Exam #1

Instructions. This is a 135-minute test. You may use your notes. You may assume anything that we proved in class or in the homework is true.

Question	Score	Points
1		10
2		10
3		10
4		10
5		10
6		10
Out Of		60

Name: ________________________________

1. We say $f : \mathbb{R} \to \mathbb{R}$ has bounded gradients if for every $x, y \in \mathbb{R}$ $|f'(x) - f'(y)| \leq L|x - y|$ for some absolute constant L that does not depend on x or y. (the expression f' is the derivative of f). Now consider the function $f(x) = |x - 2|$. Is this function convex? Explain why or why not. Does this function have bounded gradients (assume the gradient at the point $x = 2$ is defined to be 0). If so, for what value of L? Consider starting at the point $x = 1.05$ and running gradient descent to find the minimum of $f(x)$. Assume you use a learning rate equal to .1. Will you converge to the global minimum? Explain.

2. Regression problems.

(a) You are given a data set $S = (x_1, y_1), \ldots, (x_t, y_t)$ where each x_i and y_i are real numbers. You perform simple linear regression to obtain the line $\beta_0 + \beta_1 x$. Now re-scale the $y_i's$ so that $y_i' = \alpha y_i$ for some real number α. Perform simple linear regression again. How do the coefficients β_0, β_1 change for the new line, quantitatively? You may reason by drawing a picture or using formulas for these coefficients from class.

(b) What happens if the x_i's are scaled by α?

(c) For each of the following scenarios, state whether or not we can effectively use linear regression, and give a short reason.

(i) We have training data (x, y) (where $x \in \mathbb{R}^2, y \in \mathbb{R}$) satisfying $y = \alpha x_1 + \beta x_2$, and we want to learn the model parameters α, β. (That is, we have training data of the above form for various different x.)

(ii) We have training data (x, y) (where $x \in \mathbb{R}^2, y \in \mathbb{R}$) satisfying $y = \alpha x_1^2 + \beta x_2^3$, and we want to learn the model parameters α and β.

(iii) We have training data (x, y) (where $x \in \mathbb{R}^2, y \in \mathbb{R}$) satisfying $y = 2^\alpha x_1^\beta$, and we want to learn the model parameters α, β.

3.

(a) Suppose we have a data set consisting of three points in $\mathbb{R}^2$: $(1,3),(2,6),(3,9)$. How many principal components does this data set have? Write down the first principal component.

(b) Given the SVD of matrix

$$A = U\Sigma V^T = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 3 & 1 \\ 7 & 9 & 10 \end{bmatrix}$$

$$= \begin{bmatrix} -0.222 & 0.364 & -0.905 \\ -0.270 & -0.914 & -0.301 \\ -0.937 & 0.178 & 0.301 \end{bmatrix} \cdot \begin{bmatrix} 16.180 & 0 & 0 \\ 0 & 2.862 & 0 \\ 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} -0.486 & -0.599 & -0.637 \\ -0.716 & 0.145 & -0.682 \\ 0.501 & -0.788 & 0.358 \end{bmatrix}.$$

Write down the matrix that is the best rank-1 approximation to A. You don't need to calculate the exact numbers, a formula or expression is enough.

(c) Take the data set A and project onto the first k principal components. How can you write this as an expression involving only U and Σ?

4.

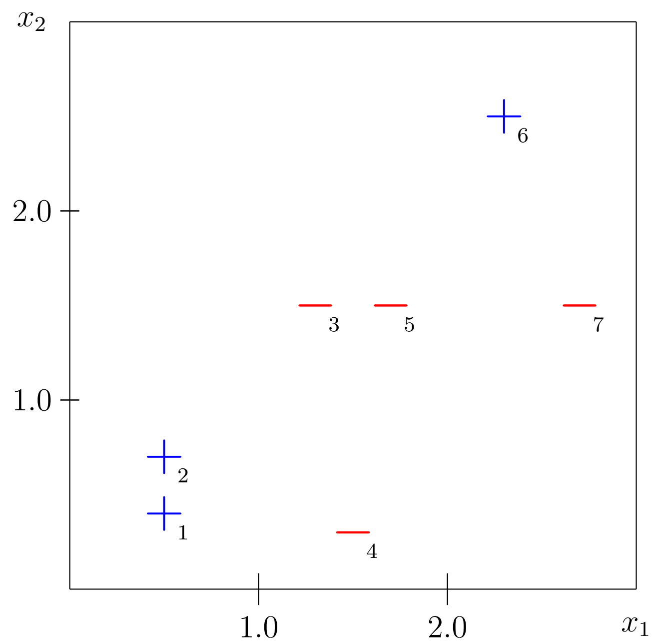
In this problem we want to learn the concept class of origin-centered concentric circles. More specifically, fix two circles (both have centers at the origin) in the plane. The learner is given draws from a distribution where each point in between the two circles is labeled positive, and every other point is labeled negative (points on the boundary of either circle are labeled positive).

(a) Give an algorithm for PAC learning this concept class. You may just describe it in words or via pseudocode.

(b) Describe the bad events for your algorithm (i.e., the events where your algorithm will fail to output an accurate hypothesis). Be as formal as you can.

(c) Now in terms of ϵ and δ analyze the sample complexity of your algorithm.

5.



(a) Assume we want to use Adaboost to classify the training examples S in the 2D plane given in the figure above. The weak learner outputs a function of the form $\mathsf{sign}(x_i - t)$ for $i \in \{1,2\}$ and $t \in \{1,2\}$ **or its negation**. You can assume $\{+1,-1\}$ labels. If you recall, at the beginning of Adaboost, the weight for each training example is the same, so $w_1 = (1,1,\ldots,1)$. What is the best hypothesis for the weak learner to output to minimize the error rate with respect to this initial weighting?

What is its error rate? Call it E_1.

(b) Following the Adaboost algorithm, we must now give new weights to each point. What is the new weighting w_2 (recall $\beta_1 = \frac{E_1}{1-E_1}$)?

Now according to this new weighting of points, what is the best hypothesis for the weak learner to output?

6.

(a) In each of the following plots, a training set of data points X in $\mathbb{R}^2$ labeled either $+$ or $-$ is given, where the original features are the coordinates (x, y). You can assume that the data is origin-centered. For each of the two training sets below, answer the following questions:

  (i) Draw a simple simple recreation of each of the two datasets below (no need for exact precision) and draw all the principal components (eyeball it).

  (ii) For each dataset, can we correctly classify the labels by using a halfspace after projecting onto one of the principal components? If so, which principal component should we project onto? If not, explain in 1–2 sentences why it is not possible.

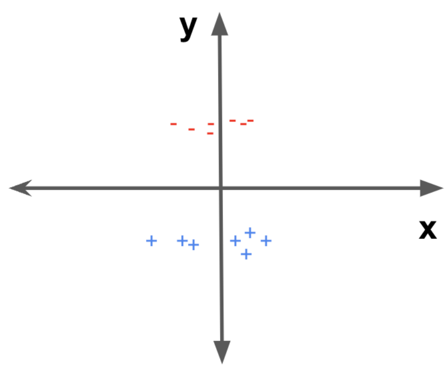


Figure 1: Dataset 1

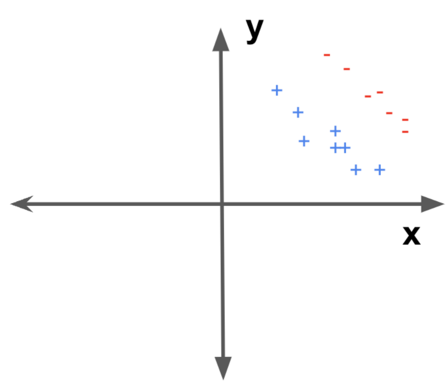


Figure 2: Dataset 2

(b) Is it possible to have a data set in $\mathbb{R}^2$ that is linearly separable by a halfspace in $\mathbb{R}^2$ but is not linearly separable after projecting onto *either* of the two principal components?

If so, give a simple example along the lines of the above data sets. If not, explain in 1–2 sentences why it is not possible.