(s) who graded the question are written by the Question number.

Monsoon 2023 - CSE 643 Artificial Intelligence
End-Sem Solutions - Dec. 08, 2023

Maximum score: 20

1. (5 points) **True/False** with justification. [Saloni]

(a) Depth-first search always expands at least as many nodes as A* search with an admissible heuristic.

(b) $P(a|b \wedge a) = 1$.

(c) In any Bayes net, the parents of a single child are always conditionally independent of each other given the child.

(d) A* is not used in robotics because the percepts, states and actions are continuous.

(e) Does logistic regression for binary classification produce a linear decision boundary? {*Hint*: Try to visualize (maybe in 1D or 2D) what the decision boundary would be like, then generalize to nD.} **Correction announced**. Logistic Regression for binary classification always produces a linear decision boundary.

Solution:

(a) **False**. A lucky DFS might expand exactly d nodes to reach the goal. A* largely dominates any graph-search algorithm that is guaranteed to find optimal solutions.

(b) **True**. The "first principles" needed here are the definition of conditional probability, $P(X|Y) = P(X \wedge Y)/P(Y)$, and the definitions of the logical connectives. It is not enough to say that if $B \wedge A$ is "given" then A must be true! From the definition of conditional probability, and the fact that conjunction is commutative, associative, and idempotent, we have

$$P(A|B \wedge A) = \frac{P(A \wedge (B \wedge A))}{P(B \wedge A)} = \frac{P(B \wedge A)}{P(B \wedge A)} = 1$$

(c) **False**. The parents will usually be conditionally dependent given the child, as one cause explains away another.

(d) **False**. A* is often used in robotics after discretizing the states, percepts and actions.

(e) **True**. The logistic regression model only has a nonlinear component for the probability mapping. The decision boundary is linear and is given as the hyperplane orthogonal to the weight vector $\mathbf{w}$ with the offset as the bias term b.

2. (5 points) **Answer in 1-2 sentences**. [Rohit and Nidhi]

(a) What impact does node ordering have on the construction of a Bayesian Network?

(b) What is the criterion for an admissible heuristic in A*?

(c) Write the mathematical expression for Naive Bayes and state the assumption it makes.

(d) Explain Transitivity & Continuity constraints for rational preferences with their expressions.

(e) What are the four different inference problems in temporal models?

Solution:

(a) The choice of node ordering influences the number of parameters required to describe the joint distribution in the Bayesian Network, which may effect the computational cost of inference as well.
Rubric: 1 if no. of parameters is stated, or both no. of parameters & computation is stated; 0.5 if only computation is stated; 0 otherwise (e.g., if only the graph topology is stated as the only impact).

(b) An admissible heuristic for A* *always* underestimates the true distance between the current node and the goal.

(c) The Naive Bayes model assumes that the effects (or features/observations) are conditionally independent of each other, given the cause (or class / state). The mathematical form is given by:

$$P(C, E_1, E_2, \ldots, E_n) = P(C) \prod_{i=1}^{n} P(E_i|C)$$

where C is the *Cause* and $E_i,\ i = 1, \ldots, n$ are the *Effects* indexed by i.

(d) **Transitivity**: Given any three lotteries, if an agent prefers A to B and prefers B to C, then the agent must prefer A to C, i.e.,

$$(A \succ B) \wedge (B \succ C) \implies (A \succ C)$$

Continuity: If some lottery B is between A and C in preference, then there is some probability p for which the rational agent will be indifferent between getting B for sure and the lottery that yields A with probability p and C with probability $1 - p$.

$$A \succ B \succ C \implies \exists p : [p, A; 1 - p, C] \sim B$$

(e) The following are the main inference tasks in temporal models:

(i) **Filtering**. Filtering or state estimation is the task of computing the belief state, i.e., $P(\mathbf{X}_t|\mathbf{e}_{1:t})$.

(ii) **Prediction**. This is the task of computing the posterior distribution over the future state, given all evidence to date, i.e., $P(\mathbf{X}_{t+k}|\mathbf{e}_{i:t})$.

(iii) **Smoothing**. This is the task of computing the posterior distribution over a past state, given all evidence up to the present, i.e., $P(\mathbf{X}_k|\mathbf{e}_{1:t})$ for some k such that $0 \leq k < t$.

(iv) **Most likely explanation**. Given a sequence of observations, the sequence of states that is most likely to have generated those observations, i.e., $\arg\max_{\mathbf{x}_{1:t}} P(\mathbf{x}_{1:t}|\mathbf{e}_{1:t})$.
Rubric: 1 point if names are listed correctly with at least two correct explanations. 0.5 points if names are listed correctly with no or incorrect explanations. 0.5 if at least three names are correctly listed, and 0 if two or more names are incorrect.

[Drishya] (2 points)

3. (3+2=5 points) ~~(a)~~ Given the Bayesian Network in Fig. 1b, (i) compute $P(Cloudy|WetGrass = true)$ and (ii) verify whether *Sprinkler* is conditionally independent of *Rain* given *Cloud*. Show work for full credit. (1 point)

Solution: 3(a)

(i) C - Cloudy, W - WetGram, R - Rain, S - Sprinkler, t - true, f - false

$$P(C|W=t) = \alpha \sum_{r,s} P(C, r, s, w)$$

$$= \alpha \sum_{r,s} P(C)\, P(s|C)\, P(r|C)\, P(w|s,r)$$

$$= \alpha\, P(C) \sum_{r,s} P(s|C)\, P(r|C)\, P(w|s,r)$$

$$P(C=t|W=t) = \alpha\, P(C=t) \left[\sum_r P(r|C) \sum_s P(s|C)\, P(W=t|s,r) \right]$$

$$= \alpha \times 0.5 \Big[P(r=t|C=t) \big(P(s=t|C=t) P(w=t|s=t, r=t) + P(s=f|C=t) P(w=t|s=f, r=t) \big)$$

$$+ P(r=f|C=t) \big(P(s=t|C=t) P(w=t|s=t, r=f) + P(s=f|C=t) P(w=t|s=f, r=f) \big) \Big]$$

$$= \alpha \times 0.5 \Big[0.8 \times \big(0.1 \times 0.99 + 0.9 \times 0.9 \big) + 0.2 \times \big(0.1 \times 0.9 + 0.9 \times 0.0 \big) \Big]$$

$$= \alpha \times 0.5 \big(0.8 \times (0.099 + 0.81) + 0.2 (0.09) \big)$$

$$P(C=t|W=t) = \alpha \times 0.5 \big(0.8 \times (0.909) + 0.018 \big) = 0.3726\, \alpha$$

$$P(C=f|W=t) = \alpha \left[\sum_r P(r|C) \sum_s P(s|C)\, P(W=t|s,r) \right]$$

$$= \alpha \times 0.5 \Big[P(r=t|C=f) \big(P(s=t|C=f) P(w=t|s=t, r=t) + P(s=f|C=f) P(w=t|s=f, r=t) \big)$$

$$+ P(r=f|C=f) \big(P(s=t|C=f) P(w=t|s=t, r=f) + P(s=f|C=f) P(w=t|s=f, r=f) \big) \Big]$$

$$= \alpha \times 0.5 \left[0.2 \times (0.5 \times 0.99 + 0.5 \times 0.9) + 0.8 \times (0.5 \times 0.9 + 0.5 \times 0.0) \right]$$

$$= \alpha \times 0.5 (0.2 \times 0.5 \times 0.99 + 0.2 \times 0.5 \times 0.9 + 0.8 \times 0.5 \times 0.9 + 0.8 \times 0.5 \times 0.0)$$

$$= \alpha \times 0.5 (0.1 \times 0.99 + 0.1 \times 0.9 + 0.4 \times 0.9 + 0)$$

$$= \alpha \times 0.5 (0.099 + 0.09 + 0.36) = \alpha \times 0.5 \times 0.549 = 0.2745\alpha$$

~~$= \alpha \times 0.5 (0.054) = \alpha \times 0.027$~~

$P(C=f \mid W=t) = \alpha \times 0.02745$ ~~$\alpha \times 0.027$~~

$\Rightarrow P(C \mid W=t) = \langle 0.576, 0.424 \rangle$ ~~$\langle 0.932, 0.068 \rangle$~~

(ii) To show the S is conditionally independent of R, given C, the simplest strategy would be to establish.

$$P(S,R \mid C) = P(S \mid C) P(R \mid C)$$

$$P(S,R \mid C) = \sum_W P(S,R,C,W)$$

For $C=t$

$$P(S=t, R=t \mid C=t) = \alpha \sum_W P(C) P(S \mid C) P(r \mid C) P(W \mid S,r)$$

$$= \alpha P(C) P(S \mid C) P(r \mid C) \sum_W P(W \mid S,r)$$

$\sum_W P(W \mid S,r)$ will always be 1 $\forall S, r$

$$= \alpha \; 0.5 \times 0.1 \times 0.8 (0.99 + 0.01)$$

$$P(S=t, R=t \mid C=t) = \alpha \times 0.05 \times 0.8 = 0.04\alpha$$

$$P(S=t, R=f \mid C=t) = \alpha \times 0.5 \times 0.1 \times 0.2 = 0.01\alpha$$

$$P(S=f, R=t \mid C=t) = \alpha \times 0.5 \times 0.9 \times 0.8 = 0.36\alpha$$

$$P(S=f, R=f \mid C=t) = \alpha \times 0.5 \times 0.9 \times 0.2 = 0.09\alpha$$

$\alpha = \frac{1}{0.5}$

$P(S=t, R=t \mid C=f) = \beta \times 0.5 \times 0.5 \times 0.2 = 0.05\beta$

$P(S=t, R=f \mid C=f) = \beta \times 0.5 \times 0.5 \times 0.8 = 0.2\beta$

$P(S=f, R=t \mid C=f) = \beta \times 0.5 \times 0.5 \times 0.2 = 0.05\beta$

$P(S=f, R=f \mid C=f) = \beta \times 0.5 \times 0.5 \times 0.8 = 0.2\beta$

$\beta = \frac{1}{0.5}$

S	R	$P(S,R \mid C=f)$	$P(S \mid C=f) \times P(R \mid C=f)$
t	t	0.1	$0.5 \times 0.2 = 0.1$
t	f	0.4	$0.5 \times 0.8 = 0.4$
f	t	0.1	$0.5 \times 0.2 = 0.1$
f	f	0.4	$0.5 \times 0.8 = 0.4$

S	R	$P(S,R \mid C=t)$	$P(S \mid C=t) \times P(R \mid C=t)$
t	t	0.08	$0.1 \times 0.8 = 0.08$
t	f	0.02	$0.1 \times 0.2 = 0.02$
f	t	0.72	$0.9 \times 0.8 = 0.72$
f	f	0.18	$0.9 \times 0.2 = 0.18$

Since for all combinations of S, R, & C, we have

$$P(S,R \mid C) = P(S \mid C)\, P(R \mid C)$$

we have verified that S & R are conditionally independent, given C.

	Corona (t)		~Corona (f)		Total
	travel (t)	~travel (f)	travel (t)	~travel (f)	
(m) mild	0.15	0.046	0.06	0.044	0.3
(s) severe	0.22	0.025	0.1	0.055	0.4
(d) death	0.055	0.004	0.24	0.001	0.3
Total	0.425	0.075	0.4	0.1	1

P(travel) = 0.825
P(~travel) = 0.175

[Suryaksa]

3 (b) Given the joint distribution table from our Corona and other diseases example, construct a Bayesian Network and write the corresponding CPTs. Use travel as root (no parent).

Solution: 3(b)

3 variables → Corona (C), Travel (T), Severity (S)

t - true
f - false

$T \in \{t, f\}$ $C \in \{t, f\}$ $S \in \{m, s, d\}$

The BN is as shown below

$P(T=t) = 0.825$

$P(C=t|T=t) = 0.425/(0.425+0.4)$

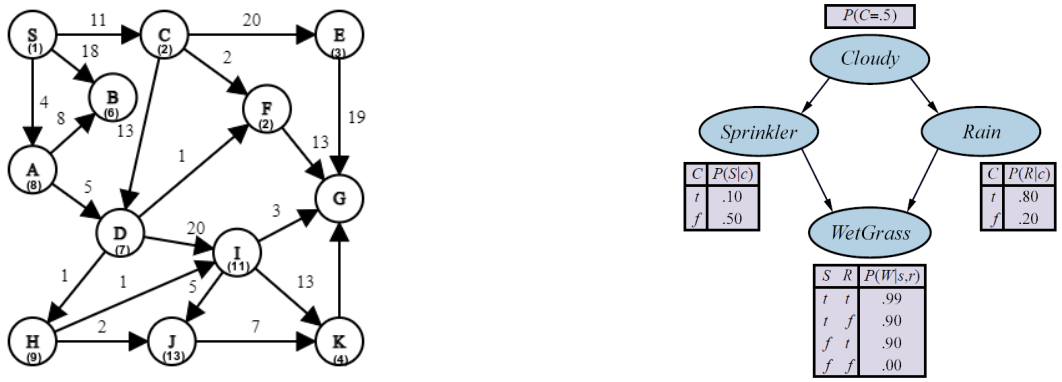
$P(C=t|T=f) = 0.075/(0.075+0.1)$

	S=m	S=s	S=d
P(S|C=t)	0.196/0.5	0.245/0.5	0.059/0.5
P(S|C=f)	0.05/0.5	0.155/0.5	0.241/0.5

4. (3+2=5 points) [Ritisha and Swatantra]

(a) Given the graph in Fig.1a, let S be the source node, and G be the goal node. Find the cost-effective path from S to G using (1) A* and (2) Best-first search. Show the steps and explain which yields a lower-cost path and why. **Note**: The numerals written on edges represents the cost between nodes. The numerals written on the nodes represents the heuristic value. Heuristic for Goal Node G is 0. Cost from K to G is 6.

(b) How is the output of the logistic regression model interpreted? If the input is $\mathbf{x}$ and the weights are w_j, write the output of logistic regression as a function of $\mathbf{x}$ and calculate its derivative w.r.t. w_j.



(a) Search Graph (b) Bayesian Network

Figure 1

Solution:

(a) Refer to the solution of HW-2 theory question.
Note: The solution for Best-first search will depend on what evaluation function has been used, e.g., if the heuristic is used as the evaluation function, the greedy best-first search will be the algorithm to consider.

(b) The output of logistic regression is interpreted as the probability p of the target variable y being *true*, i.e., $p = f(\mathbf{x})$. Let $z = \mathbf{w}^\top \mathbf{x}$. Assuming the bias term is included in $\mathbf{w}$ and the corresponding constant term is included in $\mathbf{x}$.

$$f(\mathbf{x};\mathbf{w}) = p = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-\mathbf{w}^\top\mathbf{x}}} = \frac{1}{1+e^{-\sum_j w_j x_j}}$$

$$\begin{aligned}
\frac{\partial p}{\partial w_j} &= \left(\frac{\partial p}{\partial z}\right)\cdot\left(\frac{\partial z}{\partial w_j}\right) \\
&= \left(\frac{-1}{(1+e^{-z})^2}\cdot(e^{-z})(-1)\right)\cdot(x_j) \\
&= \left(\frac{e^{-z}}{(1+e^{-z})}\cdot\frac{1}{(1+e^{-z})}\right)\cdot(x_j) \\
&= x_j\left(\frac{e^{-\mathbf{w}^\top\mathbf{x}}}{(1+e^{-\mathbf{w}^\top\mathbf{x}})}\cdot\frac{1}{(1+e^{-\mathbf{w}^\top\mathbf{x}})}\right) = x_j\cdot p\cdot(1-p).
\end{aligned}$$

5. (**Extra Credit** 3+2=5 points) [Hans]

(a) Let $\mathbf{x} \in \mathbb{R}^2$ and is distributed as a standard normal distribution ($\mathcal{N}(0, \mathbf{I}_{2\times 2})$). Let $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{b}$, where $\mathbf{A}$ and $\boldsymbol{b}$ are matrices and vectors of appropriate sizes. What would be the distribution of $\mathbf{y}$? What would be the mean and the covariance of $\mathbf{y}$? Show derivation to get full credit.

(b) What are the two key differences between a temporal model with a Bayesian Network structure and Kalman Filters?

Solution:

(a) As given, the mean and covariance of $\mathbf{x}$ are:

$$\begin{aligned} \mathbf{E}[\mathbf{x}] &= \boldsymbol{\mu}_{\mathbf{x}} = [0, 0]^\top \\ Cov(\mathbf{x}) &= \mathbf{E}[(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^\top] = \mathbf{I}_{2\times 2} \end{aligned} \tag{1}$$

Since $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{b}$ is an affine (linear function ($\mathbf{A}$)+ constant($\boldsymbol{b}$)) transformation of the random vector $\mathbf{x}$, the distribution of $\mathbf{y}$, its distribution will also be Gaussian. The mean and covariance will be given as below:

$$\begin{aligned} \mathbf{E}[\mathbf{y}] &= \mathbf{E}[\mathbf{A}\mathbf{x} + \boldsymbol{b}] = \mathbf{A}\mathbf{E}[\mathbf{x}] + \mathbf{E}[\boldsymbol{b}] = \mathbf{A} \cdot \boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{b} = \boldsymbol{b} \\ Cov(\mathbf{y}) = \mathbf{E}[(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})^\top] &= \mathbf{E}[\mathbf{y}\mathbf{y}^\top - \mathbf{y}\boldsymbol{\mu}_{\mathbf{y}}^\top - \boldsymbol{\mu}_{\mathbf{y}}\mathbf{y}^\top + \boldsymbol{\mu}_{\mathbf{y}}\boldsymbol{\mu}_{\mathbf{y}}^\top] \\ &= \mathbf{E}[\mathbf{y}\mathbf{y}^\top - \mathbf{y}\boldsymbol{b}^\top - \boldsymbol{b}\mathbf{y}^\top + \boldsymbol{b}\boldsymbol{b}^\top] \\ &= \mathbf{E}[\mathbf{y}\mathbf{y}^\top] - \mathbf{E}[\mathbf{y}\boldsymbol{b}^\top] - \mathbf{E}[\boldsymbol{b}\mathbf{y}^\top] + \mathbf{E}[\boldsymbol{b}\boldsymbol{b}^\top] \\ &= \mathbf{E}[\mathbf{y}\mathbf{y}^\top] - \mathbf{E}[\mathbf{y}]\boldsymbol{b}^\top - \boldsymbol{b}\mathbf{E}[\mathbf{y}]^\top + \mathbf{E}[\boldsymbol{b}\boldsymbol{b}^\top] \\ &= \mathbf{E}[\mathbf{y}\mathbf{y}^\top] - \boldsymbol{b}\boldsymbol{b}^\top - \boldsymbol{b}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{b}^\top \\ &= \mathbf{E}[(\mathbf{A}\mathbf{x} + \boldsymbol{b})(\mathbf{A}\mathbf{x} + \boldsymbol{b})^\top] - \boldsymbol{b}\boldsymbol{b}^\top \\ &= \mathbf{E}[\mathbf{A}\mathbf{x}\mathbf{x}^\top\mathbf{A}^\top + \boldsymbol{b}\mathbf{x}^\top\mathbf{A}^\top + \mathbf{A}\mathbf{x}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{b}^\top] - \boldsymbol{b}\boldsymbol{b}^\top \\ &= \mathbf{E}[\mathbf{A}\mathbf{x}\mathbf{x}^\top\mathbf{A}^\top] + \mathbf{E}[\boldsymbol{b}\mathbf{x}^\top\mathbf{A}^\top] + \mathbf{E}[\mathbf{A}\mathbf{x}\boldsymbol{b}^\top] + \mathbf{E}[\boldsymbol{b}\boldsymbol{b}^\top] - \boldsymbol{b}\boldsymbol{b}^\top \\ &= \mathbf{E}[\mathbf{A}\mathbf{x}\mathbf{x}^\top\mathbf{A}^\top] + \boldsymbol{b}\mathbf{E}[\mathbf{x}]^\top\mathbf{A}^\top + \mathbf{A}\mathbf{E}[\mathbf{x}]\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{b}^\top - \boldsymbol{b}\boldsymbol{b}^\top \\ &= \mathbf{A}\mathbf{E}[\mathbf{x}\mathbf{x}^\top]\mathbf{A}^\top + \boldsymbol{b} \cdot 0 \cdot \mathbf{A}^\top + \mathbf{A} \cdot 0 \cdot \boldsymbol{b}^\top \\ Cov(\mathbf{y}) &= \mathbf{A}\mathbf{A}^\top \end{aligned}$$

(b) (1) For Bayesian Networks, the state and sensor/observation variables need to be discrete, while Kalman filters use continuous random variables to model the state and observations.
(2) Bayesian Networks use a conditional probability table to model the necessary conditional probabilities, while Kalman filters make use of Gaussian distributions to model the randomness of the state and observation variables.