CS 491: Introduction to Machine Learning
Spring 2013
Midterm Exam

Name: ________________________

User ID: ________________________

Instructions:

1. Write your name and user ID above. Do not begin the exam (look at other pages) until told to do so.
2. There should be 9 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:

- $P(\mathbf{x}) = \sum_{\mathbf{y}} P(\mathbf{x}, \mathbf{y})$ (marginalization)
- $P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{x},\mathbf{y})}{P(\mathbf{y})}$ (conditioning)
- $P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}{P(\mathbf{y})}$ (Bayes theorem)
- $\mathbb{E}_{x \sim P}[f(X)] = \sum_x P(x) f(x)$ (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	/15
Q5	/15
Total	

Q1. Multiple Choice (10 questions, 20 points total)

(Circle ALL correct answers)

Q1.1: (2 points) Which of the following techniques is/are applicable for supervised learning settings?
(a) K-nearest neighbors
(b) Naïve Bayes
(c) K-means
(d) Decision Trees

Q1.2: (2 points) If $X \perp Y|Z$ (i.e., X is independent of Y given Z), which of the following is/are implied?
(a) $P(X, Y|Z) = P(X|Z)P(Y|Z)$
(b) $P(X, Y|Z) = P(X|Y)P(Y|Z)$
(c) $P(X|Y, Z) = P(X|Z)$
(d) $P(Z|X, Y) = P(Z|X)$

Q1.3: (2 points) The parameters $\theta = \{\hat{P}(Y), \hat{P}(X|Y)\}$ obtained by maximizing the likelihood of parameters for a set of i.i.d. data $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ...\}$ (maximum likelihood estimation), $\text{argmax}_\theta\, P(\mathcal{D}|\theta)$, are equivalent to:
(a) The parameter obtained by maximizing the log likelihood, $\text{argmax}_\theta \log P(\mathcal{D}|\theta)$
(b) $\hat{P}(Y) = \text{argmax}_{P(Y)} \sum_i \log P(y_i)$, $\hat{P}(X|Y) = \text{argmax}_{P(X|Y)} \sum_i \log P(x_i|y_i)$
(c) The parameters obtained using maximum a posteriori (MAP) estimation with a uniform prior distribution $P(\theta)$
(d) $\text{argmax}_\theta\, \mathbb{E}_{(x,y)\sim\mathcal{D}}[\log P(X, Y|\theta)]$ where $(x, y) \sim \mathcal{D}$ is the empirical distribution of the (training) data

Q1.4: (2 points) The expectation-maximization (EM) algorithm:
(a) is an iterative method
(b) will converge to a global optima for maximizing the marginal data likelihood $\prod_{i=1}^N \sum_{y_i\in\mathcal{Y}} P(x_i, y_i)$
(c) optimizes a lower bound on the marginal data likelihood, $\prod_{i=1}^N \sum_{y_i\in\mathcal{Y}} P(x_i, y_i)$, in the M-step
(d) cannot be applied to semi-supervised learning problems

Q1.5: (2 points) Which of the following are good ways to limit over-fitting in the Decision Tree classifier?
(a) Prune some of the branches of an over-fit tree
(b) Stop growing the tree at nodes with a small number of examples
(c) Use a subset of the training data to construct the decision tree
(d) Randomize the predictions in the tree leaves.

Q1.6: (2 points) If a family of distributions (e.g., Gaussian or Dirichlet) is a conjugate prior to a likelihood function, this implies that:
(a) The posterior distribution, $P(\theta|\mathcal{D})$, is also a member of the likelihood function's family of distributions
(b) The posterior distribution, $P(\theta|\mathcal{D})$, is also a member of that prior's family of distributions
(c) The likelihood function is also a member of the prior's family of distributions
(d) The maximum a posteriori (MAP) estimate and the maximum likelihood estimate (MLE) are the same

Q1.7: (2 points) Assuming $\mathcal{P} \neq \mathcal{NP}$, finding the most probable Bayesian Network graph structure (using a decomposable score function) for a given dataset $\mathcal{D}$ can be accompished with a polynomial time algorithm (in the number of variables) for:
(a) Tree structures
(b) All directed acyclic graphs
(c) A fixed number of parents $k \geq 2$
(d) A fixed number of parents $k \geq 2$ with an ordering such that $\forall i \in \{1, ..., N\}, \text{parents}(x_i) \subseteq (x_1, x_2, ..., x_{i-1})$

Q1.8: (2 points) To evaluate the usefulness of a classifier, the best metric to use is (circle one):
(a) Classification accuracy on the training dataset
(b) Classification accuracy on a withheld dataset not used for training
(c) Classification accuracy on a dataset created by combined the training dataset with all additional withheld data
(d) Classification accuracy on a random subset of the training dataset

Q1.9: (2 points) For the hidden Markov model where each of the hidden variables $X_1, X_2, ..., X_T$ can take on $|\mathcal{X}|$ different values, the most accurate run time characterization for the Forwards-Backwards algorithm is (circle one):
(a) $O(|\mathcal{X}|^T)$
(b) $O(T \log |\mathcal{X}|)$
(c) $O(T|\mathcal{X}|^2)$
(d) $O(T \log T|\mathcal{X}|^2)$

Q1.10: (2 points) Bayesian statistical methods are characterized by:
(a) The expectation-maximization (EM) algorithm
(b) Treating model parameters as random variables with probability distributions
(c) The use of Bayes Theorem to obtain a posterior probability
(d) Incorporating a prior probability distribution

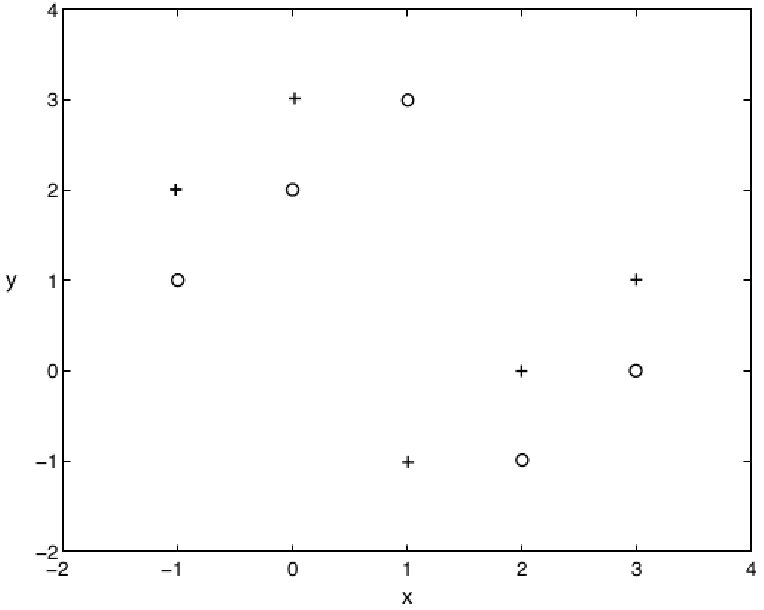
Q2. Short Answer (5 questions, 30 points total)

Q2.1: (6 points) If two binary random variables X and Y are independent, are $\bar{\text{X}}$ ($\bar{\text{X}}$ is the complement of X) and Y also independent? Prove your claim.

Q2.2: (6 points) The naïve Bayes classifier employs Bayes theorem to predict $P(y|\mathbf{x}_{1:K})$ as a function of $P(x_i|y)$ probabilities and prior $P(y)$. What are the advantages of this approach over directly estimating $P(y|\mathbf{x}_{1:K})$? (Consider conditional Multinoulli distributions for each conditional probability distribution.)

Q2.3: (6 points) Consider a Bayesian network for 4 variables: A, B, C, D. The random variables A, B, and C are binary. The random variable D can take on 5 values. What is the largest number of free parameters the Bayesian network can have if there are at most 3 edges? Draw the Bayesian network and write the number of parameters. (Hint: since probabilities must sum to one, a binary distribution has one free parameter.)

Q2.4: (6 points) Consider the following dataset with '+' and 'o' classes.



For each data point, consider a K-nearest neighbor classifier that is trained using all the other data, except for that data point, and then used to predict the label of the withheld data point.
What is the average classification accuracy when $K = 1$?

What is the average classification accuracy when $K = 3$?

Q2.5: (6 points) Consider a Naïve Bayes model, $P(y|x_{1:K}) \propto P(y) \prod_{i=1:K} P(x_i|y)$. If multiple copies of a useful, predictive variable x_i are mistakenly used as input features, i.e., $x_i = x_{i+1} = x_{i+2} = x_{i+3}$, will the resulting prediction probabilities tend to be more confident (i.e., closer to 0 or 1) or less confident (i.e., closer to uniform) that the model when those redundant input features are omitted? Explain.

Q3. Parameter Estimation (20 points total)

Consider binary-valued data, $x_i \in \{0, 1\}$, distributed according to a Bernoulli distribution (i.i.d.),

$$P(X_i = x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}, \tag{1}$$

and a prior probability distribution,

$$\begin{aligned} P\left(\theta = \frac{1}{4}\right) &= \frac{1}{4} \\ P\left(\theta = \frac{1}{2}\right) &= \frac{1}{2} \\ P\left(\theta = \frac{3}{4}\right) &= \frac{1}{4}, \end{aligned}$$

and a dataset $\mathcal{D} = \{x_1, x_2\} = \{1, 1\}$ (i.e., two "1" examples).

Q3.1: (5 points) What is the maximum likelihood estimate for θ given the dataset $\mathcal{D}$?

Q3.2: (5 points) What is the Bayesian posterior probability, $P(\theta|\mathcal{D})$ for each value of θ: $\frac{1}{4}, \frac{1}{2}, \frac{3}{4}$?

Q3.3: (5 points) What is the maximum a posteriori (MAP) estimate for θ given the dataset $\mathcal{D}$ and prior probability distribution?

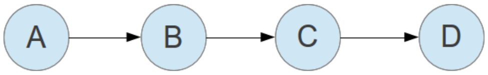
Q3.3: (5 points) What is the Bayesian posterior prediction of an additional data point, $P(x_3|\mathcal{D})$?

Q4. Bayesian Networks (15 points total)

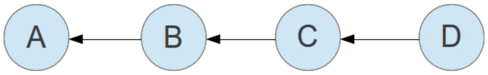
For each of the following pairs of Bayesian networks, indicate whether the two express the same independence properties. If they are not the same, write one independence property they disagree on. Otherwise, write an independence property that they share.

Q4.1: (5 points)

Bayesian Network 1



Bayesian Network 2

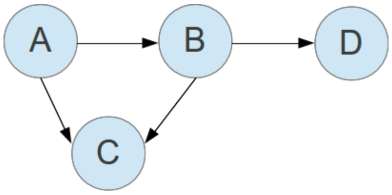


Same independence properties (circle one): Yes No

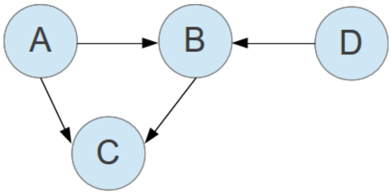
One independence property (shared if Same, disagree on if Not Same):

Q4.2: (5 points)

Bayesian Network 1



Bayesian Network 2

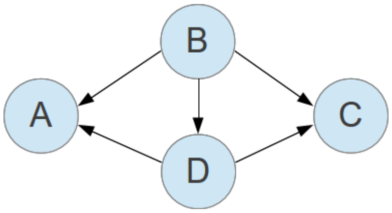


Same independence properties (circle one): Yes No

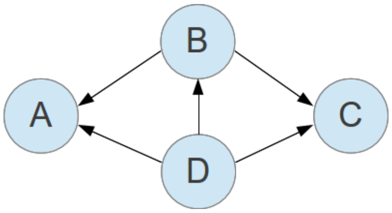
One independence property (shared if Same, disagree on if Not Same):

Q4.3: (5 points)

Bayesian Network 1



Bayesian Network 2



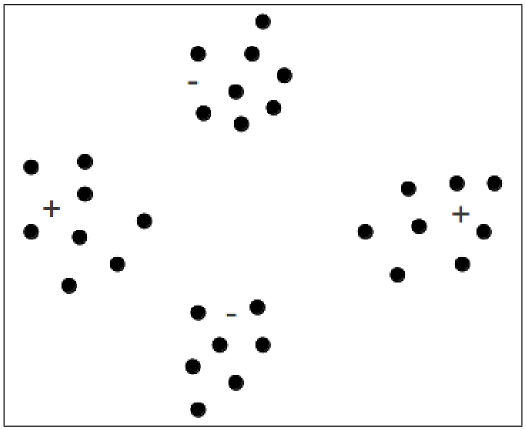
Same independence properties (circle one): Yes No

One independence property (shared if Same, disagree on if Not Same):

Q5. Semi-Supervised Clustering (15 points)

Consider a Naïve Bayes Gaussian mixture model (also known as a mixture model with independent x variables), i.e., $P(x_{1:K}|y) = \prod_{i=1}^{K} P(x_i|y)$ and $x_i|y \sim \text{Normal}(\mu_{y,i}, \sigma^2_{y,i})$.

Q5.1 (7 points) Consider the following dataset.



Initializing the Gaussian estimates with only the labeled examples (two - examples and two + examples), to what two clusters will the EM algorithm converge? Draw them on the figure (i.e., draw the regions that 95% of their probability will cover).

Q5.2 (8 points) Draw a semi-supervised dataset (using dots for unlabeled points, '+' and '-' for labeled points, as in the figure above) and draw two clusters with solid lines that EM could converge to that are local optima. Draw the global optimal clusters with dotted lines. Explain what this local optimality means.

Extra space