**CS 7301** **Name (Print):** _____
**Fall 2016**
**Final Exam**
**12/14/2016**

---

This exam contains 12 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 | 20 | |
| 2 | 25 | |
| 3 | 10 | |
| 4 | 15 | |
| 5 | 30 | |
| 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) Policy iteration converges to the unique global optimum.

   (b) (4 points) The VC dimension of linear separators using a Gaussian kernel for points in $\mathbb{R}^n$ is $n + 1$.

*(True/False continued...)*

(c) (4 points) The EM algorithm always converges.

(d) (4 points) The maximum likelihood of a specific Bayesian network structure for a given data set cannot increase whenever an edge is removed from the network.

(e) (4 points) The hidden variables in the LDA model are the distribution of topics for each document and the distribution of words for each topic.

2. **Exponential Distributions:** Consider a nonnegative, real-valued random variable $X \in [0, \infty)$ that is distributed according to the exponential distribution $X \sim \lambda e^{-\lambda x}$ for some real-valued parameter $\lambda > 0$.

   (a) (5 points) Given data samples $x^{(1)}, \ldots, x^{(m)}$, what is the maximum likelihood estimate for $\lambda$? Is the log-likelihood a concave function of $\lambda$?

   (b) (5 points) Consider a naïve Bayes model for binary classification of nonnegative vectors in $\mathbb{R}^n$ in which all of the conditional probability distributions, $p(x_i|y)$, are modeled as exponential distributions for each fixed value of the label $y$. For a given set $(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, x^{(m)})$ of data points, $x$, and their corresponding labels, $y$, what is the maximum likelihood estimate of $\lambda_i(y)$ for the $i^{th}$ feature in this naïve Bayes model?

*(Exponential distributions continued...)*

(c) (7 points) For some fixed positive integer $k$, consider fitting a mixture of $k$ exponential distributions to the data as we did for GMMs. In the E-step of the EM algorithm, what is the approximate distribution $q_i$ for the $i^{th}$ data point as a function of the current estimate of the parameters $\lambda_1, \ldots, \lambda_k$ and the current mixture probabilities $p_1, \ldots, p_k$?

(d) (8 points) What is the M-step of EM for the parameter $\lambda_j$ of the $j^{th}$ mixture component in the exponential mixture model for a fixed value of the approximating distributions $q_1, \ldots, q_m$ and mixture parameters $p_1, \ldots, p_k$?

*(Exponential distributions continued...)*

3. **Active Learning:**
Active learning is a variant of semisupervised learning in which labels for each of the data points are initially hidden, but can revealed to the learner upon request. The goal of the learner, in say a classification task, is to produce the best classifier possible while requesting the fewest number of labels.

   (a) (6 points) You are given $m$ distinct data points in one dimension whose labels are unknown. However, you are promised that the data is linearly separable. Describe an active learning strategy that finds the correct linear separator while asking for as few data labels as possible. How many labels does your strategy request in the worst case?

   (b) (4 points) Give an example of a real-world problem for which active learning would be preferred over supervised learning.

4. **Lagrange Multipliers:** Consider a set of data points $x^{(1)}, \ldots, x^{(m)}$ in $\mathbb{R}^n$. In this question, you will prove that PCA explains as much of the variance as possible. To do this, suppose that we want to find the direction in $\mathbb{R}^n$ such that the variance of the data points is largest when projected along this direction. Recall that the projection of a vector $x$ along the line defined by the unit vector $w$ is given by $(w^T x)w$.

   (a) (3 points) Formulate a constrained optimization problem that seeks to maximize the sample variance of the points $w^T x^{(1)}, \ldots, w^T x^{(m)}$ over all possible unit vectors $w \in \mathbb{R}^n$.

   (b) (3 points) Express you constrained optimization problem using Lagrange multipliers.

*(Lagrange multipliers continued)*

(c) (3 points) Compute the gradient of the Lagrangian with respect to $w$ and set it equal to zero. Express the resulting linear system in terms of the sample covariance matrix for the $x$'s.

(d) (3 points) Use the result from part (c) and the constraint from part (a) to find the zero gradient points of the Lagrangian. Hint: for vectors $a, b, c \in \mathbb{R}^n$, $(a^T(b-c))^2 = a^T(b-c)(b-c)^T a$.

(e) (3 points) Which critical point of the Lagrangian has the maximum value?

5. **Short Answer:**

(a) (5 points) Suppose you are given a Bayesian network with latent variables. Explain why latent variables that have no children in the network can be safely ignored when performing maximum likelihood estimation using the EM algorithm.

(b) (5 points) Suppose you plan to use an SVM to train a classifier for a given data set. However, given the high dimensionality of the data, you want to first perform PCA. You divide your data into a training set and a testing set. On which data should PCA be performed: the training set only, the testing set only, or both the training and the testing set? Explain.

*(Short Answer continued...)*

(c) (5 points) Consider a binary classification problem in $\mathbb{R}^n$ using a given hypothesis space $\mathcal{H}$. If $\mathcal{H}$ can shatter a set of $n$ points, what is the smallest that $|\mathcal{H}|$ can be?

(d) (5 points) We saw an example in class of an application of PCA to a face recognition task. The procedure required very few eigenvectors to be accurate in this case. On what kind of data set will PCA perform the worst (i.e., the amount of information lost by keeping only the top $k$ eigenvectors is as large as possible)? Be specific.

*(Short Answer continued...)*

(e) (5 points) We saw many machine learning algorithms that required choosing a learning rate or step size (e.g., gradient descent). Explain the problems that can occur if the step size is too large or too small.

(f) (5 points) Explain the differences and similarities between logistic regression and a single sigmoid neural network.