CS 6375 **Name (Print):** _____
Fall 2018
Final Exam
12/12/2018

This exam contains 15 pages (including this cover page) and 6 problems. Check to see if any pages are missing. Enter all requested information on the top of this page, and put your initials on the top of every page, in case the pages become separated.

You may **NOT** use books, notes, or any electronic devices on this exam. Examinees found to be using any materials other than a pen or pencil will receive a zero on the exam and face possible disciplinary action.

The following rules apply:

- **Organize your work**, in a reasonably neat and coherent way, in the space provided. Work scattered all over the page without a clear ordering will receive very little credit.
- **To ensure maximum credit** on short answer / algorithmic questions, be sure to **EXPLAIN** your solution.
- **Problems/subproblems** are not ordered by difficulty.
- **Do not** write in the table to the right.

Problem	Points	Score
1	24	
2	6	
3	19	
4	15	
5	16	
6	20	
Total:	100	

1. **True or False and Explain:** For each of the following statements indicate whether or not they are true or false and explain your reasoning. Simply writing true or false without correct reasoning will receive no credit.

   (a) (4 points) Suppose that you are fitting a Gaussian distribution to data. MAP estimation and maximum likelihood estimation result in the same parameters when the number of independently sampled data observations goes to infinity, independent of the choice of prior.

   (b) (4 points) Consider two random variables $X, Y \in \{0, 1\}$ representing independent coin flips of possibly biased coins. The only nonzero entries of the empirical covariance matrix of X and Y are the diagonal entries, which must be strictly larger than zero.

   (c) (4 points) If a given hypothesis space has VC dimension d, then there exists a set of d points such that any active learner requires d label queries, in the worst case, in order to produce a perfect classifier for these d data points.

(True and False continued)

(d) (4 points) A function $f : \mathbb{R} \to \mathbb{R}$ is quasiconvex if $f(\lambda x + (1-\lambda)y) \leq \max\{f(x), f(y)\}$ for all $x, y \in \mathbb{R}$ and all $\lambda \in (0, 1)$. Every convex function $g : \mathbb{R} \to \mathbb{R}$ is also quasiconvex.

(e) (4 points) If actions are only selected greedily in Q-learning, the Q-learning strategy cannot converge to the optimal value function.

(True and False continued)

(f) (4 points) Consider a feedforward neural network structure with 10 binary inputs, one binary output, and a single layer of 20 hidden units. Further, suppose that the activation function on the hidden units are relu's and the activation on the output is a perceptron. Let the hypothesis space H consist of every neural network that can be obtained from this structure for each different choice of the weights and biases. The VC dimension of H is at most 21.

2. **Gaussian Mixtures:** Consider fitting a K component Gaussian mixture to data points $x^{(1)}, \ldots, x^{(M)} \in \mathbb{R}^n$.

   (a) (3 points) Is there a unique setting of the mixture weights that maximizes the likelihood? Explain.

   (b) (3 points) A random initialization for the EM algorithm for Gaussian mixtures might result in a poor local optimum. How might you initialize the EM algorithm in practice to help avoid this?

3. **Laplace Maximum Likelihood:**
The Laplace distribution is a probability distribution over the real numbers defined by two parameters $\mu, b \in \mathbb{R}$ such that $b > 0$ with probability density $p(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$.

   (a) (6 points) Given data samples $x^{(1)}, \ldots, x^{(M)} \in \mathbb{R}$, compute the maximum likelihood estimator for μ and b.

   Hint: use $\frac{d}{dx}|x| = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}$.

(Laplace Maximum Likelihood continued)

(b) (8 points) Consider adding a prior probability distribution $p(\mu) = \lambda \exp(-\lambda\mu)$ for some integer λ.

i. What is the MAP estimate of μ under this prior for $\lambda = 1$.

ii. How does the choice of λ affect the MAP estimate of μ? How should you pick λ in practice?

(Laplace Maximum Likelihood continued)

(c) (5 points) Is the maximum likelihood estimator for μ unbiased? Hint: the Laplace distribution is symmetric about μ.

4. **Tic-Tac-Toe:** Suppose that you wanted to design a strategy for playing tic-tac-toe versus an adversary using reinforcement learning.

   (a) (10 points) Explain how to formulate this as a Markov decision process (the transitions can be deterministic or stochastic). You must specify all parts of the MDP for full credit. For simplicity, you can assume that you always place X's and that your opponent always places O's.

(Tic-Tac-Toe continued)

(b) (5 points) Explain how to pick the discount factor γ so that an optimal policy for your MDP results in a strategy that cannot lose, i.e., at worst a game can end in a draw.

5. **Uniform Estimation:** Consider the uniform distribution, i.e., a constant probability density function, over the interval $[0, \theta]$ for some $\theta \in \mathbb{R}_{\geq 0}$. Given M data points $x^{(1)}, \ldots, x^{(M)} \in \mathbb{R}$, consider the following questions.

   (a) (1 point) What is the probability density function for the uniform distribution as a function of θ?

   (b) (2 points) What is the log-likelihood? Is it a concave function of θ?

(Uniform Estimation continued)

(c) (3 points) What is the maximum likelihood estimator for θ?

(d) (5 points) Using that fact that for any continuous random variable X with differentiable cumulative distribution function $Pr(X \leq x)$, the probability density function $p(x) = \frac{d}{dx}Pr(X \leq x)$, show that the maximum likelihood estimator of θ is biased. How would you make it unbiased?

(Uniform Estimation continued)

(e) (5 points) Let's call a parameterized distribution, $p(\cdot|\theta)$ for some $\theta \in \Theta$, MLE PAC-learnable if there exists $f(\epsilon, \delta)$ polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$ and independent of θ such that for every $\epsilon, \delta \in (0, 1)$ and $\theta \in \Theta$, if $M \geq f(\epsilon, \delta)$ samples are drawn independently from $p(\cdot|\theta)$, then $|\theta - \theta_{MLE}| \leq \epsilon$ with probability at least $1 - \delta$. Is the uniform distribution described above MLE PAC learnable? If so, provide the function f.

6. **Short Answer:**

(a) (5 points) Consider a binary classification problem for points in $\mathbb{R}$ with a hypothesis space consisting of intervals $[a, b]$ for $a \leq b \in \mathbb{R}$ such that all points inside the interval are labeled plus and all points outside the interval are labeled minus. Even if the data can be perfectly classified, explain why an active learner could, in the worst case, require labels for all of the data points.

(b) (5 points) Consider a binary classification problem for points in $\mathbb{R}$ with a parameterized hypothesis space consisting of functions $h(x|\theta) = +1$ if $\sin(\theta x) \geq 0$ and -1 otherwise, for each $\theta \in \mathbb{R}$. What is the VC dimension of this hypothesis space on $\mathbb{R}$? Explain.

(Short Answer continued)

(c) (5 points) Explain why it is important to use a prior when fitting a naïve Bayes classifier in practice.

(d) (5 points) Suppose you are fitting a model to data observations. Explain why applying PCA to the data before fitting the model might make the learned model less "interpretable".