Name _____

COMPSCI 348 **Practice Exam 3** Spring 2019

Instructions (for the actual exam)

- *Do not open the exam until directed to by the instructor or TA*
- *Do not use books, notes, electronic devices, or other aids.*
- *Please avoid wrinkling the exam because that makes it difficult to scan.*
- *Your answers must be your own, so keep your eyes on your exam. Do not look at other students' exams.*
- *Answer each question. Note the point values and allocate your time accordingly.*
- *Be clear in marking your answers, and please place your answers in the designated spaces.*
- *Only the final answer in the designated space will be graded. However, other markings and calculations will be reviewed in support of regrade requests.*

Name _____

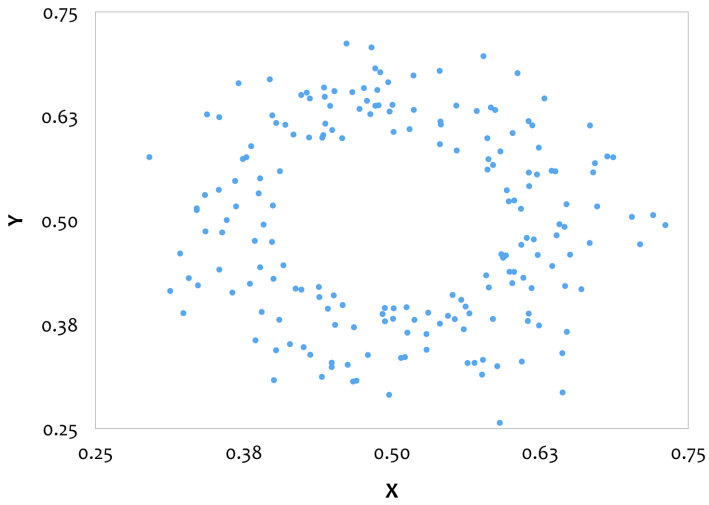
1. **Goals of data science** — For each statement below, mark it either True or False by filling in the appropriate circle. (10 points)

True	False	Question
◯	◯	Methods that model univariate probability distributions are primarily used for predictive analytics.
◯	◯	Methods for causal modeling are primary used for prescriptive analytics.
◯	◯	To perform well at predictive analytics, it is usually necessary for a model to accurately represent the internal structure of the underlying data generating process (DGP).
◯	◯	Simple Bayesian classifiers are often used for causal modeling.
◯	◯	Methods for exploratory data analysis and visualization are typically used in descriptive analytics.

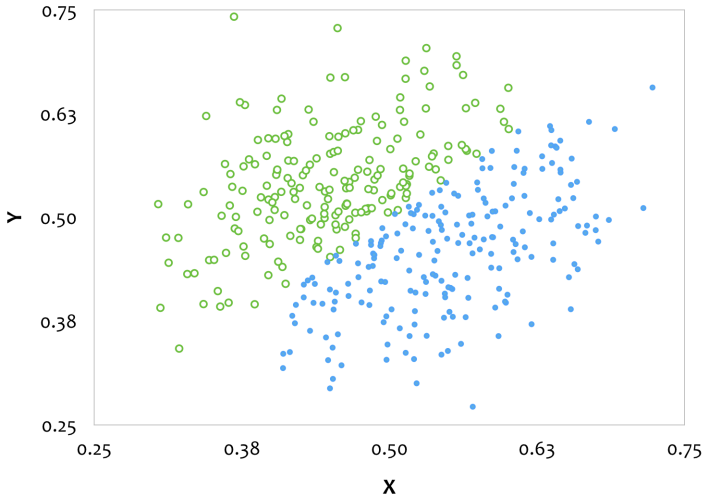
Name _____

2. Independence and dependence — Answer each question below.

a. Below is a scatterplot of data instances sampled from the joint distribution of two variables, X and Y. The horizontal and vertical axes show the values of X and Y, respectively. State in words how you would determine whether X and Y are marginally independent. (5 points)



b. Below is a scatterplot of data instances sampled from the joint distribution of three variables, X, Y, and Z. The horizontal and vertical axes show the values of X and Y, respectively. The value of a third variable, Z, is shown as the symbol and color of the plotted point. State in words how you would determine whether X and Y are conditionally independent given Z. (5 points)



Name _____

3. **Assumptions** — In your own words, briefly define each of the assumptions below. (10 points)

   a. Causal Markov assumption

   b. Causal sufficiency

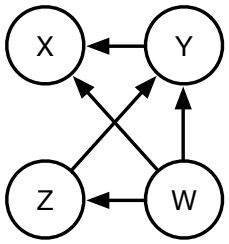
   c. Faithfulness

   d. Positivity

Name _____

4. Semantics of Bayesian networks — Answer each question below.

a. For the Bayesian network at right, use probability notation to write the equation for the joint probability distribution as a product of conditional distributions. (4 points)



b. Can two Bayesian networks with different structure represent the same joint probability distribution? Say "yes" or "no" and briefly describe why or why not. (2 points)

c. Can a graph with N variables (nodes) and N^2 dependencies (edges) be a valid Bayesian network? Say "yes" or "no" and briefly explain why or why not. (4 points)

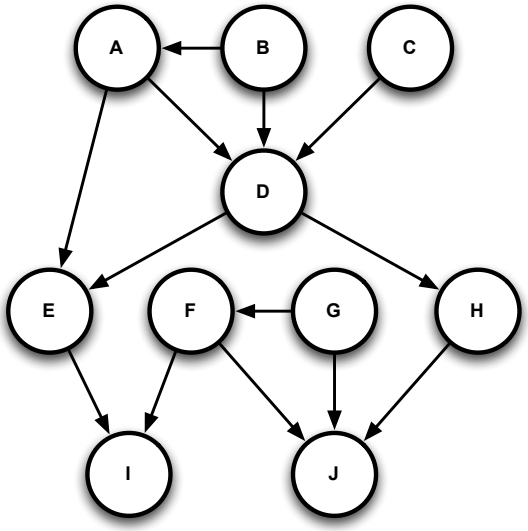
Name _____

5. **Bayesian networks** — Draw the graphical structure of the Bayesian network that represents the full joint distribution P(D,R,F,W,G,I,M,B,L) corresponding to the situation described below. (10 points)

   "Ghoulisitis is a disease (D) that can cause your eyes to turn red (R), your hair to fall out (F), and your skin to turn pasty white (W). The disease has both a genetic component (G) and an infectious component (I). Persons afflicted with the disease usually have a single recessive gene, inherited from their mother's genome (M), and they have suffered an infectious bite of a ghoul-rat (B). Initially, epidemiologists studying ghoulisitis were baffled by many cases of genetically susceptible individuals being bitten by ghoul-rats and not contracting the disease. However, recent research has revealed one additional influence on disease transmission, rat-bites that occur under the light of a full moon (L) help activate the ghoulisitis gene and make the rat-bite more infectious. Also, diagnosing the disease is complicated by the fact that being bitten by a ghoul-rat is so frightening that some individuals who have been bitten by one acquire a pasty white complexion even though they do not contract the disease."

Name _____

6. d-separation — In the Bayesian network below, use d-separation to decide whether X is conditionally independent of Y given Z. If X and Y are not conditionally independent, provide one d-connecting path by naming the sequence of nodes (e.g., ABD). The answer to the first query is provided. (2 points each)

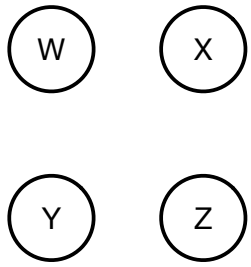


X	Y	Z	X ⫫ Y | Z? (Circle one)	*d*-connecting path (if False)
A	C	{D}	True (False)	ADC
A	C	{B,D,E}	True False	
E	J	{}	True False	
A	H	{D,I}	True False	
I	J	{F,G,D}	True False	
A	G	{E}	True False	

Name _____

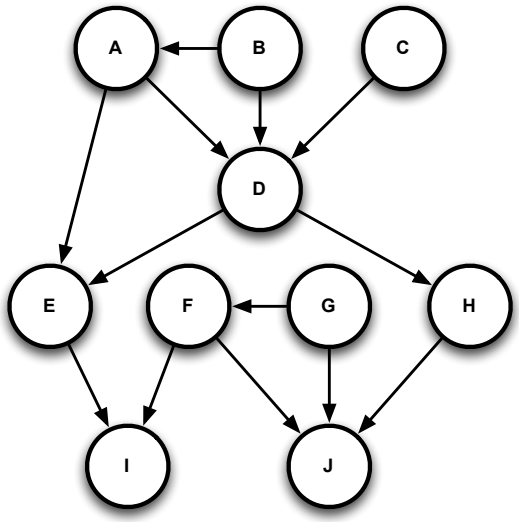
7. Inferring model structure — The conditional independence statements below provide the minimum-sized conditioning sets necessary to make any two variables in {W,X,Y,Z} conditionally independent. If no such statement is made for a given set of variables, then the variables cannot be made conditionally independent by conditioning on any set of the other variables. Use these statements to infer the structure of a graphical model consistent with the statements and draw in the directed edges between the nodes below. (10 points)

$W \perp\!\!\!\perp X$
$X \perp\!\!\!\perp Z \mid Y$
$W \perp\!\!\!\perp Z \mid Y$



Name _____

8. Causal graphical models — For the Bayesian network at right, answer each of the questions below and briefly explain your answer. (2 points each)



a. Does A cause H?

b. Does E cause H?

c. Suppose that we intervene on E, then does C cause I?

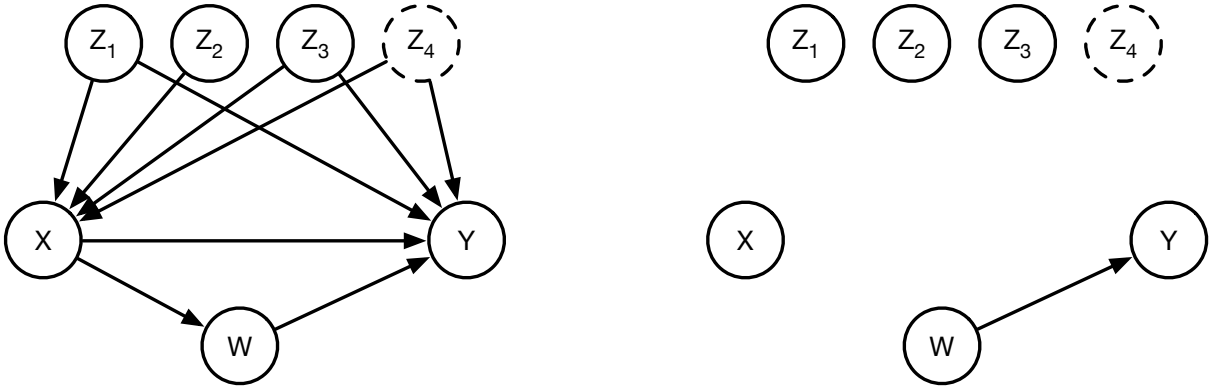
d. Suppose that we intervene on both A and D, do the variables B and C provide us any information about the other variables in the network?

e. Suppose we try to intervene on D, but aren't sure we have been successful. If you can collect data after the (potential) intervention, name one conditional independence test on that data that would help you verify whether your intervention has been successful.

Name _____

9. Experiments and quasi-experiments — Answer each question below.

a. The model in the figure below at left represents the ordinary behavior of a causal system. Assume you can perform an experiment on this system that randomly assigns the value of X. Add directed edges to figure below at right so that it represents the new causal system after random assignment of X. One valid edge (from W to Y) has already been added. (5 points)



b. Below is a brief description of an analysis that uses observational data and a propensity scores approach to determine whether consuming eggplant (E) prevents heart disease (H), given a set of possible confounding variables, including the country you were born in (C), age (A), and income (I). Fill in the missing words or letters denoting variables so that the description is complete and accurate. (5 points)

A propensity score design constructs a model that predicts the probability of _____ using the variables _____. Then, it partitions the distribution of the _____ into categories with nearly equal values. Using these partitions, it then determines if ____ is conditionally independent of _____ given _____. If it is, then E is not causal for H.

Name _____

10. Miscellaneous — For each statement below, mark it either True or False by filling in the appropriate circle. (10 points)

True	False	Question
◯	◯	Two variables X and Y are considered to be conditionally independent given Z if X and Y are independent for at least one value of Z.
◯	◯	In general, when used as a representation of the conditional probability distributions in a Bayesian network, simple Bayesian classifiers and classification trees would be equally accurate.
◯	◯	If two variables are d-connected in a Bayesian network, then they are guaranteed to be dependent.
◯	◯	A directed graphical model can be completely specified by a set of nodes (corresponding to random variables) and a set of edges (corresponding to direct causal dependence) that form a directed acyclic graph.
◯	◯	If X causes Y, then any manipulation of X will produce a change in the value of Y.