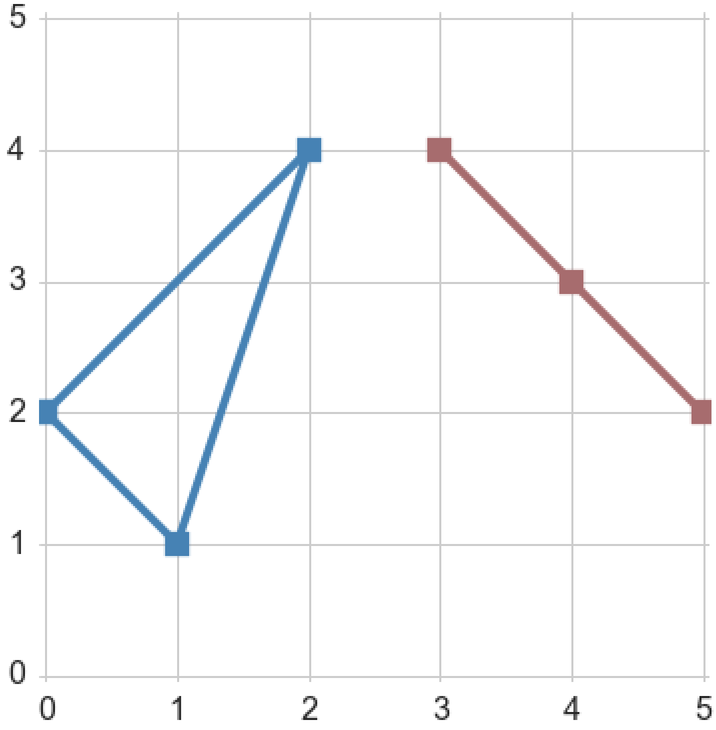
Raymond Duncan
CSCI 5622
Learnability Writeup

Problem 1

Consider the class C of concepts defined by triangles with distinct vertices of the form (i, j) where i and j are integers in the interval [0,99]. A concept c labels points on the interior and boundary of a triangle as positive and points on the exterior of the triangle as negative.

Give a bound on the number of randomly drawn training examples sufficient to assure that for any target class c in C, any consistent learner will, with probability 95%, output a hypothesis with error at most 0.15.

Note: To make life easier, we'll allow degenerate triangles in C. That is, triangles where the vertices are collinear. The following image depicts an example of a degenerate and nondegenerate triangle.



Answer

There are 100^2 possible coordinates for the vertices of the triangle to be. Since order of the points doesn’t matter the total number of hypotheses, H, is 10000 choose 3. Plugging this into the general learnability equation gives $m \geq \frac{1}{.15}(ln(H) + ln(1/.95)) \Rightarrow m \geq \frac{1}{.15}(ln(166616670000) + ln(1/.95)) \Rightarrow$ $m \geq 172.602...$

Thus, we would need at least 173 training examples to train a model which with probability 95% will output a hypothesis with max error of .15

Problem 2

This questions concerns feature vectors in two-dimensional space. Consider the class of hypotheses defined by circles *centered at the origin*. A hypothesis h in this class can either classify points as positive if they lie on the boundary or interior of the circle, *or* can classify points as positive if they lie on the boundary or exterior of the circle. State and prove (rigorously) the VC dimension of this family of classifiers.

Answer

For this question, we think of the feature vectors as $x_i = (\theta_i, r_i)$. Since the classification boundary is the circumference of the circle, we could ignore θ_i and think just of r_i when finding the VC dimension.

We could easily see that the lower bound of the VC dimension is 2 by using labels $r_1 = + \; and \; r_2 = -$ and a configuration where $r_1 < r_2$ for the classification with concept area within the circle and $r_1 > r_2$ with concept area outside the circle.

For the upper bound of the VC Dimension with concept area within the circle, we can assume that $r_1 \leq r_2 \leq r_3$ and $h = [0, r^*]$. Let $r_1, r_3 = + \; and \; r_2 = -$. Given these conditions, we find that $r^* \geq r_1, \; r^* < r_2, \; r^* \geq r_3$ and this requires that $r_1 \leq r_3 \leq r^* < r_2$. Thus the VC dimension cannot be 3.

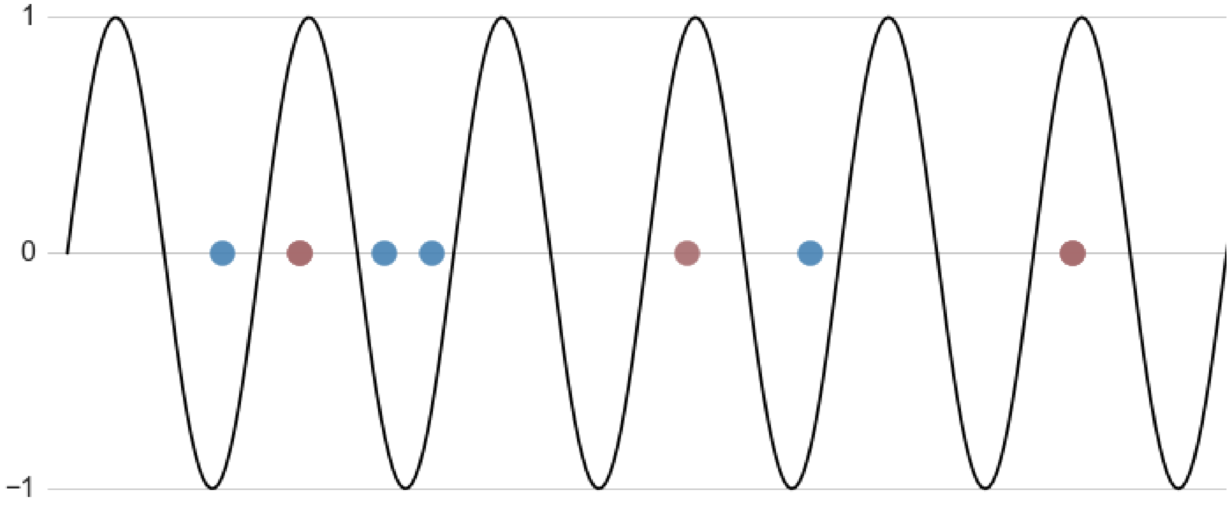
For the upper bound of the VC Dimension with concept area outside the circle, we can assume that $r_1 \leq r_2 \leq r_3$ and $h = [r^*, \infty]$. Let $r_1, r_3 = + \; and \; r_2 = -$. Given these conditions, we find that $r^* \leq r_1, \; r^* > r_2, \; r^* \leq r_3$ and this requires that $r_2 < r^* \leq r_1 \leq r_3$. Thus the VC dimension cannot be 3.

Problem 3

Consider the case of classifying 1D points on the real line with the set of hypotheses of the following form

$$h_w(x) = \begin{cases} +1 & \text{if } \sin(wx) \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

For a given value of w, the hypothesis classifies points as positive if they lie on or below the sine curve and negative if they lie above the sine curve.



It turns out that the VC Dimension of this hypothesis class is infinite. In other words, for any number of training examples there exists a configuration of points which can be shattered by the sine functions.

- Prove this fact by construction, by completing the code in *vc_sin.py* to determine the parameter w that perfectly classifies a given training set (up to floating point precision on the computer). As a hint (you'll have to do some thinking as to why this makes things easier) we've set up the code so that the training points are of the form x = 2**(-k) where k is a nonnegative integer.

- In addition to completing the code you should clearly describe your solution technique in your analysis.
- Give an example of 4 *distinct* points that cannot be shattered by this classifier. How does this relate to the VC dimension?

Answer

In calculating the parameter w, you encode the label of each data point using their location on the x axis. To do this, you add 1 to the sum of the converse of each label multiplied by n^{x_i}. You then multiply this number by pi to put it in radians. You can use this number to decode the label of the data point by multiplying it by n^{-x_i}; when you put the result in the sin function, it will output a positive classification (result is between 0 and pi), or a negative classification (result is between pi and 2*pi) and then you classify the point based on whether it is positive or negative.

There are, however, examples of points that cannot be shattered by the classifier. For example, when you have four points that are equally spaced with corresponding labels +,+,-,+ , you get the situation where it will classify them all as positive. This is because the distance between points 2 and 4 is twice the distance between points 1 and 2 (resonant), and the resulting parameter w would cause the classification boundary to cross point 3 (giving it a positive classification). This does not break the VC Dimension though because there is always a configuration for n > 1 training examples that cannot be shattered (i.e. multiple training examples are coincident on the same point but with different labels).