Homework 4:
Written Exercise Part

1 Balls and Bins [25/3 pts]

Suppose we throw balls into n bins. Each ball is thrown independently and uniformly at random.
(1) [Birthday Paradox] Suppose we throw m balls. What is the probability that at least one bin has more than one balls? Write down the expression and then use the inequality $1 - x \leq e^{-x}$ to give a lower bound.
Solution goes here.
(2) [Coupon Collecting] Let X denote the number of balls thrown until every bin has at least one ball. What is the expectation of X? Express it using the harmonic number $H_n = \sum_{i=1}^{n} 1/i$.
Solution goes here.

2 VC-dimension of Rectangles [25/3 pts]

What is the VC-dimension d of axis-parallel rectangles in R^3? Specifically, a legal target function is specified by three intervals $[x_{\min}, x_{\max}], [y_{\min}, y_{\max}], [z_{\min}, z_{\max}]$, and classifies an example (x, y, z) as positive if and only if $x \in [x_{\min}, x_{\max}], y \in [y_{\min}, y_{\max}]$, and $z \in [z_{\min}, z_{\max}]$. Justify your answer.
Solution goes here.

3 Mistake Bound Model [25/3 pts]

CNF is the class of Conjunctive Normal Form formulas in the form $C_1 \wedge C_2 \wedge \ldots$, where each clause C_i is in the form $L_1 \vee L_2 \ldots$, and each Boolean literal L_i is either a boolean feature x or its negation $\neg x$. k-CNF is the class of CNF in which each clause has size at most k. For example, $x_4 \wedge (x_1 \vee x_2) \wedge (x_2 \vee \neg x_3 \vee x_5)$ is a 3-CNF. Give an algorithm to learn 3-CNF formulas over n boolean features in the mistake-bound model. Your algorithm should run in polynomial-time per example (so the "halving algorithm" is not allowed). How many mistakes does it make at most? (Hint: modify the FIND-S algorithm.)
Solution goes here.

4 Extra Credit: VC-dimension of Linear Separators [20 pts]

In this problem, you will prove that the VC-dimension of the class H_n of halfspaces (another term for linear threshold functions $f_{w,b}(x) = \text{sign}(w^\top x + b)$) in n dimensions is $n + 1$. We will use the following definition: The convex hull of a set of points S is the set of all convex combinations of points in S; this is the set of all points that can be written as $\sum_{x_i \in S} \lambda_i x_i$, where each $\lambda_i \geq 0$, and $\sum_i \lambda_i = 1$. It is not hard to see that if a halfspace has all points from a set S on one side, then the entire convex hull of S must be on that side as well.
(a) [lower bound] Prove that VC-dim$(H_n) \geq n + 1$ by presenting a set of $n + 1$ points in n-dimension space such that one can partition that set with halfspaces in all possible ways, i.e., the set of points are shattered by H_n. (And, show how one can partition the set in any desired way.)
(b) [upper bound part 1] The following is Radon's Theorem, from 1920's.

Theorem 1. Let S be a set of $n+2$ points in n dimensions. Then S can be partitioned into two (disjoint) subsets S_1 and S_2 whose convex hulls intersect.

Show that Radon's Theorem implies that the VC-dimension of halfspaces is at most $n + 1$. Conclude that VC-dim$(H_n) = n + 1$.

(c) [upper bound part 2] Now we prove Radon's Theorem. We will need the following standard fact from linear algebra. If $x_1, \ldots, x_{n+1}$ are $n+1$ points in n-dimensional space, then they are linearly dependent. That is, there exist real values $\lambda_1, \ldots, \lambda_{n+1}$ not all zero such that $\lambda_1 x_1 + \ldots + \lambda_{n+1} x_{n+1} = 0$. You may now prove Radon's Theorem however you wish. However, as a suggested first step, prove the following. For any set of $n+2$ points $x_1, \ldots, x_{n+2}$ in n-dimensional space, there exist $\lambda_1, \ldots, \lambda_{n+2}$ not all zero such that $\sum_i \lambda_i x_i = 0$ and $\sum_i \lambda_i = 0$. (This is called affine dependence.)
Solution goes here.