上海交通大学试卷

(2019−2020 Academic Year/Fall Semester)

Class No. ____________________ Student ID No. ____________________

Name in English/Pinyin: ________________ Name in Hanzi, if applicable: ____________

Ve492 Introduction to Artificial Intelligence

Final Exam

December 9, 2019, 4:00pm-5:40pm

The exam paper has **13** pages in total.

Exam rules and information:

- The five main sections of this exam are independent.
- This is a closed book exam. You are allowed two sheets of A4 paper, which should be stapled, with your own handwritten notes. Photocopies are not allowed.
- The last two pages of the exam papers can be used as scratch papers. If you need more scratch papers, let the proctors know.
- No electronic devices are allowed. These include laptops, cell phones and smart watches.

You are to abide by the University of Michigan-Shanghai Jiao Tong University Joint Institute (UM-SJTU JI) honor code. Please sign below to signify that you have kept the honor code pledge.

THE UM-SJTU JI HONOR CODE

I accept the letter and spirit of the honor code:
I have neither given nor received unauthorized aid on this examination, nor have I concealed any violations of the Honor Code by myself or others.

Signature: ____________________

Please enter grades here:

Exercice No. 题号	Points 得分	Grader's Signature 流水批阅人签名
1.1		
1.2		
2.1		
2.2		
3		
4.1		
4.2		
5		
Total		

1 Logic

1.1 Propositional Logic

Consider a propositional logic with only four symbols: A, B, C, and D. How many models are there for each of the following sentences? **Justify** your answers.

Hint: The justification can be a truth table, or an explanation (e.g., see example below), or possibly some other convincing demonstration.

Example: The sentence $C \wedge D$ has four models: $C = D = \mathit{true}$ and A and B can be either true or false.

1. (2 points) $C \vee \neg A$

   Solution: 12: There are three possible combinations of truth values for A and C, times four possible combinations for B and D.

2. (2 points) $\neg(B \Rightarrow D)$

   Solution: 4: All models must have B = true and D = false, but A and C can be either true or false.

3. (2 points) $(B \vee C) \wedge (\neg C \vee D)$

   Solution: 8: If C is true, then D must also be true, but B (and A) can be either true or false. If C is false, then B must be true, but D (and A) can be either true or false. (12, the number of models for B $\vee$ D, is tempting but incorrect.)

4. (2 points) $(A \Rightarrow B) \wedge (B \Rightarrow C) \wedge (C \Rightarrow D) \wedge (D \Rightarrow A)$

   Solution: 2: The model in which A=B=C=D=true, and the model in which A = B = C = D = false.

5. (2 points) $(A \wedge B) \vee (B \wedge C)$

   Solution: The four models in which A = B = true (with C and D unspecified), plus the four models in which B = C = true (with A and D unspecified), minus the two models in which A = B = C = true (D unspecified), which have been double-counted.

1.2 First-Order Logic

Translate the sentences below in first-order logic using the following predicates:

- $Parent(x, p)$ which means that p is the parent of x,
- $Gender(x, g)$ which means that x is of gender g where g is $Male$ or $Female$,

or one created from any word in italics. For a given predicate P, $P(x, y)$ means y is in relation with x according to P (like for Parent).

1. (2 points) Two people who share a parent are *siblings*.

Solution: $\forall x, y\ \exists p\ Parent(x, p) \land Parent(y, p) \land \neg x = y \Rightarrow Sibling(x, y)$

2. (2 points) A *brother* is a male sibling. A *sister* is a female sibling.

Solution: $\forall x, y\ Sibling(x, y) \land Gender(y, Male) \Rightarrow Brother(x, y)$
$\forall x, y\ Sibling(x, y) \land Gender(y, Female) \Rightarrow Sister(x, y)$

3. (2 points) The brother of a parent is an *uncle*. The sister of a parent is an *aunt*.

Solution: $\forall x, u\ \exists p\ Parent(x, p) \land Brother(p, u) \Rightarrow Uncle(x, u)$
$\forall x, u\ \exists p\ Parent(x, p) \land Sister(p, a) \Rightarrow Aunt(x, a)$

4. (2 points) A parent is an *ancestor*. A parent of an *ancestor* is an *ancestor*.

Solution: $\forall x, y\ Parent(x, y) \Rightarrow Ancestor(x, y)$
$\forall x, z\ \exists y\ Ancestor(x, y) \land Parent(y, z) \Rightarrow Ancestor(x, z)$

5. (2 points) A *cousin* is a child of one's uncle or aunt.

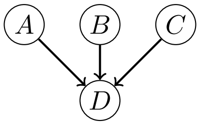
Solution:

(a) What's wrong with $Grandparent(x, g) \land Grandparent(y, g) \Rightarrow Cousin(x, y)$? Siblings would be defined as cousins, which is not correct.
(b) What's a good way to define $Cousin(x, y)$?
$\forall x, y\ \exists p, q\ Parent(x, p) \land Sibling(p, q) \land Child(q, y) \Rightarrow Cousin(x, y)$

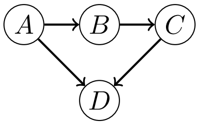
2 Bayes' Net

2.1 Representation

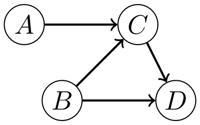
The full joint probability distribution $\mathbb{P}(A, B, C, D)$ for four binary random variables is a 16-element array of numbers adding up to 1.0, so it has 15 degrees of freedom. A Bayes net makes explicit the conditional dependencies (and therefore also the conditional independence relations) among the random variables, which simplifies the information that must be provided to determine $\mathbb{P}(A, B, C, D)$. For each of the Bayes' net diagrams below, express $\mathbb{P}(A, B, C, D)$ as the product of conditional probabilities, and give the number of degrees of freedom required to specify the conditional probability tables.



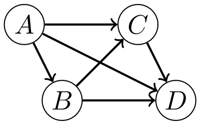
Graph 1.



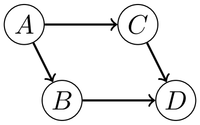
Graph 2.



Graph 3.



Graph 4.



Graph 5.

1. (3 points) Provide $\mathbb{P}(A, B, C, D)$ and the number of degrees of freedom for Graph 1.

> **Solution:** (1pt)
>
> $\mathbb{P}(A, B, C, D) = \mathbb{P}(D \mid A, B, C)\mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$ 不用附表格.
>
> (2pt)
>
> Number of degrees of freedom $= 8 + 3 = 11$

2. (3 points) Provide $\mathbb{P}(A, B, C, D)$ and the number of degrees of freedom for Graph 2.

> **Solution:** (1pt)
>
> $\mathbb{P}(A, B, C, D) = \mathbb{P}(D \mid A, C)\mathbb{P}(C \mid B)\mathbb{P}(B \mid A)\mathbb{P}(A)$
>
> (2pt)
>
> Number of degrees of freedom $= 4 + 2 + 2 + 1 = 9$

3. (3 points) Provide $\mathbb{P}(A, B, C, D)$ and the number of degrees of freedom for Graph 3.

> **Solution:** (1pt)
>
> $\mathbb{P}(A, B, C, D) = \mathbb{P}(D \mid B, C)\mathbb{P}(C \mid A, B)\mathbb{P}(B)\mathbb{P}(A)$
>
> (2pt)
>
> Number of degrees of freedom $= 4 + 4 + 1 + 1 = 10$

4. (3 points) Provide $\mathbb{P}(A, B, C, D)$ and the number of degrees of freedom for Graph 4.

> **Solution:** (1pt)
>
> $\mathbb{P}(A, B, C, D) = \mathbb{P}(D \mid A, B, C)\mathbb{P}(C \mid A, B)\mathbb{P}(B \mid A)\mathbb{P}(A)$
>
> (2pt)
>
> Number of degrees of freedom $= 8 + 4 + 2 + 1 = 15$

5. (3 points) Provide $\mathbb{P}(A, B, C, D)$ and the number of degrees of freedom for Graph 5.

Solution: (1pt)

$\mathbb{P}(A, B, C, D) = \mathbb{P}(D \mid B, C)\mathbb{P}(C \mid A)\mathbb{P}(B \mid A)\mathbb{P}(A)$

(2pt)

Number of degrees of freedom $= 4 + 2 + 2 + 1 = 9$

2.2 Conditional Independences

Consider the following Bayes' net structure. Answer "yes" or "no" to the following questions regarding (conditional) independences and **justify** your answer.

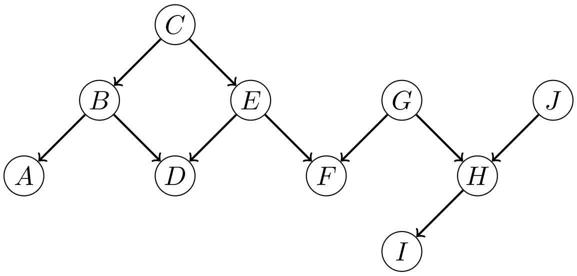


Figure 1: A Bayes' Net

1. (3 points) $G \perp\!\!\!\perp I$

Solution: (1pt for answer and 2pt for justification)

No, there is only on path (G, H, I), which is active.

2. (3 points) $E \perp\!\!\!\perp H|\ C$

Solution: (1pt for answer and 2pt for justification)

Yes, there's only one path (E, F, G, H), which is inactive because of the convergent chain in F, which is unobserved.

3. (3 points) $B \perp\!\!\!\perp H \mid A, C, D, E, F, G, J$

Solution: (1pt for answer and 2pt for justification)

Yes, there are two paths (B, C, E, F, G, H) and (B, D, E, F, G, H), both are inactive because of the divergent chain in G, which is observed.

4. (3 points) $B \perp\!\!\!\perp F \mid A, C, D, G, H$

Solution: (1pt for answer and 2pt for justification)

No, the path (B, D, E, F) is active because D, which is observed, forms a divergent chain and E, which is unobserved forms a convergent chain.

5. (3 points) $A, I \perp\!\!\!\perp C, G \mid B, H$

Solution: (1pt for answer and 2pt for justification)

Yes, all the paths from A to C or G, or from I to C or G are inactive because of the observations.

3 Markov Models

Consider the Markov chain depicted in Figure 2.

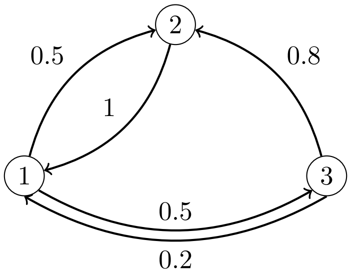


Figure 2: A Simple Markov Chain

1. (4 points) Assuming an uniform distribution over initial states, compute the probability of being in each state at time step 2.

Solution: (2pts for writing the system and 2pts for solving it)

$$(1/3\ 1/3\ 1/3) \times \begin{bmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 0.2 & 0.8 & 0 \end{bmatrix} \times \begin{bmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 0.2 & 0.8 & 0 \end{bmatrix}$$

2. (4 points) Compute its stationary distribution.

Solution: (2pts for writing the system and 2pts for solving it)

Solve the system of linear equations defined by

$$(x\ y\ z) \times \begin{bmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 0.2 & 0.8 & 0 \end{bmatrix} = (x\ y\ z)$$

and $x + y + z = 1$.

4 Machine Learning

4.1 Naïve Bayes

Your goal is to determine whether an individual email message is spam or ¬spam, based on certain words it contains. Unfortunately, spam (junk email) makes up about 60% of email traffic. (Based on a quick search, this appears to be an accurate figure.) Fortunately, by examining large numbers of spam and non-spam email messages, the frequencies of appearance of certain words, for example, “Nigerian” , “million” , and “viagra” , are significantly different across the two categories of messages. (The numbers in this table are made up for this problem, not based on data from actual spam.)

word	$\mathbb{P}(word \mid spam)$	$\mathbb{P}(word \mid \neg spam)$
“Nigerian”	0.01	0.001
“million”	0.02	0.1
“viagra”	0.015	0.001

1. (2 points) What is the prior probability $\mathbb{P}(spam)$?

Solution: The prior probability $\mathbb{P}(spam) = 0.60$.

2. (5 points) Use Bayes Rule to compute the posterior probabilities $\mathbb{P}(spam \mid \text{“million”})$ and $\mathbb{P}(\neg spam \mid \text{“million”})$.

$$P(spam \mid \text{"million"}) = \frac{P(sp\ million \mid spam)\,P(s}{P(\text{"million})}$$

Solution: The un-normalized posterior probabilities are:

$$\mathbb{P}(spam \mid \text{“million”}) = \alpha \cdot \mathbb{P}(\text{“million”} \mid spam) \cdot \mathbb{P}(spam) = \alpha \cdot .02 \cdot .60 = \alpha \cdot .012$$

$$\mathbb{P}(\neg spam \mid \text{“million”}) = \alpha \cdot \mathbb{P}(\text{“million”} \mid \neg spam) \cdot \mathbb{P}(\neg spam) = \alpha \cdot .01 \cdot .40 = \alpha \cdot .004$$

Normalizing, so they add up to 1.0, we get: $\mathbb{P}(spam \mid \text{“million”}) = 12/16 = 0.75$ and $\mathbb{P}(\neg spam \mid \text{“million”}) = 4/16 = 0.25$

3. (5 points) Use Bayes Rule to compute the posterior probabilities $\mathbb{P}(\textit{spam} \mid \text{“viagra”})$ and $\mathbb{P}(\neg \textit{spam} \mid \text{“viagra”})$.

Solution: The un-normalized posterior probabilities are:

$$\mathbb{P}(\textit{spam} \mid \text{“viagra”}) = \alpha \cdot \mathbb{P}(\text{“viagra”} \mid \textit{spam}) \cdot \mathbb{P}(\textit{spam}) = \alpha \cdot .015 \cdot .60 = \alpha \cdot .0090$$
$$\mathbb{P}(\neg\textit{spam} \mid \text{“viagra”}) = \alpha \cdot \mathbb{P}(\text{“viagra”} \mid \neg\textit{spam}) \cdot \mathbb{P}(\neg\textit{spam}) = \alpha \cdot .001 \cdot .40 = \alpha \cdot .0004$$

Normalizing, so they add up to 1.0, we get: $\mathbb{P}(\textit{spam} \mid \text{“viagra”}) = 90/94 = 0.9574$ $\mathbb{P}(\neg\textit{spam} \mid \text{“viagra”}) = 4/94 = 0.0426$.

4. (5 points) Using the Naive Bayes assumption that individual word frequencies are independent, compute the posterior probability that a message is spam (or not) if it contains all three of these words:
$\mathbb{P}(\textit{spam} \mid \text{“Nigerian”}, \text{“million”}, \text{“viagra”})$, $\mathbb{P}(\neg\textit{spam} \mid \text{“Nigerian”}, \text{“million”}, \text{“viagra”})$

Solution: The un-normalized posterior probabilities are:

$$\mathbb{P}(\textit{spam} \mid \text{“Nigerian”}, \text{“million”}, \text{“viagra”}) = \alpha \cdot .01 \cdot .02 \cdot .015 \cdot .60 = \alpha \cdot 1800 \cdot 10^{-9}$$
$$\mathbb{P}(\neg\textit{spam} \mid \text{“Nigerian”}, \text{“million”}, \text{“viagra”}) = \alpha \cdot .001 \cdot .01 \cdot .001 \cdot .40 = \alpha \cdot 4 \cdot 10^{-9}$$

Normalizing, so they add up to 1.0, we get: $\mathbb{P}(\textit{spam} \mid \text{“Nigerian”}, \text{“million”}, \text{“viagra”}) = 1800/1804 = 0.9978$ and $\mathbb{P}(\neg\textit{spam} \mid \text{“Nigerian”}, \text{“million”}, \text{“viagra”}) = 4/1804 = 0.0022$

4.2 Binary Classification

Recall a probabilistic binary classifier $f : \mathcal{X} \to \mathcal{Y}$ is defined as follows:

$$f(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \eta \geq \frac{1}{2} \\ 0 & \text{if } \eta < \frac{1}{2} \end{cases}$$

where $\boldsymbol{x}$ is known to be a non-negative vector and the two classes are encoded 0 and 1. We consider here a probabilistic binary classifier where η is defined by:

$$\eta_{\boldsymbol{\theta}} = \frac{\boldsymbol{w}_1^\intercal \boldsymbol{x}}{\boldsymbol{w}_0^\intercal \boldsymbol{x} + \boldsymbol{w}_1^\intercal \boldsymbol{x}}$$

where $\boldsymbol{\theta} = (\boldsymbol{w}_0, \boldsymbol{w}_1) \in \mathbb{R}_+^{2(D+1)}$ and assuming each instance $\boldsymbol{x} \in \mathbb{R}^{D+1}$ has been appended a constant 1.

1. (3 points) Write the equation of the decision boundary for this classifier.

Solution: $(\boldsymbol{w}_0 - \boldsymbol{w}_1)^\intercal \boldsymbol{x} = 0$

2. (3 points) What are the possible issues with this model?

Solution: The model defines a linear classifier. The weights may become negative while training, especially if the classification problem is not linearly separable.

3. We would like to learn parameters $\boldsymbol{\theta}$ from a dataset $\mathcal{D} = \{(\boldsymbol{x}^i, y^i) \,|\, i = 1, \ldots, N\}$, which has been collected in an i.i.d. way.

   (a) (3 points) Write the log-likelihood of $\mathcal{D}$ given parameter $\boldsymbol{\theta}$: $L(\boldsymbol{\theta}; \mathcal{D}) = \log \mathbb{P}(\mathcal{D} \,|\, \boldsymbol{\theta})$.

   Solution: $\mathbb{P}(\mathcal{D} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^N \big(\frac{\boldsymbol{w}_1^\intercal \boldsymbol{x}^i}{\boldsymbol{w}_0^\intercal \boldsymbol{x}^i + \boldsymbol{w}_1^\intercal \boldsymbol{x}^i}\big)^{y^i} \big(\frac{\boldsymbol{w}_0^\intercal \boldsymbol{x}^i}{\boldsymbol{w}_0^\intercal \boldsymbol{x}^i + \boldsymbol{w}_1^\intercal \boldsymbol{x}^i}\big)^{1-y^i}$
   $L(\boldsymbol{\theta}; \mathcal{D}) = \log \mathbb{P}(\mathcal{D}|\boldsymbol{\theta}) = \sum_{i=1}^N y^i \log \boldsymbol{w}_1^\intercal \boldsymbol{x}^i + (1 - y^i) \log \boldsymbol{w}_0^\intercal \boldsymbol{x}^i - \log\big((\boldsymbol{w}_0 + \boldsymbol{w}_1)^\intercal \boldsymbol{x}^i\big)$

   (b) (4 points) Write the gradient of this log-likelihood with respect to $\boldsymbol{\theta}$. Recall the gradient is given by:

$$\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathcal{D}) = \big((\frac{\partial L(\boldsymbol{\theta}; \mathcal{D})}{\partial w_{0k}})_{k=1,\ldots,D+1}, (\frac{\partial L(\boldsymbol{\theta}; \mathcal{D})}{\partial w_{1k}})_{k=1,\ldots,D+1}\big)$$

   Solution: $\frac{\partial L(\boldsymbol{\theta};\mathcal{D})}{\partial w_{1k}} = \sum_{i=1}^N \frac{y^i}{\boldsymbol{w}_1^\intercal \boldsymbol{x}^i} x_k^i + \frac{1}{(\boldsymbol{w}_0 + \boldsymbol{w}_1)^\intercal \boldsymbol{x}^i} x_k^i$
   $\frac{\partial L(\boldsymbol{\theta};\mathcal{D})}{\partial w_{0k}} = \sum_{i=1}^N \frac{1-y^i}{\boldsymbol{w}_0^\intercal \boldsymbol{x}^i} x_k^i + \frac{1}{(\boldsymbol{w}_0 + \boldsymbol{w}_1)^\intercal \boldsymbol{x}^i} x_k^i$

5 General Questions

For each question, answer ”yes” or ”no” and **justify** your answer.

1. (2 points) For any events E and F, we always have $\mathbb{P}(E \,|\, F) \geq \min(\mathbb{P}(E), \mathbb{P}(F))$

Solution: (1pt for answer, 1 pt for justification)

No, because if $\mathbb{P}(E) > 0$, $\mathbb{P}(F) > 0$ and $E \cap F = \emptyset$, then
$\mathbb{P}(E \cap F) = 0 < \min(\mathbb{P}(E), \mathbb{P}(F))$.

2. (2 points) For any events E and F, we always have $\mathbb{P}(E \mid F) \leq \max(\mathbb{P}(E), \mathbb{P}(F))$

Solution: (1pt for answer, 1 pt for justification)

No, for instance, for $E = F$ and $\mathbb{P}(E) < 1$, we have $\mathbb{P}(E \mid F) = 1$.

3. (2 points) We are given a joint distribution $\mathbb{P}(X, Y)$ over variables X and Y whose domains are respectively $\mathcal{D}_X$ and $\mathcal{D}_Y$. We always have $\sum_{x \in \mathcal{D}_X} \mathbb{P}(X = x \mid Y = y) = 1$ for any $y \in \mathcal{D}_Y$ such that $\mathbb{P}(y) > 0$.

Solution: (1pt for answer, 1 pt for justification)

Yes, $\mathbb{P}(X \mid Y = y)$ is a probability distribution.

4. (2 points) We are given a joint distribution $\mathbb{P}(X, Y)$ over variables X and Y whose domains are respectively $\mathcal{D}_X$ and $\mathcal{D}_Y$. Assume that $\forall y \in \mathcal{D}_Y, \mathbb{P}(y) > 0$. Then, we always have $\sum_{y \in \mathcal{D}_Y} \mathbb{P}(X = x \mid Y = y) = 1$ for any $x \in \mathcal{D}_X$.

Solution: (1pt for answer, 1 pt for justification)

No, $\mathbb{P}(X = x \mid Y)$ is not a probability distribution.

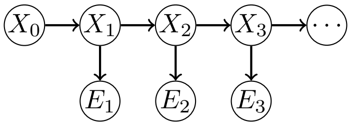


Figure 3: Example of HMM.

5. (2 points) The Viterbi algorithm has the same computational complexity as Variable Elimination if variables are selected in the chronological order.

Solution: (1pt for answer, 1 pt for justification)

Yes. Both are in O(|X| T).

6. (2 points) The Viterbi algorithm has the same space complexity as Variable Elimination if variables are selected in the chronological order.

Solution: (1pt for answer, 1 pt for justification)

No. The Viterbi algorithm needs to store more information in order to rebuild the most likely path.