Robert Werthman
CSCI 5622
Homework 5: Learnability

Problem 1

As stated by the problem, we have $\epsilon = .15$–the error of the the hypothesis h_i. The probability of hypothesis h being outputted with error ϵ is the confidence of the hypothesis which is given by the equation $1 - \delta$. In this particular problem, $1 - \delta = .95$. This means $\delta = .05$.

This problem has a finite, consistent hypothesis class H. It is finite because each hypothesis h_i is a triangle with 3 distinct vertices on the interval [0,99]–there is a maximum number of triangles. It is consistent because each h_i determines if a given a training example x_i is inside or outside of its boundaries. This is aligned with the concept c of labeling each point positive or negative depending on whether that point is interior and exterior to the triangle boundary. There is an algorithm A that exists that given a point x_i and 3 vertices that comprise a triangle, will tell if that point is inside or outside of the triangle boundary.

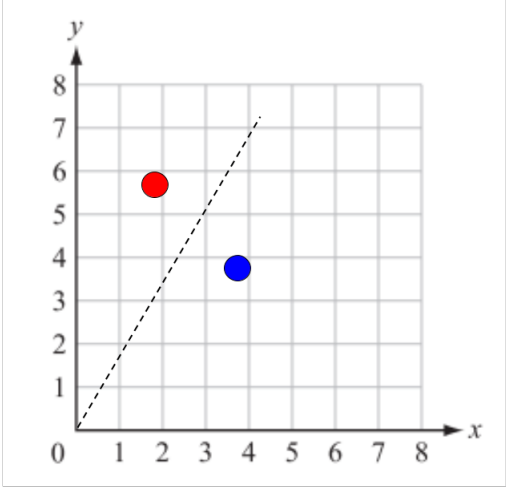
Because this problem has a finite, consistent hypothesis class we can use the following equation to find the bound on training examples m to ensure each hypothesis has a confidence of 95% and error of .15.

$$m \geq \frac{1}{\epsilon}(ln|H| + ln\frac{1}{\delta})$$

$|H|$ is the number of hypothesis which is equal to the number of traingles that can be found in the problem. The number of triangles that can be found in a given interval is the total number of combinations of the total number of points in that interval made up of 3 vertices $\binom{n}{3}$. n in the case when the interval for both x and y [0,99] is 200 points. This means there are $\binom{200}{3} = 1313400$ possible triangles. This means $|H| = 1313400$. Plugging $|H|, \epsilon, \delta$ in to the equation for above we get

$$m \geq \frac{1}{.15}(ln(1313400) + ln\frac{1}{.05}) \approx 114 \text{ training samples}$$

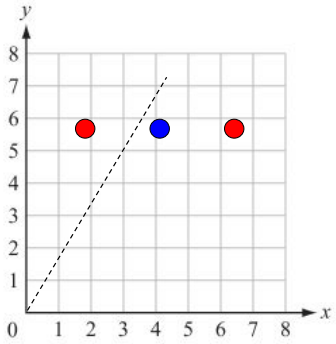
Problem 2



(a) Lower VC Dim bound example

The figure (a) is an example of a configuration of 2 points that are shattered by H. Because this is the lowest number of points H can shatter, the $VCdim(H) \geq 2$.

As for the upper bound on the VCdim I argue that it is 2 as well. As depicted in figure (b), there is no way for a hyperplane to classify all labelings of two points when the point between the other two points is a different label. This is true for all configurations of the points.



(b) Upper VC Dim bound example

In the case of 2 dimensions, the hyperplane (decision boundary) can be flattened into a line represented by the equation $y = wx$, where w is the slope of the line. The hypothesis class is then the set of all these equations for a line that classify w points correctly. Let us label points positive if they can be intersected by a line with slope $r \geq w$ and points negative if they can be intersected by a line with slope $r < w$. This gives us the function for each hypothesis

$$h(p = (x, y)) = \begin{cases} +1 & \text{if } r = \frac{y}{x} \geq w \\ -1 & \text{if } r = \frac{y}{x} < w \end{cases}$$

Proof

We want to prove that for all configurations of the points there exists a labelling that cannot be shattered by the hypothesis class H.

Let p_1, p_2, p_3 be points such that their y and x coordinates are ordered as $(x_1, y_1) \leq (x_2, y_2) \leq (x_3, y_3)$. Let the class labels for each point be set to $p_1 = +1, p_2 = -1, p_3 = +1$. Let w be the slope of the hypothesis h. Since $p_1 = +1$, the slope of p_1, r_1, must be $\geq w$. Since $p_3 = +1$, the slope of p_3, r_3, must be $\geq w$. This would mean, based on the ordering of the coordinates above, that the slope of p_2, r_2, must also be $\geq w$. But this contradicts the original labelling of p_2 as -1 which would mean $r_2 < w$. This means that there is no hypothesis h that can shatter a set of 3 points. Therefore $VCdim < 3$ or $VCdim = 2$.