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)
- $X \sim \text{Normal}(\mu, \sigma) \implies P(X = x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- $X \sim \text{Multinoulli}(\theta) \implies P(\mathbf{x}) = \prod_{i=1}^{K} \theta_i^{x_i}$

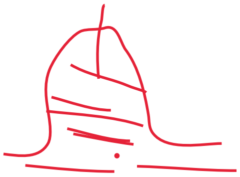
	Points
Q1	/20
Q2	/20
Q3	/15
Q4	/20
Q5	/25
Total	

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 $x \in \mathcal{X}$
(b) $f(x) > 0$ for some $x \in \mathcal{X}$
(c) $P(X \in \mathcal{X}') \leq 1$ for all $\mathcal{X}' \subseteq \mathcal{X}$
(d) $\boxed{f(x) \leq 1}$ for all $x \in \mathcal{X}$



$\frac{1}{\sqrt{2\pi\sigma}}$

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?

$P(\theta)$

(a) MLE and MAP with a uniform prior are equivalent.
(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.

MAP: $\max_\theta \left[\prod_i P(x_i|\theta)\right] P(\theta)$

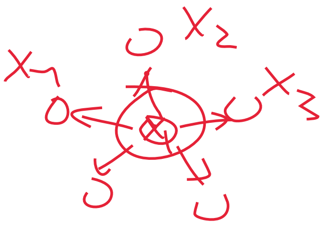
MLE $\max_\theta \prod P(x_i|\theta)$

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 $X_i \perp Y | X_j$
(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 $\mathcal{O}(|\mathcal{X}|n^2)$ time
(c) Nodes with many parents and children should generally be eliminated first
(d) Finding the optimal VE ordering is NP-hard in general



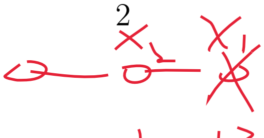
2^{n-1}

for X_2

for X_1 . .

for X_n

sum = 0



(X).

for x,

→ sum += ——

record(sum, x_2, x · x_b)

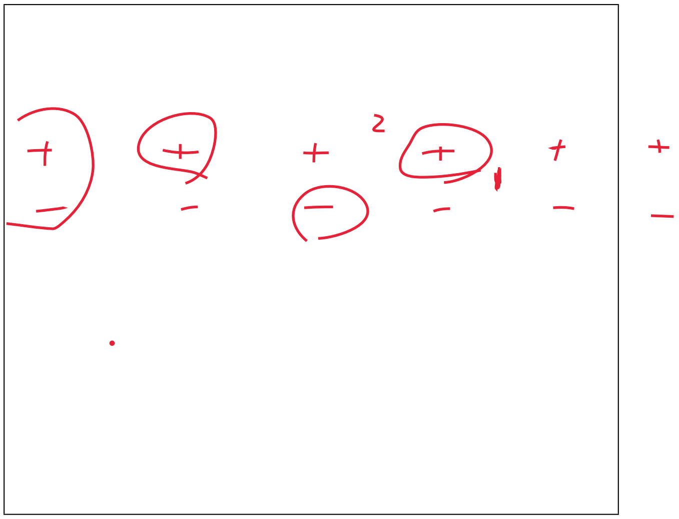
Q2. Short Answer (4 questions, 20 points total)

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)

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)

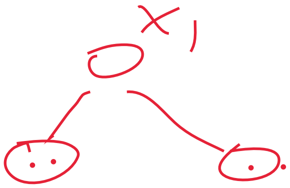
Q2.2: (5 points) Plot positive ('+') and negative examples ('-') so that one nearest neighbor (1-NN) will perform significantly worse than 3-NN when evaluated using leave-one-out cross-validation (LOOCV). Circle the examples that 3-NN LOOCV will correctly classify but 1-NN LOOCV will not.



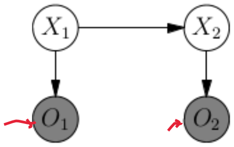
Q2.3: (5 points) Consider a decision tree with input features $X_1, X_2, ..., X_n$ and class label Y. Is it true that if X_2 is independent of Y ($X_2 \perp Y$), then no decision based on X_2 will appear in the decision tree?

Yes / No (Circle one)

Argue why this is or is not the case.



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., $\text{argmax}_{x_1,x_2} P(X_1 = x_1, X_2 = x_2 | O_1 = A, O_2 = B)$)?

$\propto$

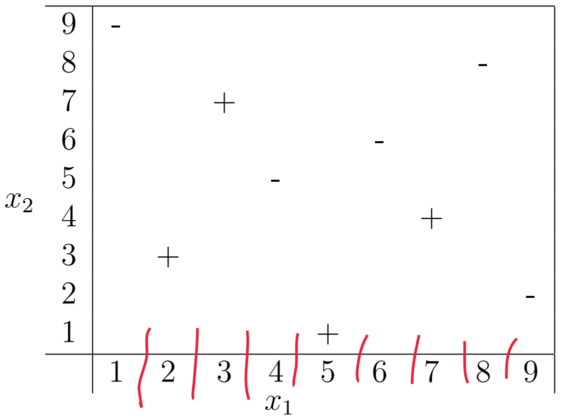
$$P(X_1,X_2|O_1.O_2) \propto P(X_1.X_2.O_1.O_2) \cancel{P(O_1,O_2)}$$

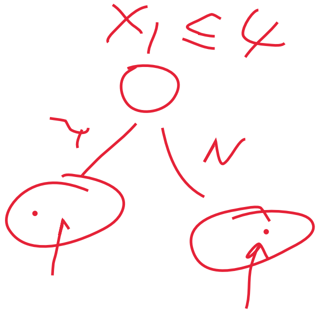
$$= P(X_1)P(X_2|X_1)P(\underset{A}{O_1}|X_1)P(\underset{B}{O_2}|X_2)$$

X_1	X_2
0	0
0	1
1	0
1	1

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.)

$\boxed{P(X) = \theta(1-\theta)^X}$ $X \in \{0, 1, 2, \cdots\}$

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

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

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.)

$$P(\theta | X_1 X_2 X_3) \propto P(X_1 X_2 X_3 | \theta) P(\theta)$$
$$= \prod_{i=1}^{3} P(X_i | \theta) P(\theta)$$
$$= \theta(1-\theta)^{X_1} \theta(1-\theta)^{X_2} \theta(1-\theta)^{X_3} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$
$$= \theta^{\alpha+3-1} (1-\theta)^{X_1+X_2+X_3+\beta-1}$$

$Beta(\alpha'. \beta')$ $\alpha' = \alpha + 3$, $\beta' = X_1 + X_2 + X_3 + \beta$

$\alpha' =$

$\beta' =$

A $Beta(\alpha, \beta)$ distribution has a mean value of $\frac{\alpha}{\alpha+\beta}$ and a mode of $\frac{\alpha-1}{\alpha+\beta-2}$.

Q4.3: **(5 points)** What is the expected value of a new datapoint x_4 given x_1, x_2, x_3 using maximum a posteriori estimation? Write your answer in terms of the posterior parameters α', β'.

$$\theta_{MAP} = \frac{\alpha'-1}{\alpha'+\beta'-2}$$

$$E[X_4]$$

$$P(X_4|\theta_{MAP}) = \theta_{MAP}(1-\theta_{MAP})^{X_4}$$

$$X_4 \sim P(X_4|X_1 \cdots X_3)$$

$$= \int_{X_4} X_4 P(X_4|\theta_{MAP}) dX_4$$

$$= \int_{X_4} X_4 \theta_{MAP}(1-\theta_{MAP})^{X_4} dX_4 = \frac{1-\theta_{MAP}}{\theta_{MAP}}$$

Q4.4: **(5 points)** What is the probability that a new datapoint has value 0 given x_1, x_2, x_3 using a full Bayesian treatment (i.e., $P(X_3 = 0|x_1, x_2, x_3)$ using Bayesian posterior prediction)? Write your answer in terms of the posterior parameters α', β'. (Hint: $P(X_3 = 0|x_1, x_2, x_3) = \int_\theta P_\theta(X_3 = 0)P(\theta|x_1, x_2, x_3)d\theta$.)

$$\int_\theta P(X_4=0|\theta)\, P(\theta|X_1 \cdots X_3)\, d\theta$$

$$= \int \theta(1-\theta)^0 Beta(\alpha', \beta')\, d\theta$$

$$= \frac{\alpha'}{\alpha'+\beta'}$$

Q5. Bayesian Networks (25 points total)

Consider the following joint probability distribution table:

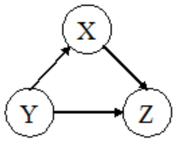
X	Y	Z	p(X,Y,Z)
0	0	0	$\frac{1}{16}$
0	0	1	$\frac{1}{16}$
0	1	0	$\frac{1}{8}$
0	1	1	$\frac{1}{4}$
1	0	0	$\frac{1}{16}$
1	0	1	$\frac{1}{16}$
1	1	0	$\frac{1}{4}$
1	1	1	$\frac{1}{8}$

For each of the following Bayesian networks, is it consistent with the probability distribution shown in the table (i.e., can the probability distribution in the table be represented using the Bayesian network)?

If so, list one independence property that the table possesses that the Bayesian network does not (if one exists).

If not, list one independence property that the Bayesian network possesses that the table does not.

Q5.1: (5 points) Bayesian Network 1

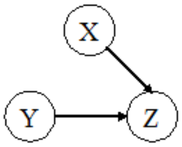


Consistent Not Consistent (circle one)

Unshared independence property:

Consistent because the graph does not enforce any consistency.

Q5.2: (5 points) Bayesian Network 2

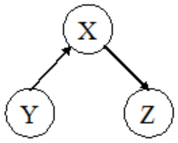


Consistent Not Consistent (circle one)

Unshared independence property:

Just need to check if X is independent of Y

Q5.3: (5 points) Bayesian Network 3

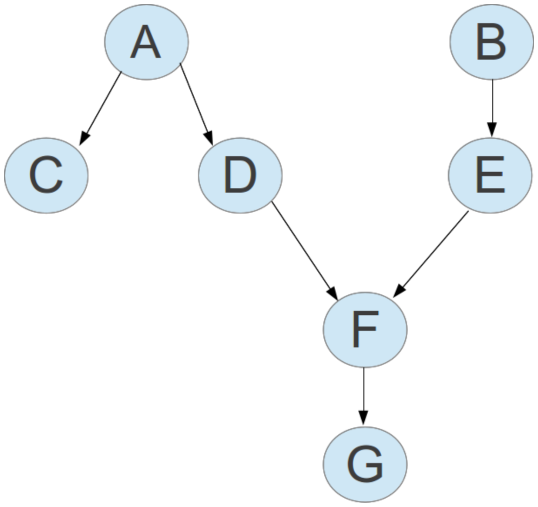


Consistent Not Consistent (circle one)

Unshared independence property:

Just check if Y is conditionally independent of Z given X

Q5.5: (10 points) Consider the Bayesian network for the following set of questions:



True or False, $A \perp B$?

True or False, $A \perp E|D$?

True or False, $B \perp C|G$?

True or False, $B \perp D$?

True or False, $C \perp G$?

Extra space