MAT 345 - Homework 1

Due Wednesday, September 19, 2018, in class

1. For each of the following tasks, state which kind of learning is involved (supervised, reinforcement, or unsupervised), and the training data to be used. If a task fits in more than one type, explain how.

   (a) recommending a book to a user in an online bookstore

   (b) playing tic-tac-toe

   (c) categorizing videogames into different types

   (d) learning to play piano

   (e) deciding on a maximum line of credit for each customer at a bank

2. Express each of the following tasks by specifying the input space $\mathcal{X}$, the output space $\mathcal{Y}$, the target function $f : \mathcal{X} \to \mathcal{Y}$ and the specifics of the data set $\mathcal{D}$ that we will learn from.

   (a) a patient walks in with a medical history and some symptoms, and you want to identify the medical problem

   (b) handwritten digit recognition

   (c) determining if an email is spam or not

   (d) a problem of interest to you for which there is no analytic solution, but you have data from which to construct a solution. State the problem you are considering.

3. Suppose we use the perceptron to detect spam messages. Let's say that each email message is represented by the frequency of occurrence of keywords, and the output is $+1$ if the message is considered spam.

   (a) List some keywords that will have a large *positive* weight in the perceptron.

   (b) List some keywords that will have a *negative* weight in the perceptron.

   (c) What parameter in the perceptron affects how many borderline messages are classified as spam?

4. Program the PLA algorithm and run it on the given data set (below). Output the final hypothesis function h and the number of iterations needed until PLA converges.

x_1	x_2	x_3	y
3	2	1	1
1	2	0	1
4	0	0	-1
2	1	2	1
3	-2	1	-1
2	2	-1	1
4	1	3	-1
0	-1	2	-1
1	0	4	-1
-2	1	2	1

5. We choose a line in the plane and set our target function f to $+1$ on one side of the line and to -1 on the other. Let $f : \mathbb{R}^2 \to \{-1, +1\}$ by

$$f(x_1, x_2) = \begin{cases} +1 & , \quad \text{if } 2x_1 + 3x_2 > 1 \\ -1 & , \quad \text{if } 2x_1 + 3x_2 < 1 \end{cases}$$

Generate a data set $\mathcal{D}$ of 20 randomly chosen points and run the PLA on your data set. Output $\mathcal{D}$, and the final hypothesis function h. Note how long it takes for PLA to converge and how well the final hypothesis h matches the target function f.