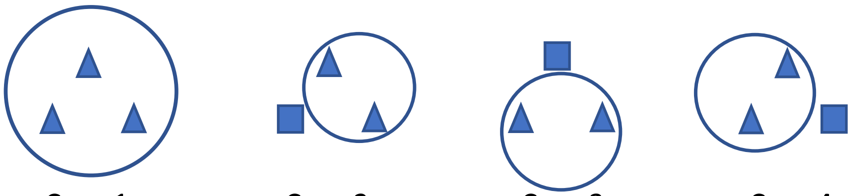
CSCI3320 Written Homework 2 Chung Tsz Ting 1155110208

1) VC dimension is defined by the cardinality of the largest set of points the hypothesis H can shatter.
   a) The VC dimension is 3. For example, giving a set of points of { (0,0), (2,0), (1,1) }, there 8 configurations of labelling, i.e. {P, P, P}, {P, P, N}, {P, N, P}, {N, P, P}, {P, N, N}, {N, P, N}, {N, N, P} and {N, N, N} for P stands for positive examples and N stands for negative examples. As the center position and radius of the circle can be varied to a suitable value, the 8 configurations can be classified.
   Let ▲ be positive example and ■ be negative example.



Case 1: Case 2: Case 3: Case 4:

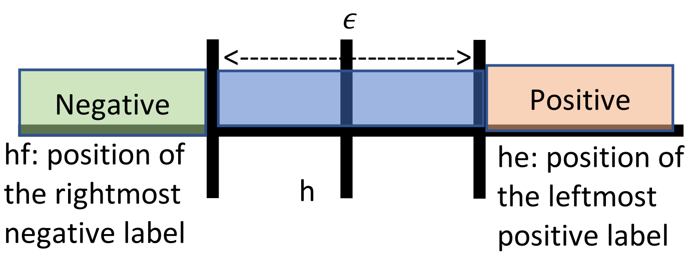
Case 5: Case 6: Case 7: Case 8:

   b) The VC dimension is also 3. The reason is same as above by making positive example as negative and vice versa.

2) The VC dimension is 2 as the largest set of points the hypothesis can shatter is 2 in which there are 2 configurations {P, N} and {N, P} for P stands for positive example and N stands for negative examples. One end point of the interval can be set between the two examples and the other end can be set on the other side of the positive example.
   The reason of VC dimension cannot be 3 (which have to satisfy the 8 configurations form by 3 examples of point with 2 classes in total) is illustrated in the below graphical representation.
   Let ▲ be positive example and ■ be negative example.

   There cannot be a interval include both of the two positive examples.

3)



   According to the hypothesis which is to place the half-line anywhere in the transition region form examples with negative label to those with positive labels, there can be a gap between the position of the rightmost example of the negative set and the position of the leftmost example of positive set.

As shown in the above picture, the green area contain all the given negative labelled examples while the red area contain all the given positive labelled examples. ‘hf’ is the leftmost vertical line which is the position of the rightmost negative example; while ‘he’ is the rightmost vertical line which is the position of the leftmost positive example. And the hypothesis will place h in between the two line.

Because of the unsampled point, the region in between hf and he is not known to be as positive or negative. Thus, the gap between is the error regions.
For ϵ good hypothesis, the probability of having examples into the interval between hf and he should be less than ϵ (i.e. the interval distance be at most ϵ).

To ensure the hypothesis will be ε good with high probability of $1-\delta$, we have $\Pr\{error\ region \ \leq \ \epsilon\} \geq 1-\delta$ which is equivalent to $\Pr\{error\ region \ > \ \epsilon\} \leq \delta$.

There are 2 cases for the happening of error, (assuming the actual correct transition position be t)

- The hypothesis place h to the left of t.
- The hypothesis place h to the right of t.

Each of them have their error region be at most ϵ for a good ϵ hypothesis.
Assume there are n number of given examples. For a ϵ bad hypothesis, the h should fall outside the region of error of the interval.

$$\Pr\{error\ region \ > \ \epsilon\}$$
$$\leq \sum_{i=1}^{2} \Pr\{no\ example\ falling\ into\ the\ interval\ of\ error\ of\ case\ i\}$$
$$\leq \sum_{i=1}^{2} \Pr\{no\ example\ falling\ into\ the\ interval\ of\ error\}$$
$$\leq \sum_{i=1}^{2} (1-\epsilon)^n$$
$$\leq 2\,(1-\epsilon)^n$$

$$By\ 1+x \ \leq \ e^x, \quad \Pr\{error\ region \ > \ \epsilon\} \leq 2\, e^{-\epsilon n}$$

$$\Pr\{error\ region \ > \ \epsilon\} \leq \delta$$
$$2\, e^{-\epsilon n} \ \leq \ \delta$$
$$ln(2\, e^{-\epsilon n}) \ \leq \ ln\delta$$
$$ln2 - \ \epsilon n \ \leq \ ln\delta$$
$$n \geq \frac{1}{\epsilon}\ ln\frac{2}{\delta}$$

Thus, the minimum number of sample points we need for is $\frac{1}{\epsilon}\ ln\frac{2}{\delta}$ that the hypothesis is PAC-learnable.