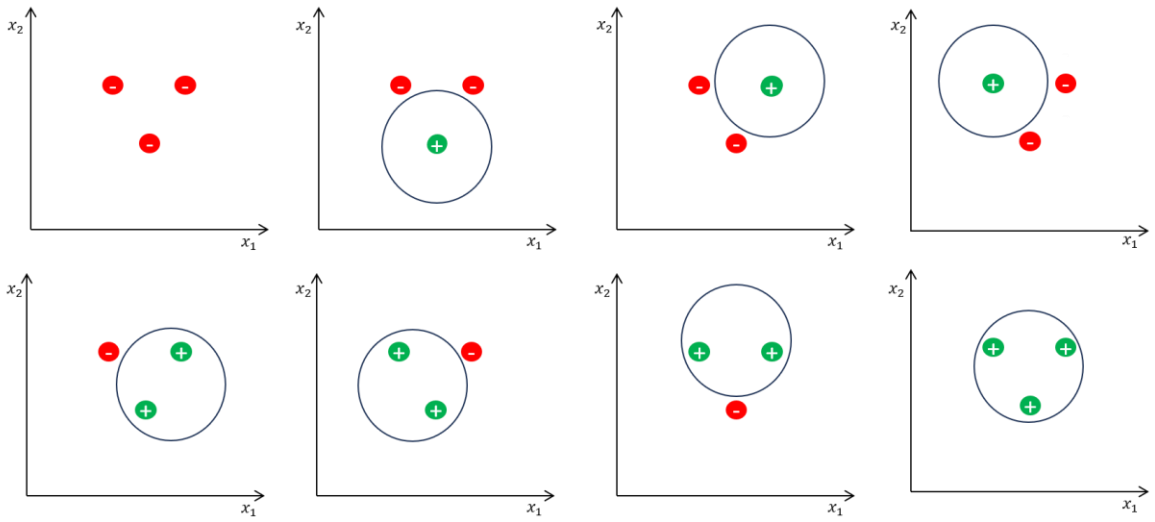
DSCI 552 – MACHINE LEARNING FOR DATA SCIENCE
HOMEWORK – 1

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

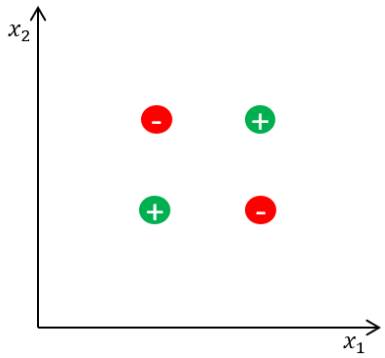
ANS:

VC Dimension of a Circle is 3. For a particular structural arrangement (triangular in our case), hypothesis h of Circle Hypothesis Class H ($h \in H$) can successfully shatter 3 points into Class A (positive) and Class B (negative). Below are all 8 possibilities (2^N) for arranging of 3 points.



If we can shatter N points using a hypothesis class H and we cannot shatter $N + 1$ points using that hypothesis class, then we can say that VC dimension of that particular hypothesis class H is N.

A circle instance cannot shatter 4 points in any structural arrangement. Hence, VC Dimension of a Circle is 3. An example with a default structural arrangement (square) is given below



It is not possible to shatter 4 points using a circle instance from Circle Hypothesis. Apart from the above example, it is not possible for any structural arrangement of 4 points. This 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.

ANS:

The advantage of using a hypothesis class that consists of the union of multiple rectangles is that it allows for more complex and flexible decision boundaries in the input space. This flexibility enables us to capture difficult patterns and relations within the data.

In this context, each rectangle represents a conjunction of conditions on the input attributes. Having multiple rectangles allows for a disjunction of these conditions. This is similar to expressing logical formulas as a disjunction of conjunctions. The positive instances can be grouped into multiple clusters in the input space, allowing the hypothesis class to represent more diverse and complex decision regions.

Now, regarding the idea of setting $(m = N)$ (where N is the number of positive instances), this means having a separate rectangle for each positive instance. While this might seem like a way to perfectly fit the training data, there are several reasons why it could be a bad idea:

1. Overfitting: Creating a separate rectangle for each positive instance could lead to overfitting. The model might capture noise or specific characteristics of individual data points that do not generalize well and adapt to unseen data

2. Loss of Generalization: The goal of machine learning is to learn patterns and relationships that generalize to unseen data. If each positive instance has its own rectangle, the model might fail to generalize to new instances that have similar patterns but are not identical to the training instances.

3. Computational Complexity: Having a separate rectangle for each positive instance could result in a large and computationally expensive model. This could make the model inefficient in terms of both training and prediction times.

In summary, while increasing m allows for more expressive hypothesis classes, setting $m = N$ might lead to overfitting, loss of generalization, and increased computational complexity, making it a secondary (not optimal) choice in practice.

Let us consider a real-world example involving the classification of objects in images. The task is to distinguish between positive instances (Ex: images of cats) and negative instances (Ex: images other than cats). We will use rectangles as our basic building blocks for the hypothesis class.

Example: Cat Classification

Single Rectangle:

Suppose we have a simple hypothesis class consisting of a single rectangle. This rectangle represents a conjunction of conditions on two features extracted from the images (Ex: pixel intensity values).

Single Rectangle Hypothesis: If pixel intensity in a certain range and another feature within specific bounds, classify as a positive instance (cat). Otherwise, classify as a negative instance.

However, this simple decision boundary might not capture the complexity of cat images, leading to misclassifications.

Union of Rectangles:

Now, consider a more flexible hypothesis class that allows the union of rectangles. Each rectangle corresponds to a conjunction of conditions on the features. This flexibility can help in capturing different patterns in different regions of the feature space.

Union of Rectangles Hypothesis: Combine rectangles to form a more complex decision boundary that adapts to different features associated with cat images.

In this example, the hypothesis class with the union of rectangles allows for a more adaptive and nuanced decision boundary, potentially improving the classification accuracy on a variety of cat images.

In a real-world scenario, using a more expressive hypothesis class can be beneficial when dealing with diverse and complex patterns in data. This allows the model to capture different attributes of the positive instances. This flexibility is crucial for achieving better generalization to unseen examples.

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*?

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

ANS:

By seeing the squared error, we can assume that there is Gaussian Noise. If a noise is identified from a distribution that has long tails, then by summing up the squared differences we end up with few outliers that corrupt the fitted line (Mean Squared Error – MSE {quadratic loss}). To decrease the effect of outliers, we can sum up the absolute value of differences instead of squaring them. For robust regression, we can use

1. **Mean Absolute Error (MAE):**

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

- $E(g|X)$: Expectation or mean value of the predicted values g given the input features X.
- N: Number of instances or data points in the training set.
- $\sum_{t=1}^{N}$: Summation over all instances in the training set, indexed by t.
- r^t: The actual (ground truth) value for the t-th instance.
- $g(x^t)$: The predicted value for the t-th instance based on the input x.
- $|r^t - g(x^t)|$: Absolute difference between the actual and predicted values for the t-th instance.

MSE is very sensitive to the outliers compared to absolute loss (Mean Absolute Error – MAE {{linear loss}}). In addition to the above, we can also use

2. **Huber Loss**

Huber Loss is a function that has the combined properties of both MSE and MAE which provides a more robust solution when we encounter outliers during regression. Huber Loss is less sensitive to the outliers and can be depicted as follows

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

- 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.

Its minimum point can be calculated by taking the partial derivatives of E with respect to w_1 and w_0, setting them equal to 0, and solving for the two unknowns:

$$w_1 = \frac{\sum_t x^t r^t - \overline{x}\,\overline{r}N}{\sum_t (x^t)^2 - N\overline{x}^2}$$
$$w_0 = \overline{r} - w_1\overline{x}$$

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.

ANS:

$$E(w_1, w_0 | X) = \frac{1}{N}\sum_{t=1}^{N}[r^t - (w_1 x^t + w_0)]^2$$

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

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

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

$$\boxed{=> \boldsymbol{w_0} = \overline{\boldsymbol{r}} - \boldsymbol{w_1}\overline{\boldsymbol{x}}} \ldots\ldots\ldots\ldots\ldots \text{eq (1)}$$

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

$$=> \sum_t r^t x^t = w_1 \sum_t (x^t)^2 + w_0 \sum_t x^t \ldots\ldots\ldots\ldots\ldots \text{eq (2)}$$

Substituting eq (1) in eq (2)

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

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

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

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

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.

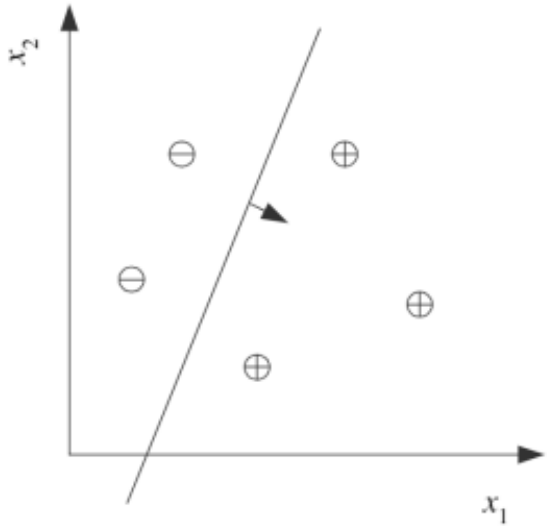
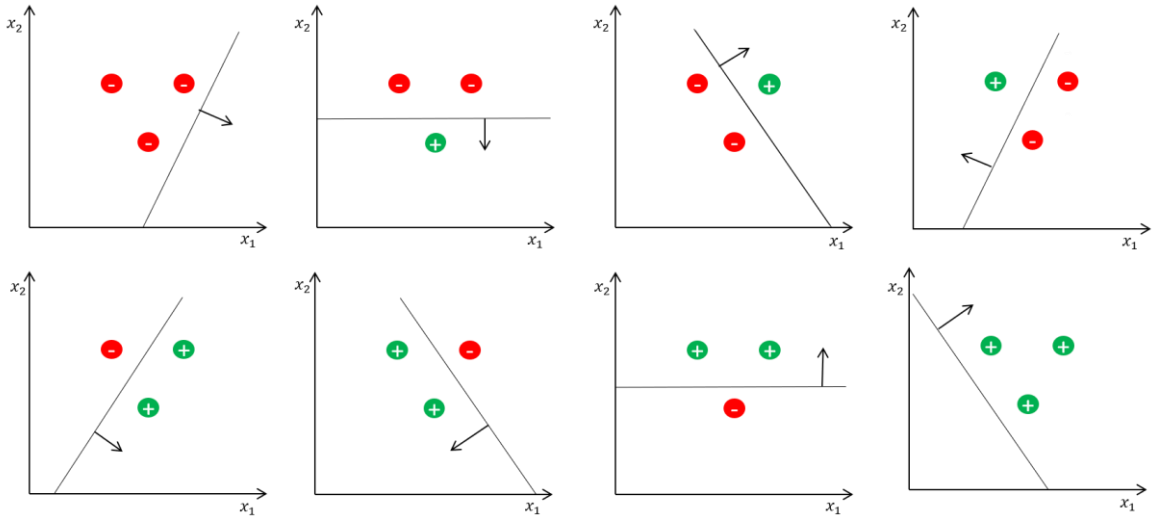


Figure 2.13 A line separating positive and negative instances.

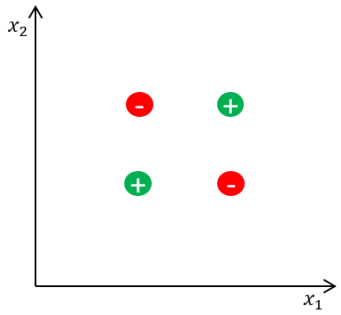
ANS:

VC Dimension of a Line is 3. For a particular structural arrangement (triangular in our case), hypothesis h of Line Hypothesis Class H ($h \in H$) can successfully shatter 3 points into Class A (positive) and Class B (negative). Below are all 8 possibilities (2^N) for arranging of 3 points.



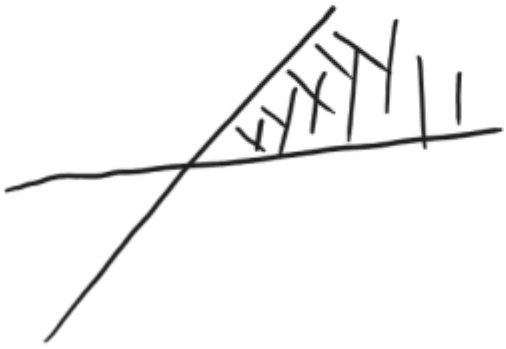
If we can shatter N points using hypothesis h from hypothesis class H ($h \in H$), and we cannot shatter $N + 1$ points using that hypothesis class, then we can say that VC dimension of that particular hypothesis class H is N.

Line cannot shatter 4 points in any structural arrangement. Hence, VC Dimension of a Line is 3. An Example with a default structural arrangement is given below



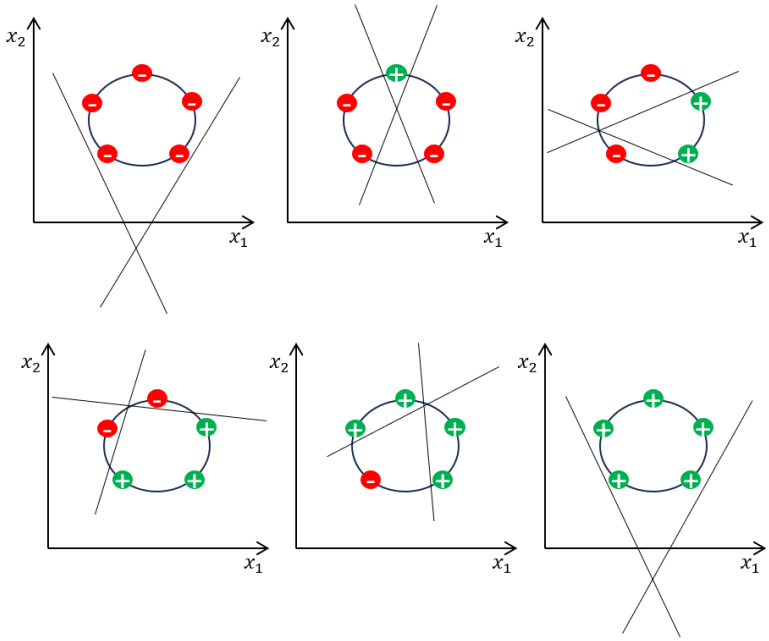
It is not possible to shatter 4 points using a line instance from Line Hypothesis. Apart from the above example, it is not possible for any structural arrangement of 4 points. This supports our claim for stating VC Dimension of Line as 3.

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.



ANS:

VC Dimension of the given wedge shape (intersections of two half-spaces) is 5. For the given pentagon-like structural arrangement (points are arranged equidistant on a circle), instances of our wedge (intersection of tow half-spaces) hypothesis can successfully shatter the 5 points into Class A (positive) and Class B (negative). Below are all few of the 32 total possibilities (2^N) for the given arrangement of 5 points.

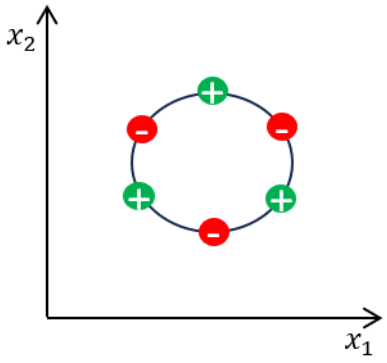


The above figure depicts how given 5 points when arranged as per following samples

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

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

Now, let's see if the given wedge hypothesis can shatter 6 points. Let's take 6 points arranged in circular fashion equidistantly as our default example.



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. This supports our claim for stating VC Dimension of the given wedge as 5.

********** **THE END** **********