Exercise 2.4

a)

We have a $(d + 1) \times (d + 1)$ nonsingular matrix, whose rows represent the d+1 points. Let the matrix be X.

Let w be the weight vector that $(w_0, w_1, w_2 ... w_{d+1})^T$.

Since X is nonsingular, according to the theorem, $X^T w$ = Y has a unique solution for every Y ∈ R^{d+1}. (The transpose of a square nonsingular is also nonsingular)

According to the PLA, we have the following: sign($X^T w$) = sign(Y).

This means that we can always find a weight vector w which is the solution of X to Y no matter how we change the configuration of points in X (keep the matrix nonsingular) or the target Y.

In this case, we can conclude that the d-dimension perceptron can shatter d+1 points.

b)

We have a $(d + 1) \times (d + 2)$ linearly dependent matrix X.

In this case, we will have some vector which is a linear combination of other vectors.

We have $X^{(i)} = \sum_{j=1}^{d+1} a_{(j)} X^{(j)}$, where $X^{(i)}$ and $X^{(j)}$is the column matrix of the X, $a_{(j)}$ is a scalar and $a_{(j)}$!= 0.

Assume that there exists a weight vector w that can shatter X, we have

$$X^{(i)T} w = \sum_{j=1}^{d+1} a_{(j)} X^{(j)T} w$$

$$sign(X^{(i)T} w) = sign(\sum_{j=1}^{d+1} a_{(j)} X^{(j)T} w)$$

We get that the class of $X^{(i)}$ is based on the summation of $X^{(j)}$s.

Consider the following situation where we have all $X^{(j)}$s in the class of +1, while $X^{(i)}$ in the class of -1. $sign(X^{(i)T} w) = sign(\sum_{j=1}^{d+1} a_{(j)} X^{(j)T} w)$ will not be true since RHS will be a summation of +1 and LHS will be -1. LHS != RHS which breaks the formula.

This forms a contradiction to our assumption.

In this case, we can conclude that the d-dimension perceptron can not shatter d+2 points.

Problem 2.3

a)

$M_H(N) = 2N$

Since we already knew that $M_H(N) = N + 1$ for positive rays, we need to add the $M_H(N)$ for negative rays which is also N + 1. Then, we need to exclude two duplicates: the data with all +1s and the data with -1s. Finally, we get 2(N+1)-2 = 2N

The break point is N = 3, hence $d_{vc} = 2$.

b)

$M_H(N) = N^2 - N + 2$

Since we already knew that $M_H(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$for positive intervals, we add $M_H(N)$ for negative intervals which is also $\frac{1}{2}N^2 + \frac{1}{2}N + 1$. Then, we remove the duplicates: the data with either +1s or -1s on two sides. (+1,+1…-1,-1) or (-1,-1…+1,+1). We observe that the relation between N and the number of this kind of rows is 2N. We get $M_H(N) = 2 \times (\frac{1}{2}N^2 + \frac{1}{2}N + 1) + 2N = N^2 - N + 2$

The break point is N = 4, hence $d_{vc} = 3$.

c)

$M_H(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$

We could choose two spots from N + 1 intervals, where points between these two spots are positive. Also, we need to consider the special case when all of the points are -1. We have $\binom{N+1}{2} + 1 = \frac{(N+1)N}{2} + 1 = \frac{1}{2}N^2 + \frac{1}{2}N + 1$

The break point is N = 3, hence $d_{vc} = 2$.

Problem 2.8
Possible growth function:
1+N: positive ray
$1 + N + \frac{N(N-1)}{2}$: positive interval
2^N: convex sets

Not possible growth function:
$2^{\sqrt{N}}$: N = 2 is the break point, d_{vc} = 1. This means that the function should be bounded by N+1; however, it is not the case since $2^{\sqrt{n}}$ will eventually grow faster than N+1. We can conclude that it is not a possible growth function.

$2^{N/2}$: N = 1 is the break point, d_{vc} = 0. This means that the function should be bounded by N^0 + 1, which is 2. However, it is definitely not the case since 2 will never grow; $2^{N/2}$ grows exponentially. We can conclude that it is not a possible growth function.

$1 + N + \frac{N(N-1)(N-2)}{6}$: N = 2 is the break point, d_{vc} = 1. This means that the function should be bounded by N+1. It is not the case since $1 + N + \frac{N(N-1)(N-2)}{6}$ has the highest polynomial 3, while the bound only has polynomial 1. We could conclude that this is not a possible growth function.

Problem 2.10

$M_H(2N) \le M_H(N)^2$

If 2N <= d_{vc}, $M_H(2N) = 2^{2N}$, we can get that N <= d_{vc}; $M_H(N)^2 = (2^N)^2$

In this case,$M_H(2N) = M_H(N)^2$

If N <= d_{vc}, $M_H(N)^2 = (2^N)^2$; there are two situations. When 2N <= $d_{vc,}$, we get the same result as above. When 2N > d_{vc}, $M_H(2N) \le 2N^{d_{vc}} + 1$. $M_H(2N) \le M_H(N)^2$ since LHS is polynomial, while RHS is exponential.

If N > d_{vc}, we have 2N > d_{vc}, we get:

LHS <= $(2N)^{d_{vc}} + 1$

RHS <= $(N^{d_{vc}} + 1)^2$ = $N^{2d_{vc}} + 2N^{d_{vc}} + 1$

LHS <= RHS

We can conclude that $M_H(2N) \le M_H(N)^2$

We have $E_{out}(g) \le E_{in}(g) + \sqrt{\frac{8}{N} ln \frac{4M_H(2N)}{\delta}} \le E_{in}(g) + \sqrt{\frac{8}{N} ln \frac{4M_H(N)^2}{\delta}}$

Problem 2.12

From equation 2.13 we have $N \geq \frac{8}{\epsilon^2} ln(\frac{4((2N)^{d_{vc}}+1)}{\delta})$

d_{vc} = 10, $\epsilon = 0.05, \delta = 0.05$

Initial guess of N=1000, we have $\frac{8}{0.05^2} ln(\frac{4((2\times1000)^{10}+1)}{0.05}) \approx 257251$

Plug the N = 257251 into the equation, then repeat the above step until N converges. (Iteration step is in problem2.12.py in which I made 10 iterations)
We find that the N converges to about 452956 after 6 iterations. (Based on the observation of the output from the program)
We can conclude that we need a sample size of about 453000 in order to satisfy the conditions stated.