Machine Learning Qualifier
Fall 2017

This exam contains 11 pages (including this cover page) and 5 problems. Check to see if any pages are missing.

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. Ask for additional paper if needed.
- **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	25	
2	15	
3	20	
4	15	
5	25	
Total:	100	

1. **Maximum Likelihood for Linear Regression:** Consider the standard linear regression problem of fitting a linear function, $f(x) = ax + b$, to data points $x^{(1)}, \ldots, x^{(M)} \in \mathbb{R}$ with continuous labels $y^{(1)}, \ldots, y^{(M)} \in \mathbb{R}$. Let's assume that our model has noise, i.e., $y^{(m)} = ax^{(m)} + b + \epsilon^{(m)}$ where $\epsilon^{(m)}$ is distributed with mean zero and variance $\sigma^2 > 0$. In other words, $p(y^{(m)}|x^{(m)}, a, b, \sigma^2)$ is normally distributed with mean $ax^{(m)} + b$ and variance σ^2.

   (a) (5 points) What is the conditional log-likelihood for the given data set under the above noise assumption?

   (b) (15 points) Find the values of a, b, and σ^2 that maximize the conditional log-likelihood.

(MLE for Linear Regression continued)

(c) (5 points) For which of a, b, and σ^2 is the maximum likelihood estimator unbiased?

2. **VC Dimension of Circles:** Consider a binary classification problem for points in $\mathbb{R}^2$ using a hypothesis space whose elements consist of a pair circles such that any point inside both of the circles is classified as a plus and the remaining points are classified as minus.

   (a) (10 points) What is the VC dimension of this hypothesis space? Justify your answer with a proof.

(VC Dimension of Circles continued)

(b) (5 points) Does the VC dimension of this hypothesis space increase if you add all of the hypotheses consisting of a pair circles such that any point inside both of the circles is classified as a minus and the remaining points are classified as plus to the hypothesis space from part (a)?

3. **Gaussian Naïve Bayes:** Consider a Gaussian naïve Bayes model for a binary classification problem with data points in $\mathbb{R}^n$, i.e., $p(X_i|Y)$ is a Gaussian distribution for each $i \in \{1, \ldots, n\}$ and each value of the class variable Y. Let $\mu_{i,1}$, $\sigma^2_{i,1}$, $\mu_{i,0}$, and $\sigma^2_{i,0}$ represent the mean and variance of each of the conditional distributions where 0 and 1 denote the value of the class label, and consider data observations $x^{(1)}, \ldots, x^{(M)} \in \mathbb{R}$ and corresponding class labels $y^{(1)}, \ldots, y^{(M)} \in \{0, 1\}$.

   (a) (5 points) Give an example of a real-life data set that can be well-approximated in this framework.

   (b) (5 points) As a function of the model parameters, what is $p(X_i > 10)$?

(Gaussian Naïve Bayes continued)

(c) (10 points) Suppose that you introduce a prior of the form $\lambda_{i,y} \exp(-\lambda_{i,y}\mu_{i,y})$ for some $\lambda_{i,y} > 0$ on each $\mu_{i,y}$ for $i \in \{1, \ldots, n\}$ and $y \in \{0, 1\}$. What is the MAP estimate for $\mu_{i,1}$ under this prior?

4. **Max-margin Linear Separators:** Consider a binary classification problem for data points in $\mathbb{R}^n$ that are linearly separable.

   (a) (10 points) Reformulate the problem of finding a linear separator as an optimization problem over the parameters of two parallel hyperplanes such that all positive data points lie above both separators and all negative data points lie below both separators and the hyperplanes are as far apart from each other as possible. Hint: the distance between a point x' and a plane $w^T x + b = 0$ is given by $\frac{|w^T x' + b|}{\sqrt{w^T w}}$.

(Max-margin Linear Separators continued)

(b) (5 points) If the optimization problem that you produced in part (a) was not convex in the parameters of the hyperplanes, explain how you can make it convex by adding a single convex constraint to your objective. Explain how to recover the standard SVM solution from the solution to your optimization problem.

5. **Short Answer:**

   (a) (5 points) Explain why a Gaussian naïve Bayes model may not be appropriate to model a binary classification problem over continuous variables in $[0, 1]^n$ for $n > 0$.

   (b) (5 points) Is it possible to have an infinite hypothesis space that has VC dimension one? Explain why or why not.

(Short Answer continued)

(c) (5 points) The adaBoost algorithm selects the hypothesis from the hypothesis space that minimizes the weighted error at each iteration. Does adaBoost still minimize the exponential loss in the limit as the number of rounds tends towards infinity if, instead, the algorithm selects the hypothesis whose weighted training error is farthest from 50%?

(d) (5 points) Suppose you are given a finite hypothesis space H consisting of distinct hypotheses. After seeing $m > 0$ data points, what are the smallest and largest number of hypotheses in H that are consistent with all of the m data points?

(e) (5 points) Explain how SVMs can be used for clustering data points into k clusters if the correct cluster assignment for some nonzero fraction of the data points in each cluster are provided as part of the training data.