CS 491: Introduction to Machine Learning
Spring 2015
Midterm Exam

Name: ______________________

Instructions:

1. Write your name above. Do not begin the exam (look at other pages) until told to do so.
2. There should be 6 pages. Count the pages (without looking at the questions).
3. Read the instructions carefully. **Q1** asks for both a TRUE or FALSE answer AND a short explanation. **Q4.1** asks for you to circle or cross out different independence properties. There is **no penalty** for guessing on these questions.
4. Partial credit will be given for incorrect answers only if you show your work.
5. Do not discuss the exam with students who have not taken the exam!

Some useful formulas:

- $P(\mathbf{x}, \mathbf{y}, \mathbf{z}) = P(\mathbf{x})P(\mathbf{y}|\mathbf{x})P(\mathbf{z}|\mathbf{x}, \mathbf{y})$ (chain rule)
- $P(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{Y}} P(\mathbf{x}, \mathbf{y})$ (marginalization)
- $P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{x},\mathbf{y})}{P(\mathbf{y})}$ (conditioning)
- $P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}{\sum_{\mathbf{x}' \in \mathcal{X}} P(\mathbf{y}|\mathbf{x}')P(\mathbf{x}')}$ (Bayes theorem)
- $\mathbb{E}_{x \sim P}[g(X)] = \sum_{x \in \mathcal{X}} P(x)g(x)$ (discrete expectation)
- $X \sim \text{Normal}(\mu, \sigma) \implies P(X = x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- $X \sim \text{Multinoulli}(\theta) \implies P(\mathbf{x}) = \prod_{i=1}^{K} \theta_i^{x_i}$

	Points
Q1	/20
Q2	/30
Q3	/20
Q4	/30
Total	

Q1. True or False (4 questions, 20 points total)

For each question: circle <u>TRUE</u> or <u>FALSE</u> (2 points) and provide a brief explanation or picture (3 points)

Q1.1: (5 points) 3-Nearest Neighbor for binary classification is guaranteed to have a lower training set error than 5-Nearest Neighbor (where the majority class of the N nearest neighbors is predicted).

TRUE or FALSE

Explanation:

Q1.2: (5 points) The MAP estimate and the maximum likelihood estimate converge to the same solution given infinite data and a reasonable prior distribution (providing non-zero probability everywhere).

TRUE or FALSE

Explanation:

Q1.3: (5 points) In the Hidden Markov Model with states $S_1, S_2, \ldots, S_T$ and observations $O_1, O_2, \ldots, O_T$, the following independence property holds: $O_t \perp O_{t+1} | S_t$ ("O_t independent of O_{t+1} given S_t").

TRUE or FALSE

Explanation:

Q1.4: (5 points) The optimal Bayesian network structure where each variable has at most <u>two</u> parents can be obtained in polynomial time.

TRUE or FALSE

Explanation:

Q2. Short Answer (3 questions, 30 points total)

Q2.1: (10 points) Bob the Bayesian has a 99% prior belief that a coin is fair (i.e., 50% probability of "heads") and a 1% prior belief that the coin is a trick coin that always lands on "heads." If Bob observes a sequence of four "heads" outcomes of coin flips, what is Bob's posterior probability that the coin is fair? (You need not compute the numerical answer; just write the equation that produces it.)

Q2.2: (10 points) Draw a set of positive ('+') and negative ('-') examples in the two-dimensional feature space for which the best decision tree of depth two makes no errors, while the best decision tree of depth one (one decision node, two leaves) makes as many errors as the best decision tree of depth zero (a single leaf).

Q2.3: (10 points) Given two classification methods and a training set of n examples: (a) describe how to accurately estimate which provides higher predictive accuracy for predictions on new data not in the training set; and (b) what assumptions are made about the training data for this to work.

Q3. Naïve Bayes (20 points total)

Consider the five-example dataset with label (Y) and three feature variables (X_1, X_2, and X_3):

X_1	X_2	X_3	Y
0	0	0	0
0	1	1	0
1	0	0	0
0	1	0	1
1	1	1	1

Q3.1: (7 points) What are the maximum likelihood estimates for the Naïve Bayes model fit from the dataset?

$\hat{P}(Y = 1) =$

$\hat{P}(X_1 = 1|Y = 0) =$

$\hat{P}(X_1 = 1|Y = 1) =$

$\hat{P}(X_2 = 1|Y = 0) =$

$\hat{P}(X_2 = 1|Y = 1) =$

$\hat{P}(X_3 = 1|Y = 0) =$

$\hat{P}(X_3 = 1|Y = 1) =$

Q3.2: (8 points) Using the estimated Naïve Bayes model from **Q3.1**:
(a) What is the joint probability of $\hat{P}(X_1 = 1, X_2 = 1, X_3 = 0, Y = 0)$?

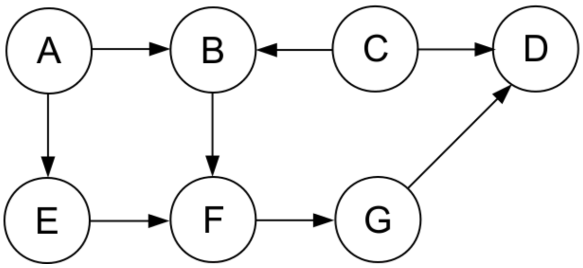
(b) What is the joint probability of $\hat{P}(X_1 = 1, X_2 = 1, X_3 = 0, Y = 1)$?

Q3.3: (5 points) Using the joint probabilities from **Q3.2**:
What is the label distribution estimate, $\hat{P}(Y = 1|X_1 = 1, X_2 = 1, X_3 = 0)$?

Q4. Bayesian Networks (30 points total)

Q4.1: (14 points) Independence Properties

Circle all of the independence properties that the Bayesian network implies and cross out all independence properties that are not implied.



$A \perp D$

$A \perp D|B$

$A \perp D|E$

$A \perp D|F$

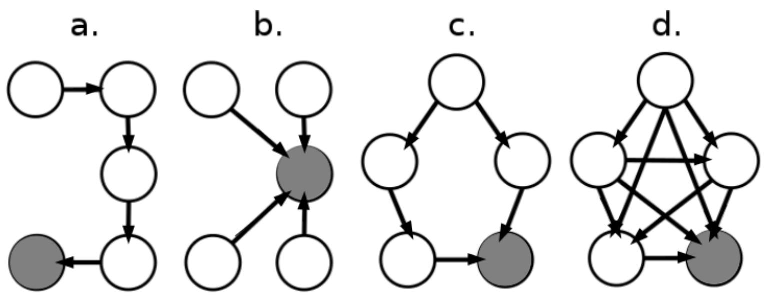
$A \perp D|G$

$A \perp D|B, E$

$A \perp D|E, F$

Q4.2: (16 points) Variable Elimination

Consider the following four Bayesian networks with observed variable shaded.



What is the time complexity that best characterizes variable elimination (using the best possible elimination order) on each of these graphs in terms of the number of variables, n, and the number of values each can take, $|X|$?

Choose from:
$O(n|X|), O(n^2|X|), O(n|X|^2), O(n^3|X|), O(n^2|X|^2), O(n|X|^3), O(n^{|X|})$, and $O(|X|^n)$ time.

a.

b.

c.

d.

Extra Space