This homework is to be done in groups of 2-3. Each group should work independently, and must acknowledge all sources of inspiration, techniques, and/or helpful ideas (web, people, books, etc.) other than the instructor and class text.

1. Consider the hypothesis class of homogeneous half-spaces in the $\mathbb{R}^d$ (i.e. each instance $x \in \mathbb{R}^d$ and $\mathcal{H}$ consists of all $h_w$ having the form $h(x) = +1$ if and only if $w \cdot x > 0$, so $0$ is always mapped $-1$). Determine the VC-dimension of homogeneous half-spaces in $\mathbb{R}^d$. In other words, for each $d \geq 1$ find a set $S \subset \mathbb{R}^d$ is shattered by $\mathcal{H}$ and show that no set $S' \subset \mathbb{R}^d$ with $|S'| > |S|$ is shattered by $\mathcal{H}$. (Since if $S'$ is shattered, then every subset of $S'$ is also shattered, it suffices to show that no set $S'$ with $|S'| = |S| + 1$ is shattered). You may use the fact that in any set $S$ of $d+1$ points in $\mathbb{R}^d$, there is at least one point $x \in S$ that can be expressed as a linear combination of the other points in $S$.

2. Consider the following experiment. A magician has three coins in his pocket, a two-headed coin, a two-tailed coin, and a fair coin. The magician picks a coin from his pocket with each coin equally likely. The magician then flips the coin twice, and sees what the comes up (either hh, ht, th, or tt). To make this more formal, consider three random variables $\text{coin} \in \{0, \frac{1}{2}, 1\}$, $\text{flipA} \in \{0, 1\}$, and $\text{flipB} \in \{0, 1\}$ where the value of $\text{coin}$ gives the probability that a "head" results when the coin is flipped (and $1 - \text{coin}$ is the probability that a "tail" results) and $\text{flipA}$ and $\text{flipB}$ are indicator functions for the events "the first flip is a head" and "the second flip is a head" respectively. Let each triple of values ($\text{coin}, \text{flipA}, \text{flipB}$) correspond to a point (atomic event) in $\Omega$, and assume that $P(\text{coin} = 0) = P(\text{coin} = 1/2) = P(\text{coin} = 1) = 1/3$.

First, what is $|\Omega|$? Second, how many points in $\Omega$ have zero probability? Third (main part), what is $P(\text{flipB} = 1|\text{flipA} = 1)$?

3. (The Monty Hall problem) Consider a game show where a contestant is shown three doors numbered 1, 2, and 3, and behind one door is a good prize, while nothing is behind the other two doors. Assume that the good prize is equally likely to be behind each of the doors. A contestant gets to choose a door. The host reveals that nothing is behind a different door, and then the contest may either keep door one or switch to the third door. What should the contestant do?

To simplify things, assume the contestant initially chooses door 1. If the prize is behind door 1, the host is equally likely to reveal either door 2 or door 3. If the prize is behind door 2 then the host reveals door 3, and vice-versa. Calculate the probability of the contestant winning given that they keep door 1, and the probability of the contestant winning given that they switch to door 2.

4. Perceptron algorithm. Implement the Perceptron algorithm presented in class in 2 dimensions and perform the following experiments where concept $C$ is defined by $C(x) = +1$ if $x_1 + x_2 > 0$ and $C(x) = -1$ otherwise.

Experiment 1:
Generate a 10 example training set by picking points uniformly at random from the unit circle and generating labels ($y$-values) according to $C$. Calculate the gap of the best separating line (this is not likely to be $C$'s decision boundary). Run
the Perceptron algorithm and note how many "mistakes" it makes before finding a consistent hypothesis, and how many iterations through the data are required before it finds a consistent hypotheses.

Perform experiment 1 10 times. Do you see a relationship between the gaps and the number of iterations of number of mistakes made?

Experiment 2:

Generate a 100 example training set by picking points uniformly at random from the unit circle, with noisy labels. For each example \( x = (x_1, x_2) \) in the training set, generate a random number \( r \) in \([0, 1]\). If \( x_1 + x_2 + 2r - 1 > 0 \) then set the label of \( x \) to 1. If \( x_1 + x_2 + 2r - 1 \leq 0 \) then set the label of \( x \) to -1. Also generate a 100 example test set the same way.

This generates a noisy version of \( C \) where the noise tends to be concentrated around the decision boundary.

Run the following version of the perceptron algorithm for 500 iterations where each iteration uses a random point from the training set (rather than cycling through the training set) and save the weight vector \( w_i \) after each iteration \( i \). Compare how well the following prediction rules perform on the test set.

(a) Last hypothesis: predict on \( x \) with \( \text{sign}(w_{500} \cdot x) \), using the hypotheses from the last iteration.

(b) Voted hypothesis: predict on \( x \) using the majority of the \( w_i \), i.e. \( \text{sign} \left( \sum_{i=1}^{500} \text{sign}(w_i \cdot x) \right) \).

(c) Longest survivor: Each \( w \) values is created on some iteration \( t \), predicts correctly for a while, and then makes a mistake at some later iteration \( t' \). The survival time of \( w \) is \( t' - t \). Let \( w_\ell \) be the longest surviving hypothesis from the 500 iterations (pick the first one in case of ties, and assume that the last \( w \) makes an incorrect prediction on iteration 501).

Create 10 different training and testing sets, and run the perceptron algorithm on each training set. How well do each of the three prediction rules do on the test sets?