DSCI 552 -- ASSIGNMENT -1

NAME: KANNA MOHAN SAI GANESH **USC ID:1288424487**

1. Alpaydin 4th edition, Chapter 2, Exercise 1. What is the VC dimension of a circle?

SOLUTION:

If we can shatter maximum of N points with our hypothesis class say H, but cannot shatter more than N points, then we say that the VC dimension of that particular hypothesis class H is N.

VC Dimension of a Circle is 3.

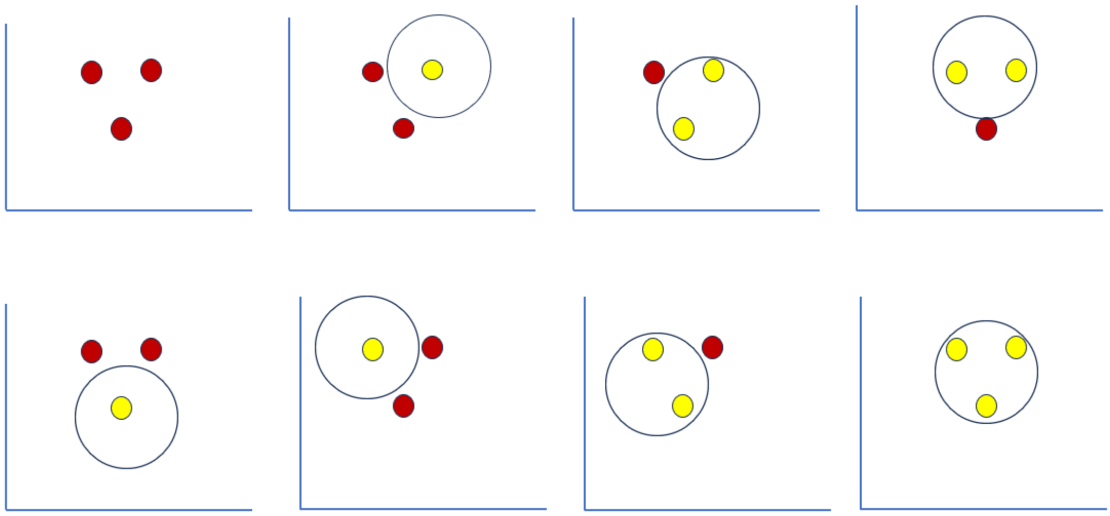
Let's consider a triangular arrangement as below and two classes A and B.

The instances of the circle hypothesis can successfully shatter the points into Class A and Class B.

The points with the yellow color are Class A (Positive). The points with the brown color are Class B (Negative).

The following are the 8 possibilities for the given structural arrangement of 3 points.

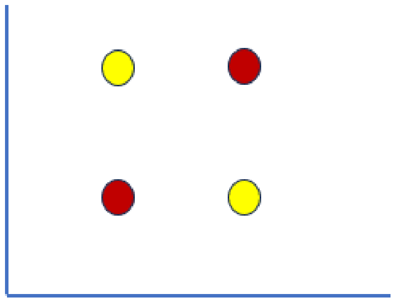
For N points we have 2^N possibilities.



However, the circle cannot shatter 4 points in any structural arrangement.

Therefore, the VC dimension of the circle is 3.

Following is the example that shows the structural arrangement of 4 points to prove that VC dimension of circle is not 4.



It is not possible to shatter 4 points for any structural arrangement apart from the above. which supports our claim for stating VC Dimension of Circle as 3.

2. Alpaydin 4th edition, Chapter 2, Exercise 2. Why would setting $m = N$ (where N is the number of positive instances) be a bad idea.

SOLUTION:

The hypothesis class consisting of union of multiple rectangles allows complex and flexible decision boundaries in the input space.

Capturing difficult patterns and relations within the data can be done with this flexibility.

Each rectangle represents a conjunction of conditions on the input attributes.

Having multiple rectangles allows for a disjunction of these conjunctions.

In the input space positive instances can be grouped into multiple clusters.

It allows the hypothesis class to represent diverse and complex decision regions.

Suppose, if the value of m is set to N, each positive instance has a separate rectangle.

This seems like a way to fit the training data perfectly, but there are few reasons why this could be a bad idea.

The following are the reasons satisfying the above statements.

1. **Computational complexity:**
   Having a separate rectangle for each positive instance may result in computationally large and expensive model. This might make the model inefficient in terms of both prediction and training times.
2. **Overfitting:**
   Having a separate rectangle for each positive instance might lead to overfitting. This model may capture the noise or specific features of the individual data points that are not generalized and may not be adaptable to the unseen data.
3. **Loss of Generalization:**
   The main objective of machine learning is to learn relationships and patterns that generalize to future or unseen data. By having own individual rectangle for each positive instance, the model may fail to generalize the new instances that are having similar patterns but are not identical to the training instances.

Therefore, increasing the m allows more expressive hypothesis classes, but setting $m = N$ might lead to increased computational complexity, overfitting and loss of generalization making it a secondary choice in practice.

3. Alpaydin 4th edition, Chapter 2, Exercise #6

In equation 2.13, we summed up the squares of the differences between the actual value and the estimated value. This error function is the one most frequently used, but it is one of several possible error functions. Because it sums up the squares of the differences, it is not robust to outliers. What would be a better error function to implement *robust regression*?

SOLUTION:

Given the equation 2.13 the **Mean Squared Error (MSE)**is as follows,

$$E(g|\, \chi) = \frac{1}{N}\sum_{t=1}^{N}[r^t - g(x^t)]^2$$

We assume that there is a Gaussian Noise by seeing the squared error.

If the distribution has long tails and a noise is identified from such a distribution, then by summing up the squared differences as provided from the equation 2.13, we may end up having the outliers that may corrupt the fitted line.

In order to decrease the effect of outliers, instead of squaring them we can use the absolute value of differences.

Below is the formula for the **Mean Absolute Error** (**MAE**).

$$E(g|\, \chi) = \frac{1}{N}\sum_{t=1}^{N}|\, r^t - g(x^t)\,|$$

$g(x^t)$ -- The predicted value of the t^{th} instance based on the input x.

$E(g|\, \chi)$ -- Expectation or mean value of g for the given input features χ.

N – Number of data points in the training data.

$\sum_{t=1}^{N}$ -- Summation of all data points of the train data, starting from to N, indexed by t.

r^t – The ground truth for the t^{th} data point.

$|\, r^t - g(x^t)\,|$ -- Absolute difference between the actual and predicted value of t^{th} data point.

MSE is more sensitive to the outliers compared to that of MAE.

There is a robust solution that combines the properties of both MSE and MAE known as **Huber Loss**.

The Huber Loss function is as follows:

$$E(g|\,\chi) = \begin{cases} \frac{1}{N}\sum_{t=1}^{N}\frac{1}{2}.(r^t - g(x^t)^2 & ,if\ |r^t - g(x^t)| \leq\ \delta \\ \frac{1}{N}\sum_{t=1}^{N}\delta.(|r^t - g(x^t)|\ - \frac{1}{2}\delta) & ,if\ |r^t - g(x^t)|\ >\ \delta \end{cases}$$

δ – Transition point between quadratic and linear regions of loss function.

Larger values of δ makes the Huber Loss function more robust to the outliers.

4. Alpaydin 4th edition, Chapter 2, Exercise #7

Derive equation 2.17.

$$
\begin{aligned}
w_1 &= \frac{\sum_t x^t r^t - \overline{x}\,\overline{r}N}{\sum_t (x^t)^2 - N\overline{x}^2} \\
w_0 &= \overline{r} - w_1\overline{x} \qquad (2.17)
\end{aligned}
$$

where $\overline{x} = \sum_t x^t/N$ and $\overline{r} = \sum_t r^t/N$. The line found is shown in figure 1.2.

SOLUTION:

Minimum point or maximum point for a graph can be found at the point where slope of the graph at that point is 0.

So, minimum point can be calculated by taking the partial derivatives of E with respect to w_1 and w_0 and equating it to 0.

Given equation, $E(w_1, w_0 | \chi) = \frac{1}{N}\sum_{i=1}^{N}[r^t - (w_1x^t + w_0)]^2$

Applying partial derivative on above equation with respective to w_0.

$\frac{\partial E}{\partial w_0} = \sum_t\ [r^t - (w_1x^t + w_0)] = 0$

$\sum_t r^t - w_1\sum_t x^t - \Sigma_t w_0 = 0$

$$N\,w_0 = \sum_t r^t - w_1\sum_t x^t$$

$w_0 = \sum_t \frac{r^t}{N} - w_1\sum_t \frac{x^t}{N}$

$w_0 = \overline{r} - w_1\overline{x}$

Therefore, $\boldsymbol{w_0 = \overline{r} - w_1\overline{x}}$. Here, $\overline{r}$ *and* $\overline{x}$ are the corresponding means.

Applying partial derivative on above equation with respective to w_1.

$\frac{\partial E}{\partial w_1} = \sum_t\ [r^t - (w_1x^t + w_0)]\,x^t = 0$

$\Sigma_t\, r^t x^t - w_1\sum_t (x^t)^2 - w_0\sum_t x^t = 0$

$\sum_t r^t x^t = w_1\sum_t (x^t)^2 + w_0\sum_t x^t$

Substituting the value of w_0, we get

$\sum_t r^t x^t = w_1\sum_t (x^t)^2 + (\overline{r} - w_1\overline{x})\sum_t x^t$

$\sum_t r^t x^t = w_1(\sum_t (x^t)^2 - \overline{x}\sum_t x^t) + \overline{r}\sum_t x^t$

$\sum_t r^t x^t = w_1(\sum_t (x^t)^2 - \overline{x}N\overline{x}) + \overline{r}N\overline{x}$

On rearranging the terms, we get,

$$w_1 = \frac{\sum_t x^t r^t - \overline{x}\overline{r}N}{\sum_t (x^t)^2 - N\overline{x}^2}$$

Therefore, $\boldsymbol{w_1} = \frac{\sum_t \boldsymbol{x^t r^t} - \boldsymbol{\overline{x}\overline{r}N}}{\sum_t (\boldsymbol{x^t})^2 - \boldsymbol{N\overline{x}^2}}$ **.** Here, $\overline{r}$ *and* $\overline{x}$ are the corresponding means.

5. Alpaydin 4th edition, Chapter 2, Exercise #9

Assume our hypothesis class is the set of lines, and we use a line to separate the positive and negative examples, instead of bounding the positive examples as in a rectangle, leaving the negatives outside (see figure 2.13). Show that the VC dimension of a line is 3.

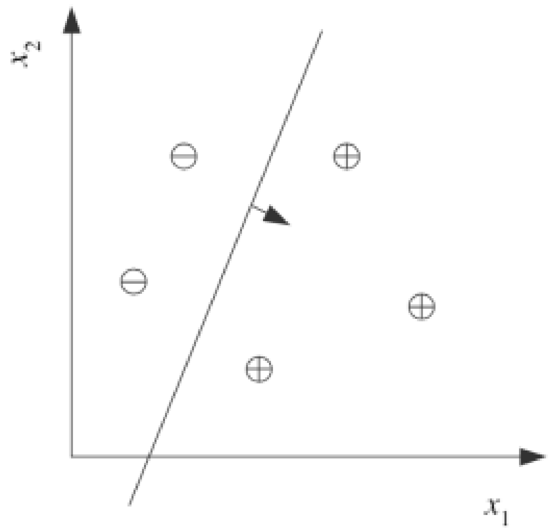


Figure 2.13 A line separating positive and negative instances.

SOLUTION:

If we can shatter maximum of N points with our hypothesis class say H, but cannot shatter more than N points, then we say that the VC dimension of our hypothesis class H is N.

VC Dimension of a Line is 3.

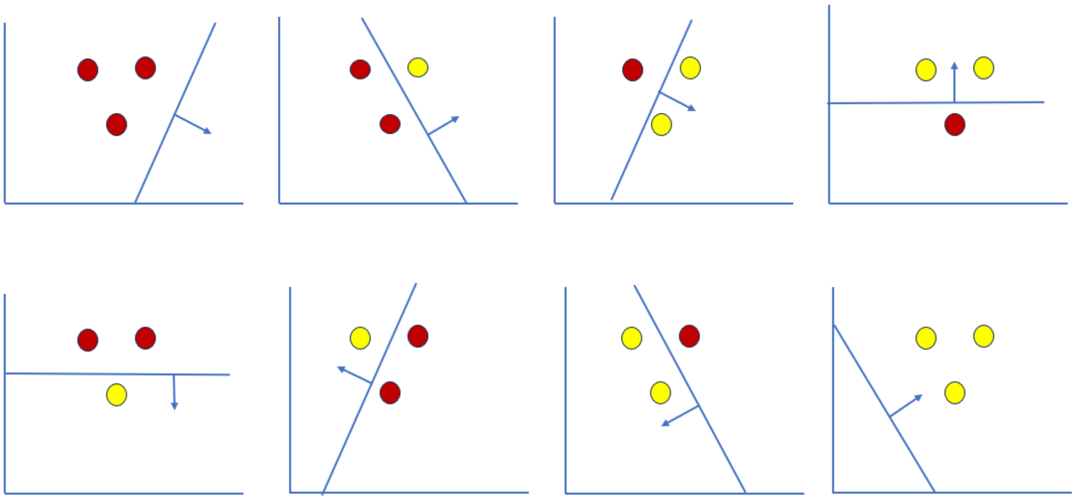
Let's consider a triangular arrangement as below and two classes A and B.

The instances of the line hypothesis can successfully shatter the points into Class A and Class B.

The points with the yellow color are Class A (Positive). The points with the brown color are Class B (Negative).

The following are the 8 possibilities for the given structural arrangement of 3 points.

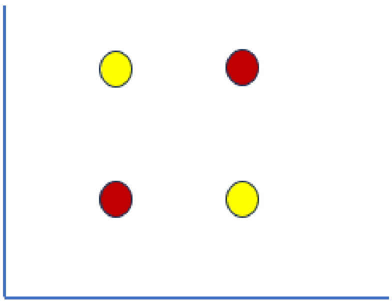
For N points we have 2^N possibilities.



However, the line cannot shatter 4 points in any structural arrangement.

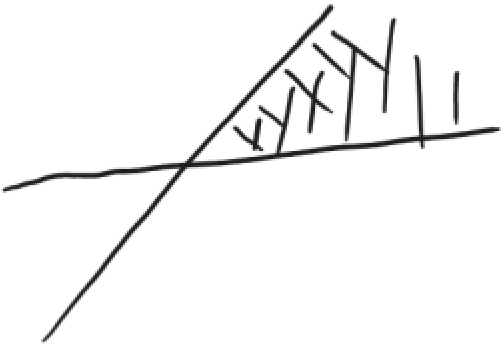
Therefore, the VC dimension of the line is 3.

Following is the example that shows the structural arrangement of 4 points to prove that VC dimension of line is not 4.



It is not possible to shatter 4 points for any structural arrangement apart from the above. which supports our claim for stating VC Dimension of Line as 3.

6. Bonus: Let a wedge shape be defined as the intersection of two half-spaces, see the image below. Show the VC dimension of a wedge is 5. Hint: place the five points equidistant on a circle.



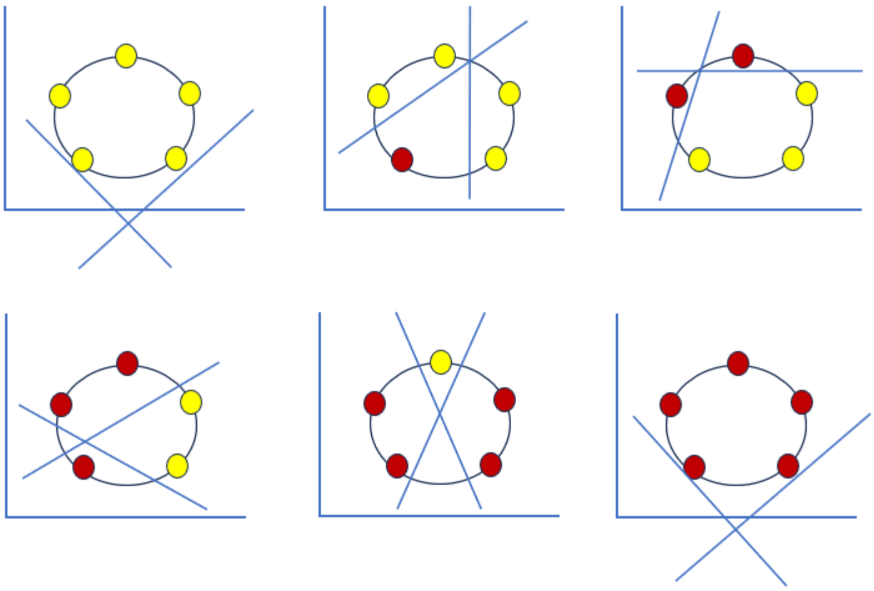
SOLUTION:

Wedge shape (intersection of two half spaces) has a VC Dimension of 5.

Considering a pentagon like structural arrangement where points are arranged equidistant on a circle, the instances of our wedge hypothesis can successfully shatter the 5 points into 2 classes say Class A and Class B.

Let Class A be positive and Class B be negative.

Following are the few of 32 possibilities. As we are having 5 points, the possibilities are 2^5.

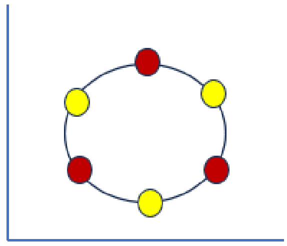


The above diagram depicts how given 5 points are arranged

1. 5 Positive, 0 Negative
2. 4 Positive, 1 Negative
3. 3 Positive, 2 Negative
4. 2 Positive, 3 Negative
5. 1 Positive, 4 Negative
6. 0 Positive, 5 Negative

Though we scramble the labels for the above possibilities, we end up shattering 5 points in all cases for the given arrangement, where all the 5 points are equidistantly arranged in circular fashion.

When 6 points are taken and arranged them as below in a circular fashion, we see that it is not possible to shatter 6 points using a wedge instance from the given Wedge Hypothesis.



Apart from the above example, it is not possible for any structural arrangement of 6 points.

Therefore, our claim is true that the VC Dimension of the given wedge is 5.