Machine Learning Qualifier
Fall 2016

This exam contains 12 pages (including this cover page) and 4 problems. Check to see if any pages are missing.

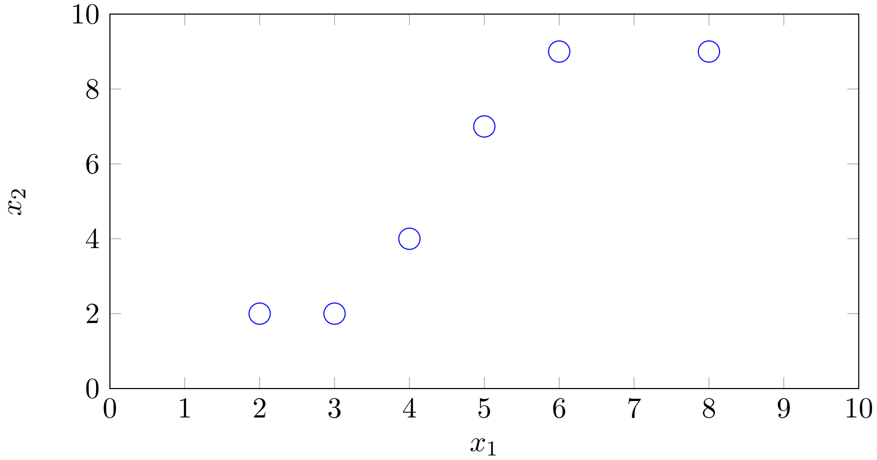
You may **NOT** use books, notes, or any electronic devices on this exam. Examinees found to be using any materials other than a pen or pencil will receive a zero on the exam and face possible disciplinary action.

The following rules apply:

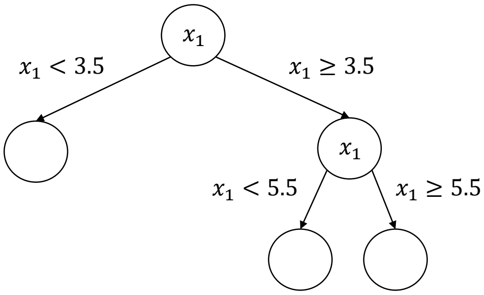
- **Organize your work**, in a reasonably neat and coherent way, in the space provided. Work scattered all over the page without a clear ordering will receive very little credit. Ask for additional paper if needed.
- **To ensure maximum credit** on short answer / algorithmic questions, be sure to **EXPLAIN**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

Problem	Points	Score
1	20	
2	35	
3	30	
4	15	
Total:	100	

1. **Regression Trees:** Suppose that we want to perform regression with decision trees. The resulting regression trees are exactly the same as binary decision trees except that, at the leaf nodes, the predicted value is determined averaging all of the observed predictions for the data points at that leaf (instead of majority vote in a standard decision tree). For this question, consider the following data set of points in $\mathbb{R}^2$: $\{(2,2),(3,2),(4,4),(5,7),(6,9),(8,9)\}$. The goal of this problem is to predict the second component x_2 given the first component, x_1.



(a) (5 points) Draw the following regression tree in the plot above. The values at the leaf nodes should be calculated using the above data set. What is the error of this estimator, in terms of squared error?



(b) (5 points) What is the optimal (in terms of squared error) regression tree with exactly one non-leaf node for the data .

(c) (5 points) Is there a polynomial time algorithm for the problem of finding the optimal regression tree with exactly one non-leaf node (in terms of squared error) on a general data set in $\mathbb{R}^n$? Again, the regression tree attempts to predict n^{th} component of the vector given the first $n-1$ components. Explain why or why not.

(d) (5 points) Explain how one-dimensional k-means clustering is a special case of the regression tree learning problem.

2. **Cylindrical Separators:** Consider data points in $\mathbb{R}^2$ that are labeled with either a + or a minus $-$. An infinite cylinder in $\mathbb{R}^2$ is the set of all points that are a distance at most r from a given line l for some $r > 0$ and some line l. Consider the hypothesis space of all infinite cylinders in $\mathbb{R}^2$. For a given cylinder of radius r, points are classified as + if they within a distance r of the defining line and all other points are classified with a $-$.

   (a) (15 points) What is the VC dimension of this hypothesis space? Prove it.

(Cylinders continued)

(b) (5 points) Does the VC dimension change (i.e., does it increase, decrease, or remain the same) if the lines of the form $y = mx + b$ in the definition of the cylinders are replaced with curves of the form $y = ax^2 + bx + c$ for some $a, b, c \in \mathbb{R}$?

(c) (5 points) Consider the data set $(1, 1, +), (0, 0, +), (0, 1, -), (1, 0, -)$. Using the hypothesis space from part (a), give a max-margin classifier for this data set.

(Cylinders continued)

(d) (10 points) Describe a feature map such that if the labeled data in $\mathbb{R}^2$ is separable by a cylinder, then the image of the data points under the feature map will be linearly separable.

Hint: The distance of a point (x_0, x_1) in $\mathbb{R}^2$ to the line such that $ax + by + c = 0$ is given by $\frac{|ax_0+bx_1+c|}{\sqrt{a^2+b^2}}$.

3. **Poisson Maximum Likelihood Estimation:** Consider a nonnegative, integer-valued random variable X that is distributed according to a Poisson distribution $X \sim \frac{\lambda^x e^{-\lambda}}{x!}$ for some real-valued parameter $\lambda > 0$.

   (a) (3 points) Given data samples $x^{(1)}, \ldots, x^{(m)}$, what is the maximum likelihood estimate for λ?

   (b) (12 points) Suppose you are interested in bounding the sample complexity. Using the Chernoff bound below, derive a lower bound on the number of samples needed to guarantee that $\lambda_{MLE} \leq \lambda + \epsilon$ for some $\epsilon > 0$ with probability at least $1 - \exp(-5)$. For full credit, your bound should be the best possible.

   Chernoff bound: for a random variable Y and a real number $t > 0$,

$$\Pr(Y \geq a) \leq \frac{\mathrm{E}(\exp(tY))}{\exp(ta)}$$

   where $E(\exp(tY))$ is the expected value of $\exp(tY)$.
   Hint: $\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}$.

(MLE continued)

(MLE continued)

Suppose now that you introduce a prior probability distribution, $\lambda \sim \frac{1}{5}\max\{-\lambda/10 + 1, 0\}$.

(a) (10 points) What is the MAP estimate under this prior probability distribution?

(MLE continued)

(b) (5 points) Compare the above prior distribution with the prior distribution $\exp(-\lambda)$. Which would you prefer in general? Give at least two reasons for your choice.

4. **Short Answer:**

   (a) (5 points) Both neural networks and decision trees are capable of perfect binary classification on noise free data, yet neural networks are more popular in practice for applications such as handwritten digit recognition, why is this?

   (b) (5 points) Explain how boosting where the weak learners are decision trees with exactly one non-leaf node can be used to perform feature selection.

(Short answer continued)

(c) (5 points) Describe two ways in which you can use a Boolean classifier to perform multi-category (or multi-class) classification. What are the pros and cons of each?