CS 412: Introduction to Machine Learning
Fall 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 8 pages. Count the pages (without looking at the questions).
3. Read the instructions carefully. **Q1** asks for both a TRUE or FALSE answer <u>AND</u> a short explanation. 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 (5 questions, 20 points total)

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

Q1.1: (4 points) X independent of Y ($X \perp Y$) and Y independent of Z given X ($Y \perp Z|X$) implies that Y is independent of Z ($Y \perp Z$).

TRUE or FALSE. Explanation:

Q1.2: (4 points) Bayes Theorem, $P(y|x) = \frac{P(x|y)P(y)}{P(x)}$ for $P(x) > 0$, only holds sometimes and maximum likelihood estimation should be employed to compute $P(y|x)$ when it is not valid.

TRUE or FALSE Explanation:

Q1.3: (4 points) The decision tree of depth n that maximizes classification accuracy can be obtained in polynomial time (in terms of n).

TRUE or FALSE Explanation:

Q1.4: (4 points) If $X_3 \perp Y$, including X_3 as an input for the <u>naïve Bayes</u> model will never improve classification accuracy.

TRUE or FALSE Explanation:

Q1.5: (4 points) If $X_3 \perp Y$, including X_3 as an input for the <u>decision tree model</u> will never improve classification accuracy.

TRUE or FALSE Explanation:

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

Q2.1: (10 points) In one sentence, please describe "overfitting." **(5 points)**

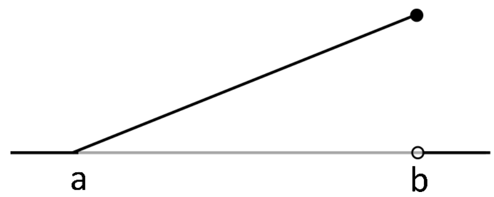
In another single sentence, please describe a situation in which it is likely to occur. **(5 points)**

Q2.2: (10 points) Draw a set of positive ('+') and negative ('-') examples in two dimensions so that the testing error (obtained by withholding some data and evaluating on the withheld data) of a decision tree of depth 3 is much worse than 1-Nearest Neighbor.

Q2.3: (10 points) Consider the joint distribution, $P(x, y, z)$ where each is a binary-valued variable. If $X \perp Y$, how many free parameters does this distribution have? (Hint: without this independence property there are $2^3 - 1$ free parameters.)

Q3. Parameter Estimation (20 points total)

Consider the "ramp" continuous probability distribution. It is defined by two parameters, a and b. For $x < a$, the probability density is 0. Between a and b, the probability density increases with a fixed slope until it is maximized at b. For $x > b$, the probability density is again 0.



Q3.1: (4 points) What is the probability density function of this distribution at $x = b$ (as a function of a and b)?

$f(x = b) =$

Q3.2: (4 points) Given two datapoints, x_1 and x_2, what is the maximum likelihood estimate for b? (Hint: no calculus is required.)

$\hat{b} =$

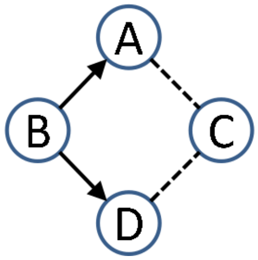
Q3.3: (4 points) If we estimated using mean of the Bayesian posterior and a reasonable prior, would this Bayesian mean estimate of b be <u>SMALLER</u> than the MLE estimate, <u>THE SAME</u> as the MLE estimate, or <u>LARGER</u> than the MLE estimate? Why?

Q3.4: (8 points) Given this maximum likelihood estimate for b, what is the likelihood function given the pair of datapoints, $P(x_1, x_2 | a, b)$?

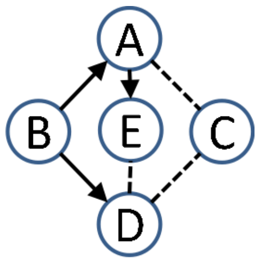
Q3.5: **(EXTRA CREDIT: 10 points)**
What is the maximum likelihood estimate for parameter a given the same two datapoints x_1 and x_2? (Hint: calculus is required and logarithms may be useful.)

Q4. Bayesian Networks (30 points total)

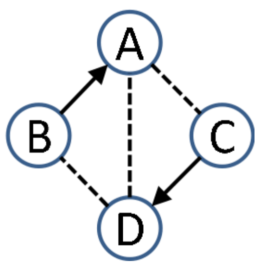
Q4.1: **(20 points) Independence Properties**
Draw directed edges for each of the undirected dotted edges to complete a Bayesian network so that the following independence properties ($\not\perp$ means not independent) hold or declare that it is IMPOSSIBLE to construct a Bayes Net with those independence properties.



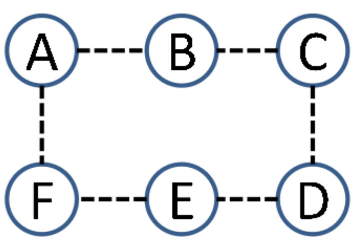
$B \perp C|D$



$B \perp C|D$

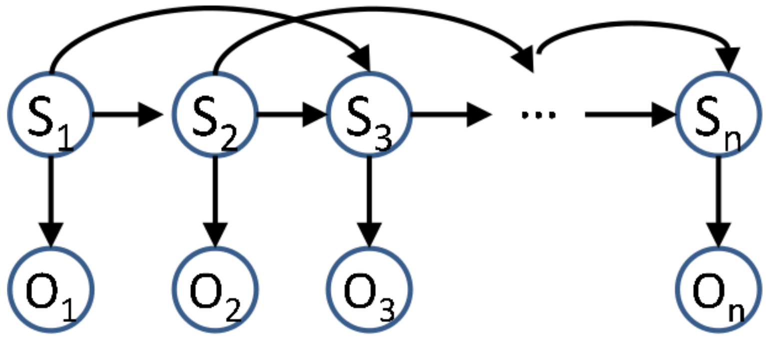


$B \perp C|A, D$



$A \perp D; B \not\perp E;$ and $C \not\perp F$

Q4.2: **(10 points) Variable Elimination**
Given a second-order Hidden Markov model with the following Bayesian network structure,



what is the time complexity that best characterizes using variable elimination to estimate $P(s_1, s_2, \ldots, s_n | o_1, o_2, \ldots, o_n)$ in terms of the number of variables, n, and the number of values each can take, $|\mathcal{S}|$?

Extra Space