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	

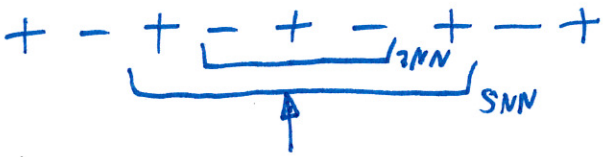
Q1. True or False (4 questions, 20 points total)

For each question: circle TRUE or FALSE (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:

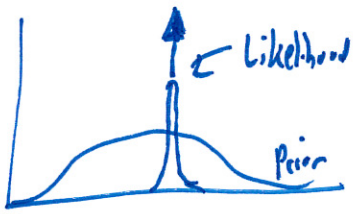


Q1.2: (5 points) The MAP estimate and the maximum likelihood estimate converge to the same solution given infinite data and a reasonable prior distribution (providing non-zero probability everywhere).

TRUE or FALSE

Explanation: MAP: $\log P(D|\theta)\,P(\theta)$

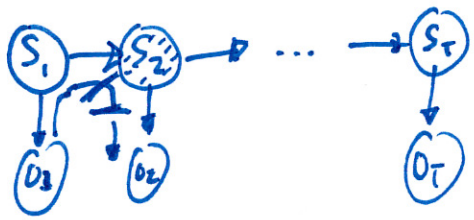
ML: $\log P(D|\theta)$



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 two parents can be obtained in polynomial time.

TRUE or FALSE

Explanation: NP-hard. Trees are poly time, but parents ≤ 1

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

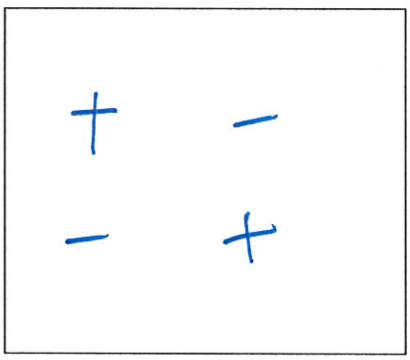
Q2.1: (10 points) Bob the Bayesian has a 99% prior belief that a coin is fair (i.e., 50% probability of "heads") and a 1% prior belief that the coin is a trick coin that always lands on "heads." If Bob observes a sequence of four "heads" outcomes of coin flips, what is Bob's posterior probability that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

P(fair) 99% P(heads | fair) 50%

P(trick) 1% P(heads | trick) 100%

$$P(\text{fair} \mid \text{heads}, \text{heads}, \text{heads}, \text{heads}) = \frac{P(\text{fair}, 4\times\text{heads})}{P(\text{fair}, 4\times\text{heads}) + P(\text{trick}, 4\times\text{heads})} = \frac{.99 \cdot .5^4}{.99 \cdot .5^4 + .01 \cdot 1^4}$$

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.

LOOCV, CV...

Withheld a testing dataset for eval...

→ IID

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?

$\hat{P}(Y = 1) =$ 2/5

$\hat{P}(X_1 = 1|Y = 0) =$ 1/3

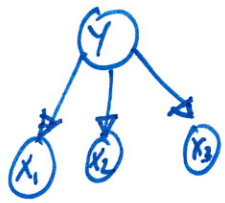
$\hat{P}(X_1 = 1|Y = 1) =$ 1/2

$\hat{P}(X_2 = 1|Y = 0) =$ 1/3

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

$\hat{P}(X_3 = 1|Y = 0) =$ 1/3

$\hat{P}(X_3 = 1|Y = 1) =$ 1/2



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

$= \hat{P}(Y=0) \cdot \hat{P}(X_1=1|Y=0)\ \hat{P}(X_2=1|Y=0)\ \hat{P}(X_3=0|Y=0)$

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

$= \hat{P}(Y=1)\ \hat{P}(X_1=1|Y=1)\ \hat{P}(X_2=1|Y=1)\ \hat{P}(X_3=0|Y=1)$

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

$P(Y=1, X_1, \ldots)$

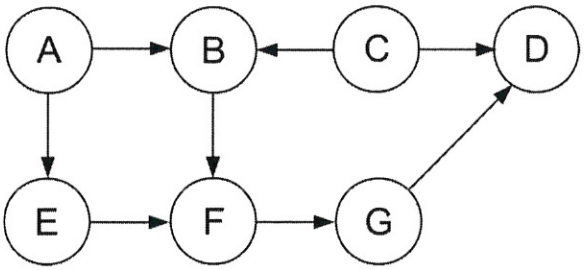
$=$

$P(Y=1, X_{1:3}=(1\ 1\ 0)) + P(Y=0, X_{1:3}=(1\ 1\ 0))$

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.



$A \perp D$

$A \perp D|B$

$A \perp D|E$

$A \perp D|F$

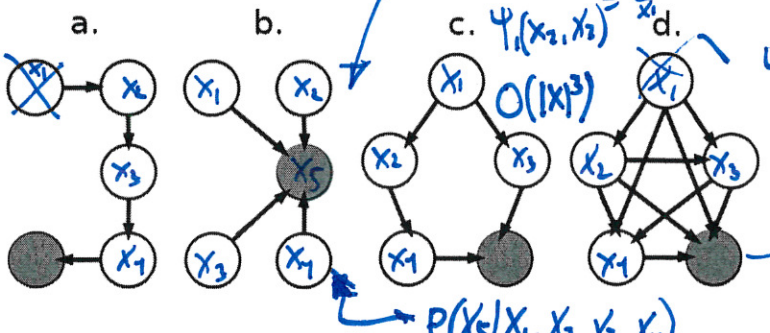
$A \perp D|G$

$A \perp D|B, E$

$A \perp D|E, F$

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. $O(n|X|^2)$

b. $O(|X|^{n-1})$

c. $O(n|X|^3)$

d. $O(|X|^{n-2})$

CS 491: Introduction to Machine Learning
Spring 2014
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 11 pages. Count the pages (without looking at the questions).
3. Q1 contains multiple choice problems. Circle every answer that you believe is correct.
4. Do not discuss the exam with students who have not taken the exam!

Some useful formulas:

- $P(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{Y}} P(\mathbf{x}, \mathbf{y})$ (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)
- $\mathbb{E}_{x \sim f}[g(X)] = \int_{x \in \mathcal{X}} f(x)g(x)dx$ (continuous 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	/20
Q3	/15
Q4	/20
Q5	/25
Total	

Q1. Multiple Choice (5 questions, 20 points total)
(Circle ALL correct answers)

Q1.1: (4 points) Consider a continuous-valued random variable, X, that can take values from the set $\mathcal{X}$. Which of the following is always true (i.e., for any choice of set $\mathcal{X}$ and distributions over the random variable X) for probability mass functions P and probability density functions f:

(a) $P(x) > 0$ for some $x \in \mathcal{X}$ — $P(X=x) = 0$

(b) $f(x) > 0$ for some $x \in \mathcal{X}$ — $f(x)$

(c) $P(X \in \mathcal{X}') \leq 1$ for all $\mathcal{X}' \subseteq \mathcal{X}$

(d) $f(x) \leq 1$ for all $x \in \mathcal{X}$

Q1.2: (4 points)
Assume you have a fair coin and flip it three times, X_1, X_2, X_3 are three random variable denote the result of each flip separately ($X_i = 1$ if the result of flip is head, $X_i = 0$ otherwise) which of the following is/are correct?

(a) $P(\min(X_1, X_2, X_3) = 0) < P(\max(X_1, X_2, X_3) = 1)$

(b) $P(X_1 < X_2 \leq X_3) = P(X_1 > X_2 \geq X_3)$

(c) $P(\max(X_1, X_2, X_3) > \min(X_1, X_2, X_3)) = 1$

(d) $P(X_1^2 + X_2^2 + X_3^2 = X_1 + X_2 + X_3) = 1$

Q1.3: (4 points) Consider maximum likelihood estimation (MLE), maximum a posteriori estimation (MAP), and Bayesian estimation (Bayes) with a prior with non-zero probability for all model parameters. Which of the following are true?

(a) MLE and MAP with a uniform prior are equivalent. — $\max_\theta P(D|\theta)$ $\max_\theta P(D|\theta)P(\theta)$ unif

(b) MLE and Bayes with a uniform prior are equivalent.

(c) With infinite amounts of data, MLE and Bayes will converge to the same estimates.

(d) With infinite amounts of data, MLE and MAP will converge to the same estimates.

Q1.4: (4 points) The Naïve Bayes classifier that predicts the class Y given the feature $X_1, ..., X_k$:

(a) uses the independence assumption $X_i \perp Y | X_j$ — F

(b) uses the independence assumption $X_i \perp X_j | Y$

(c) will always have better classification accuracy on the training set if more features are added

(d) tends to not overfit as badly when Bayesian parameter estimation is used rather than maximum likelihood

Q1.5: (4 points) Consider a set of random variables $X_1, X_2, \ldots, X_n$ each taking on $|\mathcal{X}|$ possible values. Which of the following is/are true of variable elimination (VE)?

(a) VE for worst-case Bayesian network graph structures takes $\mathcal{O}(|\mathcal{X}|^n)$ time

(b) VE for chain Bayesian network graphs takes $\mathcal{O}(|\mathcal{X}|^{2n})$ time

(c) Nodes with many parents and children should generally be eliminated first

(d) Finding the optimal VE ordering is NP-hard in general

Q2. Short Answer (4 questions, 20 points total)

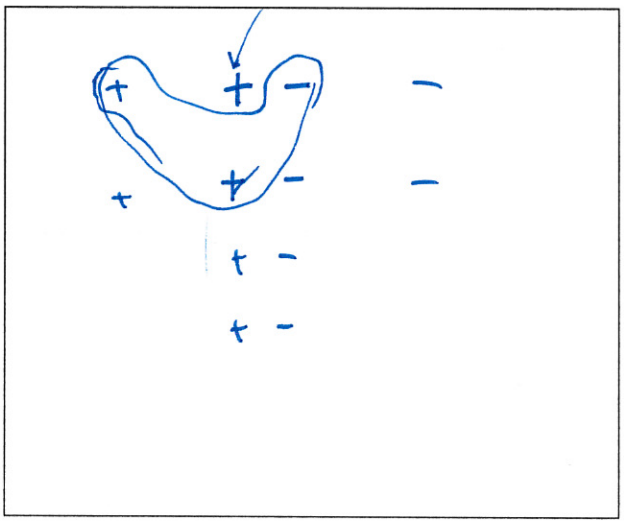
Q2.1: (5 points) Assume box A contains 2 black balls and 3 white balls, 3 of these 5 balls are selected randomly and put into box B which was originally empty. Then one ball is drawn randomly from box B.

What is the probability that the ball drawn from box B is black? (2 points)

P(WWW)
P(WWB)
P(WBB)

What is the probability that 1 black ball and 2 white balls were drawn from box A given the ball drawn from box B is black? (3 points)

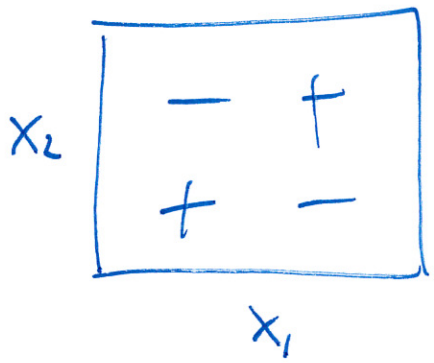
Q2.2: (5 points) Plot positive ('+') and negative examples ('-') so that one nearest neighbor (1-NN) will perform significantly worse than 3-NN when evaluated using leave-one-out cross-validation (LOOCV). Circle the examples that 3-NN LOOCV will correctly classify but 1-NN LOOCV will not.



Q2.3: (5 points) Consider a decision tree with input features $X_1, X_2, ..., X_n$ and class label Y. Is it true that if X_2 is independent of Y ($X_2 \perp Y$), then no decision based on X_2 will appear in the decision tree?

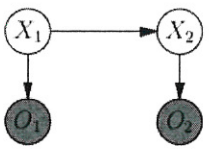
Yes / No (Circle one)

Argue why this is or is not the case.



$X_2 \not\perp Y \mid X_1$

Q2.4: (5 points) Consider the following Hidden Markov Model.



X_1	$\Pr(X_1)$
0	0.3
1	0.7

X_t	X_{t+1}	$\Pr(X_{t+1}\|X_t)$
0	0	0.4
0	1	0.6
1	0	0.8
1	1	0.2

X_t	O_t	$\Pr(O_t\|X_t)$
0	A	0.9
0	B	0.1
1	A	0.5
1	B	0.5

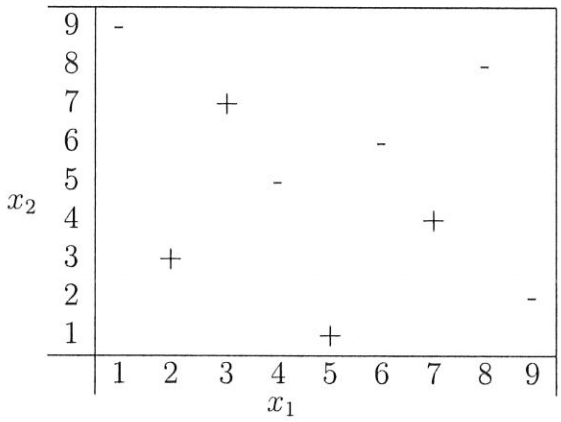
Suppose that $O_1 = A$ and $O_2 = B$ are observed. What is the most likely pair of hidden state values (i.e., $\text{argmax}_{x_1,x_2} P(X_1 = x_1, X_2 = x_2 | O_1 = A, O_2 = B)$)?

$O_1 = A \quad O_2 = B$

X_1	X_2	$P(X_1=x_1, X_2=x_2, O_1=A, O_2=B)$
0	0	$.3 \cdot .9 \cdot .4 \cdot .1$
0	1	$.3 \cdot .9 \cdot .6 \cdot .5$ ✓
1	0	$.7 \cdot .5 \cdot .8 \cdot .1$
1	1	$.7 \cdot .5 \cdot .2 \cdot .5$

Q3. Decision Tree (15 points total)

Q3.1: (5 points) Consider the dataset of positive (+') and negative ('-') examples:

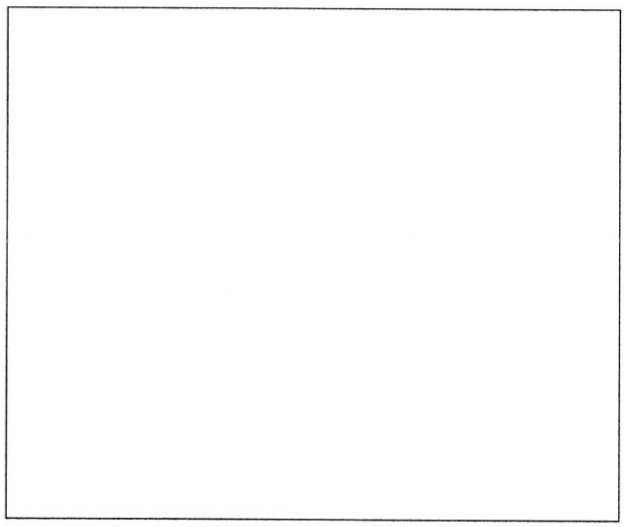


Draw the decision boundaries for a decision tree that is greedily selected using (all of the following):

- classification accuracy as the decision criterion rule;
- ties can be broken as you wish;
- thresholds in either feature dimension for defining the decision splits; and
- continues until reaching perfect classification accuracy.

Q3.2: (5 points) For the decision boundaries in **Q3.1**, draw the corresponding decision tree with decisions (e.g., $x_1 < 2.5$) in each node and prediction labels for each decision tree leaf.

Q3.3: (5 points) Draw a binary dataset in the two-dimensional feature space for which greedily choosing decisions based on classification accuracy will produce bad results, while choosing decisions based on the impurity of the decision split will perform significantly better. Explain why this is the case.



Q4. Statistical Estimation (20 points total)

For the following problems, consider the geometric distribution, $P_\theta(x) = \theta(1-\theta)^x$, for $x \in \{0, 1, 2, \ldots\}$ and given parameter $\theta \in [0, 1]$. It has mean $\frac{1-\theta}{\theta}$ and mode 0. Three i.i.d. datapoints x_1, x_2, x_3, are assumed to be drawn from the geometric distribution $P_\theta(x)$,

Q4.1: **(5 points)** What is the maximum likelihood estimate $\hat{\theta}$ in terms of x_1, x_2, x_3? (Hint: Start by writing the [log-]likelihood.)

$$\log \prod_{i=1}^{3} \theta(1-\theta)^{x_i} = \sum_{i=1}^{3} \log\theta + x_i \log(1-\theta)$$

$$= 3\log\theta + (\Sigma x_i)\log(1-\theta)$$

$$\frac{3}{\theta} + \frac{-(\Sigma x_i)}{1-\theta} = 0$$

$$3 - 3\theta - \theta\Sigma x_i = 0$$

$$3 = (3 + \Sigma x_i)\theta$$

$$\theta = \frac{3}{3 + \Sigma x_i}$$

Q4.2: **(5 points)** The Beta distribution is the conjugate prior of the geometric distribution. It has probability density function:

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

What are the parameters of the posterior Beta distribution: $\theta|x_1, x_2, x_3 \sim Beta(\alpha', \beta')$ given prior distribution $Beta(\alpha, \beta)$? (Hint: Ignore the constant terms.)

$$\prod_{i=1}^{3} \theta(1-\theta)^{x_i} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$= \theta^3(1-\theta)^{\Sigma x_i} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{\alpha+3-1}(1-\theta)^{\Sigma x_i+\beta-1}$$

$\alpha' =$ $\alpha + 3$

$\beta' =$ $\Sigma x_i + \beta$