CS 7301
Fall 2016
Midterm
10/11/2017

Name (Print): _____

This exam contains 9 pages (including this cover page) and 4 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 necessarily ordered by difficulty. Be sure to read each of the questions carefully.
- **Do not** write in the table to the right.

Problem	Points	Score
1	32	
2	25	
3	10	
4	33	
Total:	100	

1. (32 points) **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) Bagging can reduce bias.

   (b) The error on the training set converges to zero as the number of rounds of boosting tends towards infinity.

   (c) If a hypothesis space has at least two distinct hypotheses, then it has VC dimension one.

   (d) The set of subgradients of a function f at a point x is always nonempty.

(True or False continued)

(e) Increasing the size of the test data set reduces the variance of the estimated classifier.

(f) Consider a regression problem with one real-valued input and one real-valued output such that you are fitting functions of the form $f_{a_1,a_2,b_1,b_2}(x) = \max\{a_1x + b_1, a_2x + b_2\}$ with parameters $a_1, a_2, b_1, b_2 \in \mathbb{R}$. If the (x, y) observations are $(1, 2), (-10, 34)$, and $(100, 13)$, the minimum value of the squared loss function is 4.

(g) With small modification, the k-nearest neighbor algorithm can be used for regression.

(h) The maximum height of a decision tree for a binary classification problem with data points in $\mathbb{R}^n$ is n.

2. **n-Dimensional Spheres:** Consider a binary classification problem over points in $\mathbb{R}^n$ with a hypothesis space consisting of spheres such that points inside the sphere are classified as $+$ and points outside the sphere are classified as $-$. We can represent each of these hypotheses in terms of two variables: the center $c \in \mathbb{R}^n$ and the radius $r \in \mathbb{R}$.

   (a) (5 points) Express whether or not a data point $x \in \mathbb{R}^n$ is inside the sphere defined by c and r as a constraint.

   (b) (10 points) Suppose that you want to find a sphere that classifies a given data set as well as possible. Specifically, as many points as possible labeled plus should be inside the sphere and as many points as possible labeled minus should be outside the sphere. To do this, you may need to weaken the constraints to allow some violation. Given $M > 0$ training data points $x^{(1)}, \ldots, x^{(M)} \in \mathbb{R}^n$ with corresponding labels $y^{(1)}, \ldots, y^{(M)} \in \{+, -\}$, formulate the learning task as a constrained minimization problem over c, r, and slack variables such that a linear penalty is paid for each nonzero slack variable.

(n-Dimensional Spheres continued)

(c) (5 points) To be convex, a minimization problem in standard form must have a convex objective, all convex functions for inequality constraints, and all equality constraints must be linear. Is the constrained minimization problem that you constructed in part (b) convex in c, r, and the slack variables? Explain.

(d) (5 points) Does an SVM with a polynomial kernel of degree at most 2 solve the learning problem from part (b)? Explain.

3. **Short Answer:**

   (a) (5 points) What are the advantages and disadvantages of using Gaussian kernels for classification?

   (b) (5 points) What are the advantages of a "bushy" decision tree versus a tall skinny decision tree? Is a bushy decision tree always preferable to a skinny one?

4. **Pairs of Linear Separators:** Consider a binary classification problem with data points in $\mathbb{R}^n$. Let H be the hypothesis space consisting of pairs of linear separators in $\mathbb{R}^n$, denoted by the vectors $w_1, w_2 \in \mathbb{R}^n$ and the constants $b_1, b_2 \in \mathbb{R}$, such that for a given point $x \in \mathbb{R}^n$

   - if $w_1^T x + b_1 \geq 0$ and $w_2^T x + b_2 \geq 0$, then the point is classified as a $+$
   - if $w_1^T x + b_1 < 0$ and $w_2^T x + b_2 < 0$, then the point is classified as a $+$
   - Otherwise, the point is classified as a $-$

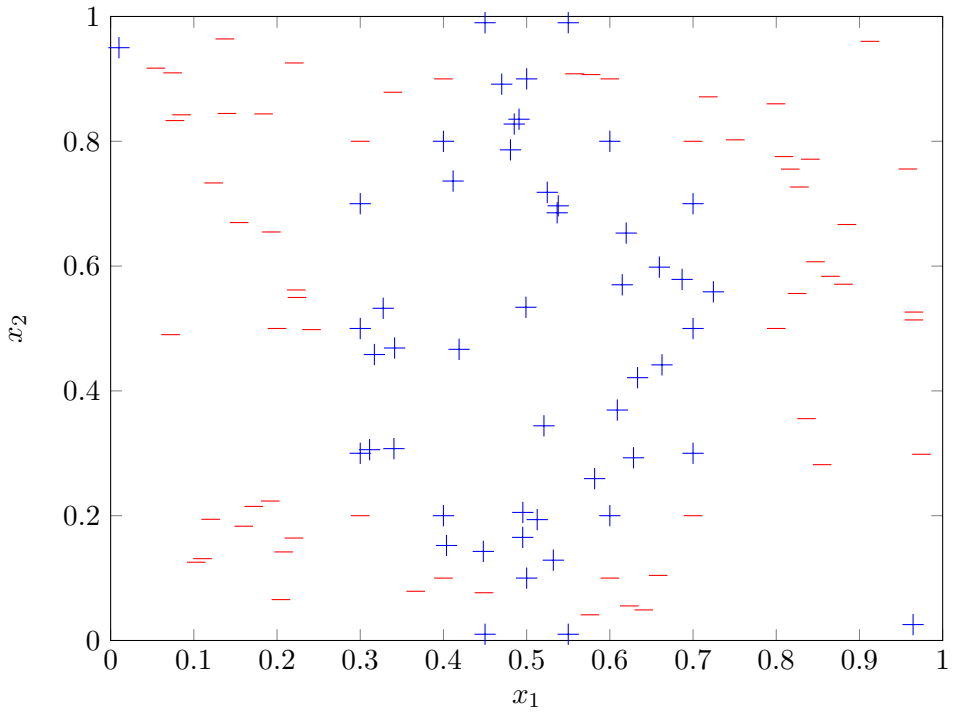
   (a) (15 points) What is the VC dimension of this hypothesis space on $\mathbb{R}^2$? Prove it.

(Pairs of Linear Separators continued)

(b) (10 points) Suppose that you are given training data that is linearly separable under this hypothesis space. Formulate the problem of finding a perfect classifier from the above hypothesis space as an optimization problem.

(c) (3 points) Is your loss function from part (b) convex in w_1, w_2, b_1, b_2? Explain.

(d) (5 points) Suppose you are given the following data points in $\mathbb{R}^2$ as training data.



The above data is not separable under the original hypothesis space - but it is separable under the previous hypothesis space using the feature map,

$$\phi(x_1, x_2) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix}.$$

Draw a perfect classifier under this feature map in the figure in which you clearly indicate the two chosen linear separators in the feature space.