CS 6375
Fall 2017
Final Exam
12/13/2017

Name (Print): ______________________________

This exam contains 10 pages (including this cover page) and 5 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	30	
2	20	
3	20	
4	10	
5	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) (5 points) Logistic regression effectively computes a linear separator, if one exists, but does not necessarily return a max-margin classifier.

   (b) (5 points) Every matrix $A \in \mathbb{R}^{n\times n}$ admits a decomposition of the form $A = QDQ^T$ for a diagonal matrix $D \in \mathbb{R}^{n\times n}$ and some orthogonal matrix $Q \in \mathbb{R}^{n\times n}$, i.e., the columns of Q are orthonormal.

   (c) (5 points) If the Q-value function is represented exactly as a multidimensional array, then Q-learning with an ϵ-greedy strategy converges to a global optimum, i.e., the same one attained by value iteration, for all $\epsilon \in (0, 1]$ and learning rates $\alpha \in (0, 1)$.

(True and False continued)

(d) (5 points) A mixture of Gaussian mixtures is itself a Gaussian mixture.

(e) (5 points) The VC dimension of the hypothesis space of Gaussian naïve Bayes classifiers for data points $x \in \mathbb{R}$ and binary class labels, i.e., $p(x|y)$ is a normal distribution for each value of y, is at least three.

(f) (5 points) Consider fitting a K-component Gaussian mixture to data points in $\mathbb{R}^n$. If the means and variances are known, and hence fixed during learning, the maximum likelihood estimation problem is concave in the remaining parameters.

2. **Mixtures of Poisson Distributions:** Consider a nonnegative, integer-valued random variable X that is distributed according to a K-component mixture of Poisson distributions: $X \sim \sum_{k=1}^{K} p_k \frac{\lambda_k^x e^{-\lambda_k}}{x!}$ for real-valued parameters $\lambda_1, \ldots, \lambda_K > 0$ and $p_1, \ldots, p_K \in [0, 1]$ such that $\sum_k p_k = 1$.

   (a) (5 points) Given integer data points $x^{(1)}, \ldots, x^{(M)}$, what is the log-likelihood of this data under the K-component mixture of Poisson distributions above? What is the lower bound optimized by the EM algorithm as a function of the distributions $q_m(k)$ introduced for the m^{th} sample?

   (b) (7 points) What is the optimal value of $q_m(k)$ computed in the E-step of EM for a fixed setting of the parameters?

(Mixtures of Poisson Distributions continued)

(c) (8 points) What are the value of the parameters $\lambda_1, \ldots, \lambda_K$ and $p_1, \ldots, p_K$ computed by the M-step of EM for a fixed set of q functions.

3. **Active Learning on Circles:**
Consider a binary classification problem for points $x \in \mathbb{R}^2$ such that $||x||_2 = 1$.

   (a) (6 points) Suppose the hypothesis space consists of linear separators with zero bias and that the training data consists of n points evenly spaced around the unit circle that are separable under this hypothesis space. Describe an efficient, in terms of the number of queries and n, active learning strategy to find a hypothesis whose error on the training set is at most $\epsilon > 0$. How many queries does your algorithm make in the worst case?

   (b) (2 points) Provide a lower bound on n that would guarantee that the true error of the hypothesis produced by your algorithm in part (a) is at most ϵ.

(Active Learning on Circles continued)

(c) (7 points) Suppose instead that the hypothesis space consists of **arbitrary** linear separators in $\mathbb{R}^2$. Argue that, under the assumption that the data on the circle is linearly separable, $O(\frac{1}{\epsilon})$ labels are required, in the worst case, by any active learner to produce a hypothesis with true error at most ϵ. Hint: consider subsets of the hypothesis space of size $1/\epsilon$.

(d) (5 points) More generally, suppose you are given n data points that can be perfectly classified with a hypothesis space with VC dimension $d < \infty$. Explain why an active learner only needs $O(\frac{d}{\epsilon}\log\frac{1}{\epsilon})$ labels in the worst case to find, with probability at least 80%, a hypothesis with error at most $\epsilon > 0$.

4. **VC Dimension:** Consider a binary classification problem for data points in $\{0,1\}^n$.

   (a) (10 points) What is the VC dimension of the family of neural networks consisting of a single perceptron with n binary inputs and one binary output under the constraint that the weights and bias can only take values in $\{0,1\}$?

5. **Short Answer:**

(a) (5 points) Let $M = xx^T$ for some column vector $x \in \mathbb{R}^n$. Show that M is a positive definite matrix. What is the maximum eigenvalue?

(b) (5 points) A star is a Bayesian network with one central node and a directed edge from the central node to each other variable node in the network. Describe a polynomial time algorithm to find the star-structured network and parameters that maximize the likelihood of given training data in $\{0, 1\}^n$.

(Short Answer continued)

(c) (5 points) Explain the difference between generative and discriminative models. In what situation would you prefer a generative model?

(d) (5 points) Explain why the missing not at random assumption makes learning the pattern of missing data challenging in practice.