FIRST SEMESTER EXAMINATIONS 2021/22

Advanced Artificial Intelligence

TIME ALLOWED : One Hour and Thirty Minutes

INSTRUCTIONS TO CANDIDATES

NAME OF CANDIDATE .... SEAT NO ....

USUAL SIGNATURE ....

READ THE FOLLOWING CAREFULLY:

1. Each of the following questions comprise 5 statements, from which you should select one appropriate answer by placing ticks in the appropriate boxes.
2. The exam mark is based on the overall number of correctly answered questions. The more questions you answer correctly the higher your mark, incorrectly answered questions do not count against you.
3. Enter your name and examination number IN PENCIL on the computer answer sheet according to the instructions on that sheet.
4. When you have completed this exam paper, read the instructions on the computer answer sheet carefully and transfer your answers from the exam paper. Use a HB pencil to mark the computer answer sheet and if you change your mind be sure to erase the mark you have made. You may then mark the alternative answer.
5. At the end of the examination, be absolutely sure to hand in BOTH this exam paper AND the computer answer sheet.
6. Calculators are permitted.

THIS PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM

Part 1: Basic Knowledge

1. Which learning task best suits the following description: given a set of training instances $\{(\mathrm{x}^{(1)}, y^{(1)}), ..., (\mathrm{x}^{(n)}, y^{(n)})\}$ of an unknown target function f, where $\mathrm{x}^{(i)}$ is the feature vector and $y^{(i)}$ is the label for $i \in \{1, ..., n\}$, it outputs a model f that divides the training instances into several groups such that those instances within a group are similar and those instances across groups are dissimilar.

☐ **A.** Anomaly Detection

☐ **B.** Supervised learning

☐ **C.** Reinforcement learning

☐ **D.** Dimensionality Reduction

■ **E.** none of the above

2. Which of the following statements is incorrect?

☐ **A.** In an active learning scheme, the learner can actively select instances for training.

☐ **B.** The hypothesis space of a learning algorithm is the function space $\mathcal{H}$ such that each element in $\mathcal{H}$ is a possible model the learning algorithm will end up with.

☐ **C.** dimensionality reduction is to find a model $f \in \mathcal{H}$ that represents each instance $\mathbf{x}$ with a lower-dimension feature vector while still preserving key properties of $\mathbf{x}$.

■ **D.** Training, test, and validation datasets cannot overlap to make sure that a training algorithm is not biased.

☐ **E.** Anomaly detection is to learn model $f \in \mathcal{H}$ that represents "normal" instances, so that the model can later be used to determine whether a new data x looks normal or anomalous.

	Intelligence = Low	Intelligence = High
Grade = A	0.07	0.21
Grade = B	0.28	0.09
Grade = C	0.27	0.08

Table 1: Joint probability for student grade and intelligence

3. According to the table in Table 1 about two random variables *Intelligence* and *Grade*, please select a value for x to make the following expression hold (rounded to 2 decimal places):

$$P(Grade = C \mid Intelligence = Low) = x$$

☐ **A.** x = 0.27
■ **B.** x = 0.44
☐ **C.** x = 0.65
☐ **D.** x = 0.35
☐ **E.** x = 0.28

4. Compute the following conditional probability according to the table in Table 1 (rounded to 2 decimal places)

$$P(Intelligence = low \mid Grade \in \{B, C\}) =$$

☐ **A.** 0.54
☐ **B.** 0.33
☐ **C.** 0.77
■ **D.** 0.76
☐ **E.** 0.65

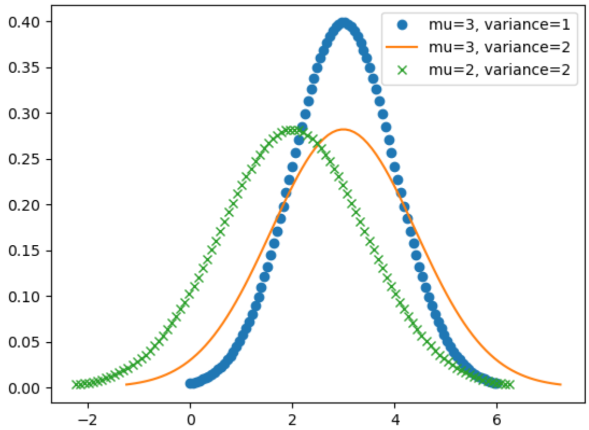


Figure 1: Diagram of three Gaussian distributions with different means and variances

5. Assume we have three functions f_1, f_2 and f_3, representing the three Gaussian distributions as in Figure 1. Specifically, we have $f_1 = \mathcal{N}(3, 1^2)$, $f_2 = \mathcal{N}(3, 2^2)$, and $f_3 = \mathcal{N}(2, 2^2)$. Which of the following statements is correct.

■ **A.** $\max_x f_1(x) \approx 0.4$

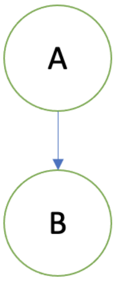
☐ **B.** $\max_x f_1(x) = \max_x f_2(x)$

☐ **C.** $\forall x : f_1(x) \geq f_2(x)$

☐ **D.** $\forall x : f_2(x) = f_3(x + 1)$

☐ **E.** $\arg\max_x f_2(x) = \arg\max_x f_3(x)$

P(A)	
a^0	a^1
0.7	0.3



P(B|A)		
	b^0	b^1
a^0	0.1	0.9
a^1	0.4	0.6

Figure 2: Probabilistic Graph of Diseases (A) and Symptom (B)

6. Use the information provided in Figure 2 to compute the following joint probability

$$P(A = a^1, B = b^0) =$$

■ **A.** 0.12
☐ **B.** 0.36
☐ **C.** 0.3
☐ **D.** 0.48
☐ **E.** 0.16

7. Use the information provided in Figure 2 to compute the following expression

$$\max_{A,B} P(A, B) =$$

☐ **A.** 0.48
☐ **B.** a^1, b^1
☐ **C.** 0.5
☐ **D.** a^0, b^1
■ **E.** 0.63

8. Use the information provided in Figure 2 to compute the following maximum a posteriori expression

$$MAP(A, B) =$$

- ☐ **A.** 0.36
- ☐ **B.** a^1, b^1
- ☐ **C.** 0.5
- ■ **D.** a^0, b^1
- ☐ **E.** a^1, b^1

9. Understanding simple numpy command.
Assume that $a = np.arange(100).reshape((5, 20))$. Then $a[:, 2:5].T.shape =$

- ■ **A.** (3,5)
- ☐ **B.** 10
- ☐ **C.** (2,5)
- ☐ **D.** (5,2)
- ☐ **E.** 5

10. Let $x = (5, 2, 3, -4)$ be a vector. Then its L^2 norm $||x||_2 =$

- ☐ **A.** 10
- ☐ **B.** $\sqrt{30}$
- ☐ **C.** 4
- ☐ **D.** $\sqrt{22}$
- ■ **E.** $\sqrt{54}$

11. Let $x = (1, 5, 3, -4)$ be a vector. Then its L^1 norm $||x||_1 =$

- ☐ **A.** 5
- ☐ **B.** $\sqrt{30}$
- ☐ **C.** 10
- ■ **D.** 13
- ☐ **E.** 6

Part 2: Simple Learning Models

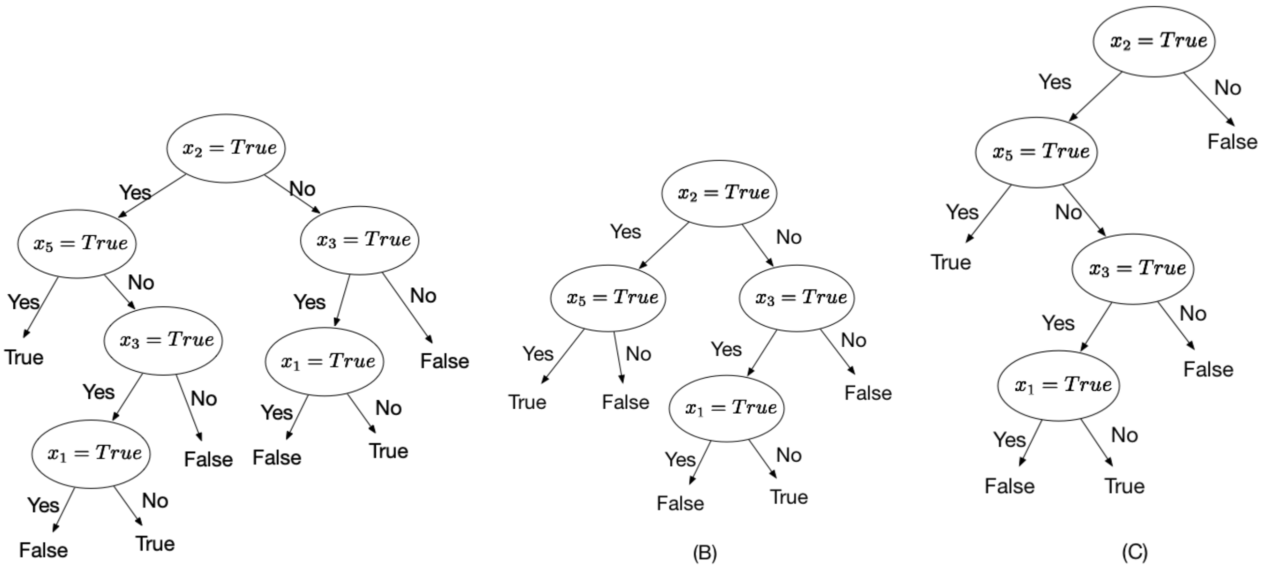


Figure 3: Decision Trees

12. Which decision trees in Figure 3 can represent the Boolean formula $(x_2 \wedge x_5) \vee (x_3 \wedge \neg x_1)$?

☐ **A.** A and B

☐ **B.** B

☐ **C.** C

☐ **D.** A and C

■ **E.** none of the above answers is correct

13. Figure 4 gives an example dataset D about Iris flowers. Please indicate which of the following expressions is used to compute its entropy $H_D(Y)$, where Y is the random variable for labelling:

☐ **A.** $\frac{8}{14}\log_2(\frac{8}{14}) + \frac{6}{14}\log_2(\frac{6}{14})$

☐ **B.** $-\frac{4}{14}\log_2(\frac{4}{14}) - \frac{4}{14}\log_2(\frac{4}{14}) - \frac{4}{14}\log_2(\frac{4}{14}) - \frac{2}{14}\log_2(\frac{2}{14})$

☐ **C.** $\frac{4}{14}\log_2(\frac{4}{14}) + \frac{10}{14}\log_2(\frac{10}{14})$

■ **D.** $-\frac{5}{14}\log_2(\frac{5}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) - \frac{4}{14}\log_2(\frac{4}{14})$

☐ **E.** none of the above

Index	Sepal Length	Sepal Width	Pedal Length	Pedal Width	Iris Class
1	5	3	1	0.5	0
2	4	3	1	0.5	0
3	4	3	1	0.5	0
4	5	3	1	0.5	0
5	4	3	1	0.5	0
6	7	3	4	1	1
7	6	3	4	1	1
8	6	3	4	1	1
9	4	2	3	1	1
10	6	3	6	2	2
11	5	2	5	2	2
12	7	3	5	2	2
13	5	2	5	2	2
14	6	2	5	1	2

Figure 4: Dataset for Iris Flowers

14. Figure 4 gives an example dataset D about Iris flower. Please compute the information gain of splitting over the feature Sepal Length, i.e., $\textbf{InfoGain}(D, SepalLength) = H_D(Y) - H_D(Y \mid SepalLength)$ (over two decimal places):

☐ **A.** 0.98

☐ **B.** -0.18

☐ **C.** 0.46

■ **D.** 0.63

☐ **E.** 0.05

15. Figure 4 gives an example dataset D about Iris flowers. Please compute the information gain of splitting over the feature Sepal Width $\textbf{InfoGain}(D, SepalWidth) = H_D(Y) - H_D(Y \mid SepalWidth)$ (rounded to two decimal places):

■ **A.** 0.28

☐ **B.** -0.19

☐ **C.** 0.12

☐ **D.** 0.15

☐ **E.** 0.13

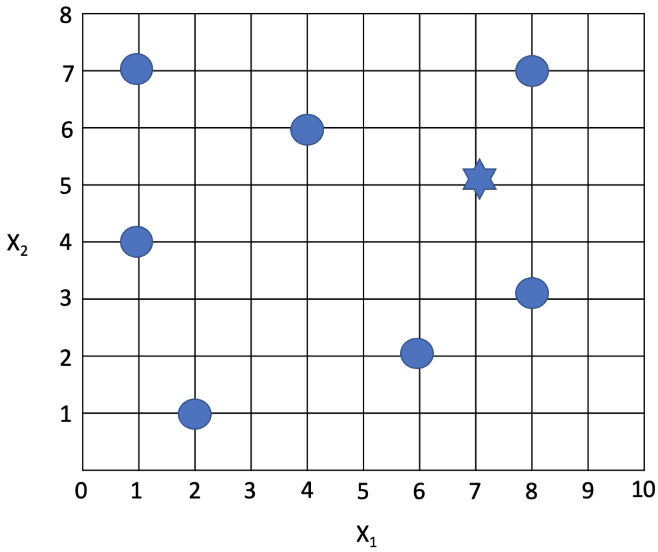


Figure 5: A set of two-dimensional input samples

16. Assume that, as shown in Figure 5, we have a set of training instances with two features X_1 and X_2:

$$\{(1,7),(1,4),(2,1),(4,6),(6,1.9),(8,3),(8,7)\}$$

such that

- the instance $(1,7)$ is labeled with value 0,
- the instances $(1,4),(2,1)$ are labeled with value 1,
- the four instances $(6,2),(8,3)$ are labeled with value 2, and
- the instance $(4,6),(8,7)$ are labeled with value 3.

Now, we have a new input $(7.1, 5.1)$. Please indicate which of the following statement is correct, according to the L^∞ distance.

☐ **A.** Both (4, 6) and (6,2) are *not* considered for the 3-nn (3-nearest neighbor) classification

☐ **B.** Both (8,3) and (8,7) are *not* considered for the 3-nn (3-nearest neighbor) classification

☐ **C.** (8,3) is *not* considered for the 3-nn (3-nearest neighbor) classification

☐ **D.** Both (6, 2) and (8,3) are *not* considered for the 3-nn (3-nearest neighbor) classification

■ **E.** Both (2,1) and (6,2) are *not* considered for the 3-nn (3-nearest neighbor) classification

17. Continue with the above. Now, for new input $(7.1, 5.1)$, please compute its regression result for the 3-nn (3-nearest neighbor) regression, according to the L^∞ distance.

☐ **A.** 5/3

☐ **B.** 6.1/3

■ **C.** 8/3

☐ **D.** 2.1

☐ **E.** 7/3

18. Please select the correct statement from the following:

☐ **A.** Validation dataset is another terminology for training dataset

■ **B.** Validation dataset is often used for early stopping

☐ **C.** Validation dataset is part of the test dataset

☐ **D.** Validation dataset cannot be used for regularization

☐ **E.** Test dataset can be overlapped with the training dataset

		Actual Class	
		positive	negative
Predicted Class	positive	a	b
	negative	c	d

Figure 6: A confusion matrix for the two-class problem

19. Assume a two-class problem where each instance is classified as either 1 (positive) or -1 (negative). We have a training dataset of 1,000 instances, such that 500 of them are labeled as 1 and 500 of them are labeled as -1. After training, we apply the trained model to classify the 1,000 instances and find that 900 instances are classified correctly. Moreover, we know that, 500 instances are classified as 1 and, within the 500 instances, 50 instances are actually labeled as -1. Please indicate which numbers should be filled in to (*A*, *B*, *C*, *D*) in Figure 6.

☐ **A.** (450, 50, 150, 350)

☐ **B.** (450, 50, 100, 400)

☐ **C.** (400, 100, 100, 400)

■ **D.** (450, 50, 50, 450)

☐ **E.** (400, 50, 150, 400)

20. Continue with the above question. Please compute the error rate of the trained model.

☐ **A.** 50/1000

☐ **B.** 150/1000

■ **C.** 100/1000

☐ **D.** 150/2000

☐ **E.** 150/850

21. Given a set of 4 training data $\{((0,0),0),(0,1),0),(1,0),0),(1,1),1)\}$, where each instance has two features X_1 and X_2 and a label y, linear regression is used to find a linear function $f(\mathbf{x}) = w^T\mathbf{x} + b$. Please indicate which of the following statement is correct:

■ **A.** if $f(\mathbf{x}) = X_1 + X_2 - 1.5$ then the training accuracy is 1.0.

☐ **B.** if $f(\mathbf{x}) = 2X_1 + X_2 - 1.5$ then the training accuracy is 1.0.

☐ **C.** if $f(\mathbf{x}) = 2X_1 + X_2 + 1.5$ then the training accuracy is 1.0.

☐ **D.** if $f(\mathbf{x}) = X_1 + X_2 + 0.9$ then the training accuracy is 1.0.

☐ **E.** none of the above statement is correct.

22. Let $f(X) = 3X_1^2 + 4e^{sin(X_2)} + 5e^{-X_3}$ be a function, where X_1, X_2 and X_3 are three variables. Please indicate which of the following gradient expressions is correct:

■ **A.** $\nabla_X f(X) = (6X_1, 4cos(X_2)e^{sin(X_2)}, -5e^{-X_3})$

☐ **B.** $\nabla_X f(X) = (6, 4cos(X_2)e^{sin(X_2)}, -5e^{-X_3})$

☐ **C.** $\nabla_X f(X) = (6X_1, 4sin(X_2)e^{sin(X_2)}, -5e^{-X_3})$

☐ **D.** $\nabla_X f(X) = (6X_1, 4e^{sin(X_2)}, 5e^{-X_3})$

☐ **E.** $\nabla_X f(X) = (6X_1, 4cos(X_2)e^{sin(X_2)}, 5e^{-X_3})$

Index	Sepal Length	Sepal Width	Pedal Length	Pedal Width	Iris Class
1	5	3	1	0.5	0
2	4	3	1	0.5	0
3	4	3	1	0.5	0
4	5	3	1	0.5	0
5	4	3	1	0.5	0
6	7	3	4	1	1
7	6	3	4	1	1
8	6	3	4	1	1
9	4	2	3	1	1
10	6	3	6	2	2
11	5	2	5	2	2
12	7	3	5	2	2
13	5	2	5	2	2
14	6	2	5	1	2

Figure 7: Dataset for Iris Flowers

23. Assume a dataset as in Figure 7, and a new data instance $(6, 3, 5, 1)$, please indicate which class the new data instance will be classified into, if we consider Naive Bayes method:

☐ **A.** Class 0

☐ **B.** Class 1

■ **C.** Class 2

☐ **D.** Either Class 1 or 2

☐ **E.** Naive Bayes cannot be applied to this dataset

Part 3: Deep Learning

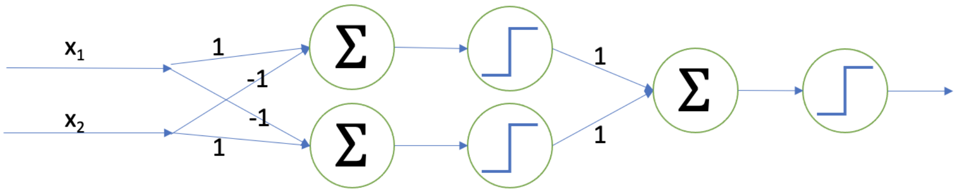


Figure 8: A simple two-layer perceptron

24. Given a simple two-layer perceptron as in Figure 8, where each layer has a linear transformation – denoted as Σ in the figure – together with an activation function ReLU. ReLU activation function is defined as $ReLU = max(0, x)$. Assume that we have a data instance $(0, 1)$, please indicate which of the following option is the output of the network:

☐ **A.** 0

■ **B.** 1

☐ **C.** 2

☐ **D.** 3

☐ **E.** None of the above is correct

25. Given a simple two-layer perceptron as in Figure 8, where each layer has a linear transformation – denoted as Σ in the figure – together with an activation function ReLU. ReLU activation function is defined as $ReLU = max(0, x)$. Assume that we have a data instance $(1, 1)$, please indicate which of the following options is the output of the network:

■ **A.** 0

☐ **B.** 1

☐ **C.** 2

☐ **D.** 3

☐ **E.** none of the above

26. Which of the following statements is correct?

☐ **A.** The success of deep learning is based on various architectures such as the well-known recurrent neural network LSTM

☐ **B.** Exclusive or (XOR) can be solved by single perceptron

☐ **C.** One of the main contributions of Frank Rosenblatt is the concept of Perceptron

☐ **D.** Rosenblatt's algorithm can be applied to train deep neural networks

■ **E.** None of the above statement is correct

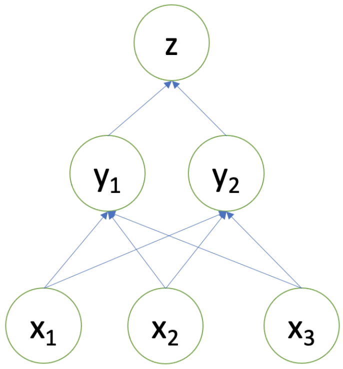


Figure 9: A simple 3-layer neural network

27. Figure 9 gives a simple 3-layer neural network with 3 inputs x_1, x_2, x_3 and a single output z. Let $z = 3y_1 + 4y_2 + 2$, $y_1 = 2x_1 + 3x_2 + x_3 + 1$, $y_2 = 3x_1 + x_2 + 5x_3 - 2$. Please indicate which of the following expressions is correct for the gradient ?

☐ **A.** $\dfrac{\partial z}{\partial x_1} = 17$

☐ **B.** $\dfrac{\partial z}{\partial x_3} = 6$

☐ **C.** $\dfrac{\partial z}{\partial x_2} = 12$

☐ **D.** $\dfrac{\partial z}{\partial x_2} = 18$

■ **E.** $\dfrac{\partial z}{\partial x_3} = 23$

input

4	0	1	7
5	6	9	-5
-3	8	3	6
2	-2	-1	4

filter

7	-3
5	2

Figure 10: A two-dimensional input and a convolutional filter

28. The following four questions are related to Figure 10. In Figure 10, we have a two-dimensional input and a convolutional filter. Given stride = 1, please indicate which of the following statements is correct if zero-padding is applied ?

☐ **A.** the result is a one dimensional array of length 9

☐ **B.** the result is a one dimensional array of length 16

☐ **C.** the result is a two dimensional array of shape $(3, 3)$

☐ **D.** the result is a two dimensional array of shape $(4, 3)$

■ **E.** None of the above is correct

29. Continue with the above question related to Figure 10, where there are a two-dimensional input and a convolutional filter. Given stride=1, please indicate which of the following statements is correct for the result of applying the convolutional filter on the input if zero-padding is applied ?

■ **A.** there is an element 45

☐ **B.** the smallest element is 39

☐ **C.** there is an element 49

☐ **D.** the greatest element is 69

☐ **E.** None of the above is correct

30. Take the same input as in Figure 10 and apply max-pooling on 2×2 filter with a stride 2. Please indicate which of the following statements is correct ?

☐ **A.** the result is a one dimensional array of length 2

☐ **B.** the result is a one dimensional array of length 4

■ **C.** the result is a two dimensional array of shape $(2, 2)$

☐ **D.** the result is a two dimensional array of shape $(3, 3)$

☐ **E.** None of the above is correct

31. Continue with the above question. Please indicate which of the following statement is correct for the result of applying max-pooling on 2×2 filter (stride 2) on the input ?

☐ **A.** there is a single element with value 7

■ **B.** there is a single element with value 9

☐ **C.** there are two elements with value 7

☐ **D.** there are three elements with value 9

☐ **E.** None of the above is correct

32. Which of the following statements is incorrect with respect to the features and the feature manifolds?

☐ **A.** In an end-to-end learning of feature hierarchy, initial modules capture low-level features, middle modules capture mid-level features, and last modules capture high level, class specific features.

☐ **B.** The computation of the coordinates of the data with respect to feature manifolds enables an easy separation of the data.

☐ **C.** Generation of feature manifolds for a given dataset is a non-trivial task, and the dimensionality reduction techniques such as PCA and t-SNE cannot work well on complex datasets.

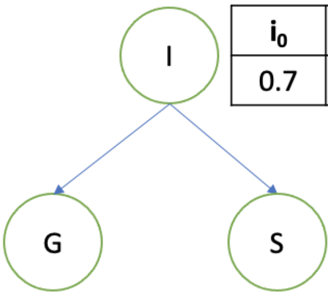
☐ **D.** It is very often that high-dimensional data lie on lower dimensional feature manifolds.

■ **E.** All of the above statements are correct.

Part 4: Probabilistic Graphical Models

33. Which of the following statements is correct:

- ☐ **A.** Naive Bayes cannot be represented as a Bayesian network
- ☐ **B.** Bayesian network is an alias of Bayesian neural network
- ☐ **C.** Every node a in a Bayesian network is associated with a conditional probability table $P(a|S)$ with S being a nonempty set of other nodes
- ☐ **D.** Graph and joint probability distribution are two key factors for Bayesian networks to represent joint probability distribution
- ■ **E.** None of the above statement is correct



	i_0	i_1
	0.7	0.3

	g_1	g_2	g_3
i_0	0.2	0.34	0.46
i_1	0.74	0.17	0.09

	s_0	s_1
i_0	0.95	0.05
i_1	0.2	0.8

Figure 11: Simple Probabilistic Graphical Model

34. Figure 11 provides a simple probabilistic graphical model of three variables S, G, and I. We already know that

$$P \models (S \perp G \mid I)$$

Which of the following is the value of $P(i_1, s_1, g_3)$?

☐ **A.** 0.0410

☐ **B.** 0.246

☐ **C.** 0.0408

■ **D.** 0.0216

☐ **E.** 0.317

X	Y	Z	P(X,Y,Z)
x_0	y_0	z_0	0.02
x_0	y_1	z_0	0.08
x_1	y_0	z_0	0.03
x_1	y_1	z_0	0.12
x_0	y_0	z_1	0.06
x_0	y_1	z_1	0.24
x_1	y_0	z_1	0.09
x_1	y_1	z_1	0.36

(a)

X	Y	Z	P(X,Y,Z)
x_0	y_0	z_0	0.01
x_0	y_1	z_0	0.04
x_1	y_0	z_0	0.015
x_1	y_1	z_0	0.06
x_0	y_0	z_1	0.07
x_0	y_1	z_1	0.28
x_1	y_0	z_1	0.105
x_1	y_1	z_1	0.42

(b)

Figure 12: Joint probability of three random variables

35. Figure 12 (a) provides a joint probability P. Let $I(P)$ to be the set of conditional independence assertions of the form $(X \perp Y | Z)$ that hold in P. Which of the following is correct?

☐ **A.** $(X \perp Y \mid Z) \notin I(P)$ but $(X \perp Y) \in I(P)$

☐ **B.** $(Y \perp Z \mid X) \in I(P)$ but $(X \perp Y) \notin I(P)$

■ **C.** $(X \perp Y) \in I(P)$

☐ **D.** $I(P) = \emptyset$

☐ **E.** None of the above is correct

36. Figure 12 (b) provides a joint probability P. Let $I(P)$ be the set of conditional independence assertions of the form $(X \perp Y | Z)$ that hold in P. Which of the following is correct?

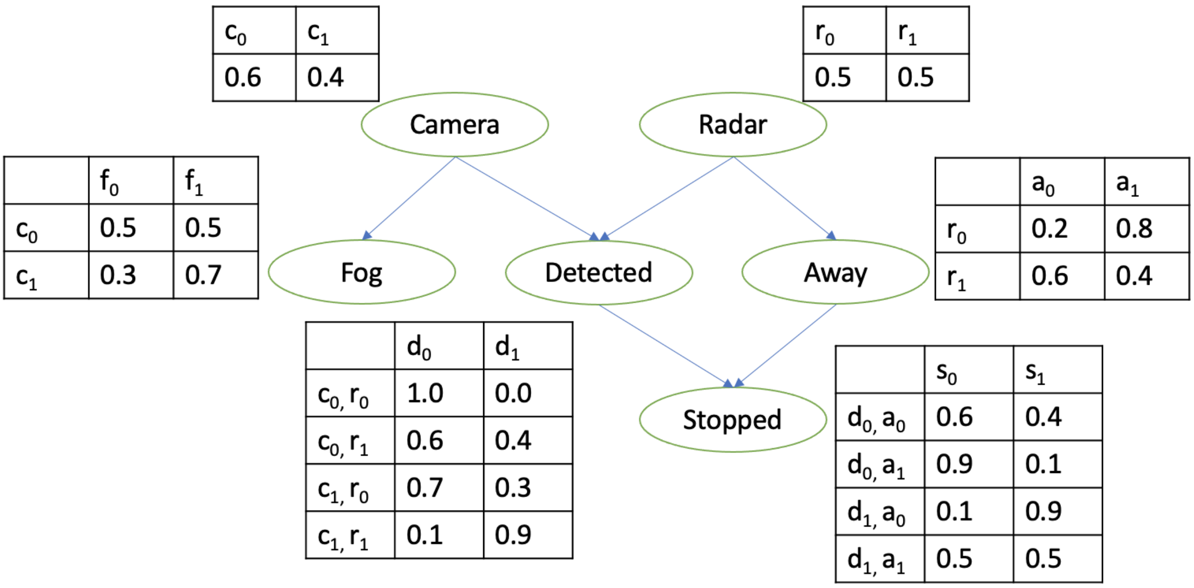
☐ **A.** $(X \perp Z \mid Y) \notin I(P)$

☐ **B.** $(Y \perp Z \mid X) \notin I(P)$

☐ **C.** $(X \perp Y) \notin I(P)$

☐ **D.** $I(P) = \emptyset$

■ **E.** None of the above is correct



c_0	c_1
0.6	0.4

r_0	r_1
0.5	0.5

	f_0	f_1
c_0	0.5	0.5
c_1	0.3	0.7

	a_0	a_1
r_0	0.2	0.8
r_1	0.6	0.4

	d_0	d_1
c_0, r_0	1.0	0.0
c_0, r_1	0.6	0.4
c_1, r_0	0.7	0.3
c_1, r_1	0.1	0.9

	s_0	s_1
d_0, a_0	0.6	0.4
d_0, a_1	0.9	0.1
d_1, a_0	0.1	0.9
d_1, a_1	0.5	0.5

Figure 13: A Bayesian network G

37. Consider the Bayesian network model G in Figure 13 and indicate which of the following is not in the I-map $I(G)$:

☐ **A.** $(Fog \perp Radar, Detected, Away, Stopped \mid Camera)$

■ **B.** $(Detected \perp Radar, Fog, Stopped \mid Camera, Away)$

☐ **C.** $(Camera \perp Radar)$

☐ **D.** $(Away \perp Camera, Fog, Detected \mid Radar)$

☐ **E.** All the above are in $I(G)$

38. Consider the Bayesian network model G in Figure 13 and calculate the following value

$$P(c_0, r_1, a_1, d_0, f_0, s_1) =$$

■ **A.** 0.0036

☐ **B.** 0.0342

☐ **C.** 0.0024

☐ **D.** 0.0252

☐ **E.** none of the above answers is correct

39. Consider the Bayesian network model G in Figure 13 and calculate the following conditional probability (rounded to two decimal places)

$$P(c_0|s_0, d_0, a_0) =$$

☐ **A.** 0.34

☐ **B.** 0.78

☐ **C.** 0.02

■ **D.** 0.81

☐ **E.** none of the above answers is correct

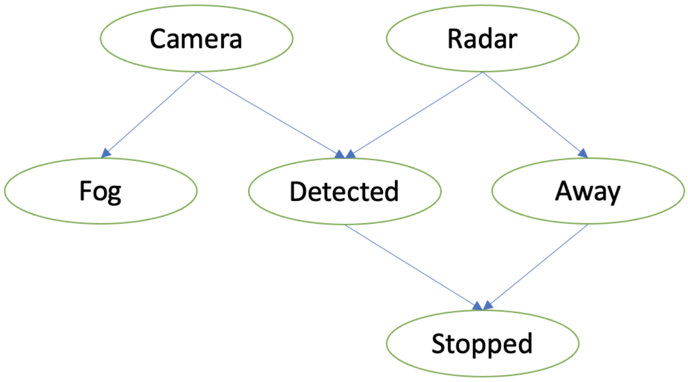


Figure 14: A simple probabilistic graphical model

40. Consider the probabilistic graphical model in Figure 14. Please indicate which of the following statement is incorrect.

■ **A.** *Fog* can influence *Radar* when only *Away* is observed.

☐ **B.** *Camera* can influence *Radar* when only *Detected* is observed.

☐ **C.** *Camera* can influence *Away* when only *Detected* is observed.

☐ **D.** *Away* can influence *Fog* when only *Stopped* and *Radar* are observed.

☐ **E.** *Fog* can influence *Radar* when only *Stopped* is observed.

This page collects some formulas/expressions that may be used in this exam.

1. entropy:

$$-\sum_{y \in values(Y)} P(y) \log_2 P(y)$$

2. conditional entropy:

$$H(Y|X) = \sum_{x \in values(X)} P(X = x) H(Y|X = x)$$

where

$$H(Y|X = x) = -\sum_{y \in values(Y)} P(Y = y|X = x) \log_2 P(Y = y|X = x)$$