CS446: Machine Learning	**Fall 2014**
Problem Set HW 4	
Siting Chang	*Handed In: 10/17/2014*

1 VC Dimesion

1.1 part a

VC dimension of H is 3.

It is trivial to show that there exists one or two points on the plane that can be shattered by H. We now show that H is able to shatter 3 points which implies that the VC dimension of H is at least 3.

Select three points with coordinates A(1,1), B(-1,1) and C(-1, -1). There are 8 possible combination of labels which H are able to classify all. The orgin and radius choices are listed as follows for each combination. The order is: label of point A, label of point B, label of point C, origin, radius.

- +, +, +, (0,0), 2
- -, -, -, (0,0), 0.5
- +, -, -, (1,0), 1
- +, -, +, (1,-1), 2
- -, +, -, (-1,1), 1
- -, +, +, (-1,0), 1
- -, -, +, (-1,-1), 0.5
- +, +, -, (0,1), 1

Next, we show that H is not able to shatter any four points on the plane which means that the VC dimension of H is less than 4.

There are two possible situations when randomly choose four points:

- four points form a convex hull. This situation cannot be classified by any hypothese in H when the opposing points with the largest distance both have positive lables and the other two have negative labels.
- three points form a convex hull and one point is internal. This situation cannot be classified when the first three points (on the convex hull) have positive label and the fourth point has negative label.

Therefore, we proved that the VC dimension of H is 3.

1.2 part b

VC dimension of H is 2k.

We start with the base case with one point x_1 on the real line. It is easy to see that we are able to shatter this point no matter it has a positive label (choose $a_1 < x_1 < b_1$ or a negative label (choose $b_1 < x_1 < a_2$).

The most complicated which is also the hardest case to classify is when neighboring points have opposite labels. For example, the most complicated case of four points is when x_1 and x_3 have positive labels and x_2 and x_4 have negative labels. The reasont that it is the hardest case to classify is because each of these four points need to be placed or assigned to a disjoint interval. This leads to the requirement of four intervals. As long as we have enough intervals, at least four intervals, to cope this situation, the points are shattered.

Given 2k points on the real line which has k disjoint intervals within which points are labeled as potive and k more disjoint intervals within which points are negative. The largest number of points that can be shattered is 2k since we are able to assign 2k points with neighboring points have opposite labels into these 2k intervals.

However, H cannot shatter 2k+1 points. Again, consider the most complicated situation with the leftmost point with a positive label. Since the leftmost point must fall in the region with $x_{leftmost} < a_1$, there is no way any hypothesis from H can classify these points, especially the leftmost point.

Therefore, we proved that the VC dimension of H is 2k.

2 Decision Lists

2.1 part a

$\neg c =< (c_1, \neg b_1), ..., (c_l, \neg b_l), \neg b >$

2.2 part b

First, show k-DNF⊆k-DL. Since each term of k-DNF can be trasformed into an item of a decision list with value 1, then clearly k-DNF⊆k-DL. Next, since we can always find some k-DNF that complements any k-CNF along with the fact that k-DL is closed under complementation (shown in part a), we say that k-CNF⊆k-DL. With each component a subset of k-DL, we say there union is also a subset of k-DL denoted as k-DNF ∪ k-CNF⊆k-DL