CS57800: Statistical Machine Learning
Homework 4

Due: Dec 05, 2018 on Wednesday

1 Boosting

We first initialize the weights of all the examples as 1/10 since there are 10 examples and all are weighted equally at the start of the first iteration. We normalize the weights so that they sum to 1 thus assigning 1/10 weight to each example.

1.1 Iteration 1

In the first round, we select the weak hypothesis H_1 as:

$$H_1 = x_1 > 6$$

We classify points as $+1$, if they obey the weak hypothesis otherwise we classify them as -1.
The distribution D_1 of points is such that all have weights $1/10$.
We get incorrect prediction for 2 points i.e. $(8, 8)$ and $(4, 10)$. The error for this hypothesis is

$$Error^1 = 1/10 * 1 + 1/10 * 1$$

i.e.

$$Error^1 = 0.2$$

Now, we calculate the value of alpha. We have

$$\alpha_1 = 1/2 * ln((1 - Error^1)/Error^1)$$

i.e.

$$\alpha_1 = 1/2 * ln(0.8/0.2)$$

i.e.

$$\alpha_1 = 1/2 * ln4$$

.

Finally, we update the weights for both correctly classified points by downvoting them i.e. for all points except $(8, 8)$ and $(4, 10)$, the weight update is shown below:

$$d_1 = 1/10 * e^{-1/2*ln4*1*1}$$

i.e.

$$d_1 = 1/10 * e^{-ln2}$$

i.e.

$$d_1 = 1/20$$

and upvote the weights for incorrectly classified points i.e. for points $(8, 8)$ and $(4, 10)$ as shown below:

$$d_2 = 1/10 * e^{-1/2*ln4*-1*1}$$

i.e.

$$d_2 = 1/10 * e^{ln2}$$

i.e.

$$d_2 = 1/5$$

Now, we find the normalization factor Z for the above mentioned updated weights as:

$$Z = 1/20 * 8 + 1/5 * 2 = 4/5$$

Normalizing d_1 and d_2 by dividing by Z, we get

$$d_1 = 1/20 * 5/4 = 1/16$$

$$d_2 = 1/5 * 5/4 = 1/4$$

1.2 Iteration 2

In the second iteration, we select the weak hypothesis H_2 to be

$$H_2 = x_2 > 8$$

Again, we classify points as $+1$ if they obey the hypothesis and otherwise mark them as -1.
The distribution D_2 of points in iteration 2 is such that $(8, 8)$ and $(4, 10)$ have weights $1/4$. The remaining points have weights $1/16$. Using this, we incorrectly classify 4 points which have weights $1/16$ each. The points are $(1, 10), (3, 16)$ and $(7, 7), (8, 7)$
The error for this hypothesis is:

$$Error^2 = 1/16 * 2 + 1/16 * 2 = 0.25$$

Now, we calculate the value of alpha. We have

$$\alpha_2 = 1/2ln(0.75/0.25) = 1/2ln3$$

.

1.3 Final Hypothesis

The final hypothesis that we have after 2 rounds is

$$H_{final}(x) = sign(\alpha_1 H_1(x) + \alpha_2 H_2(x))$$

i.e.

$$H_{final}(x) = sign(1/2 \text{ ln } 4 \ H_1(x) + 1/2 \text{ ln } 3 \ H_2(x))$$

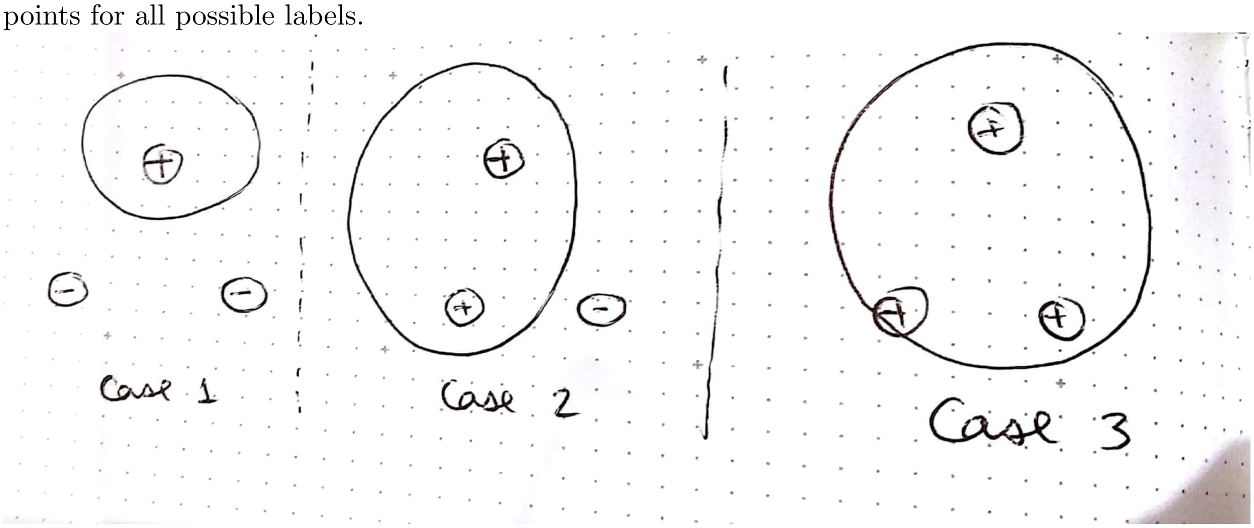
i.e.

$$H_{final}(x) = sign(0.693H_1(x) + 0.549H_2(x))$$

2 PAC Learning

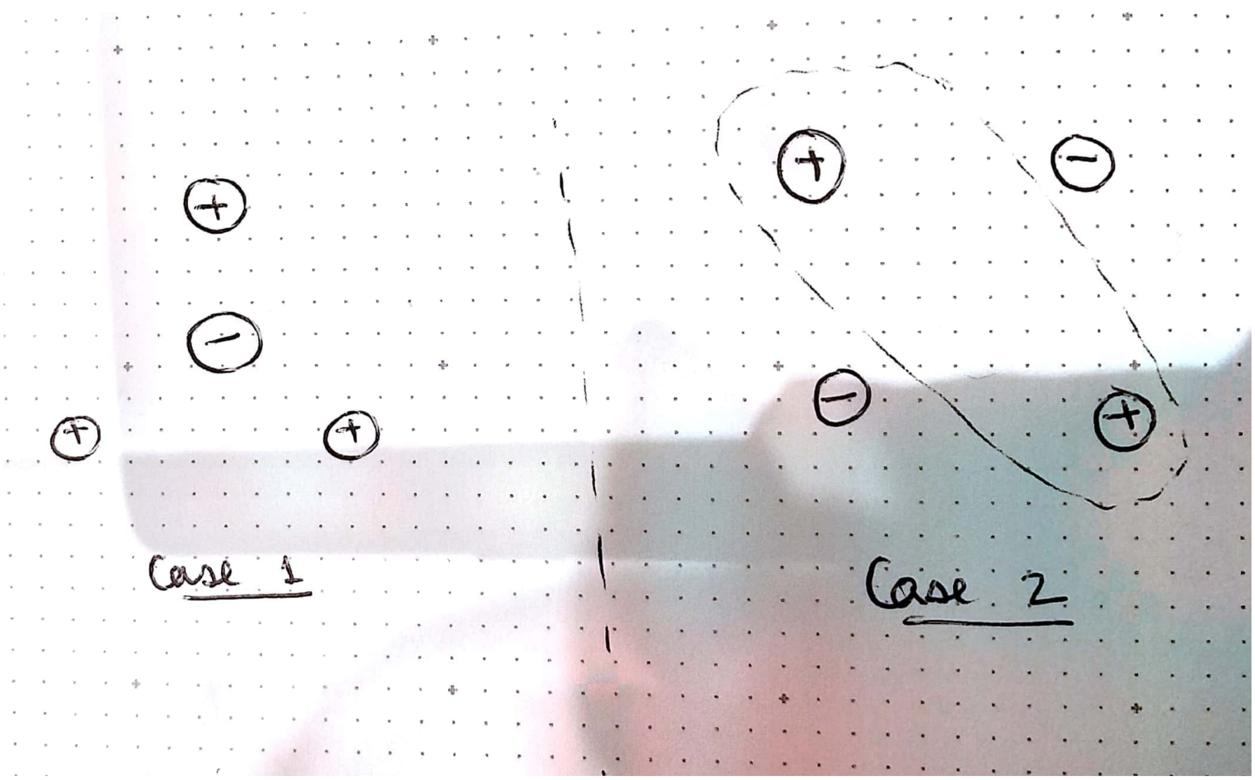
2.1 Circles

Circles can correctly classify 3 points which dont lie on a straight line. The possible labels for such set of 3 points is given below. As we can see from figure below, a circle can shatter such 3 points for all possible labels.



Now, we show that VC dimension of circles is 3 since they can't shatter set of 4 points. We observe that when we add another point to the previous set of 3 points, then either the 4th point can lie within the triangle formed by the previous 3 points or it can combined with the other 3 points form a convex polygon with 4 sides.

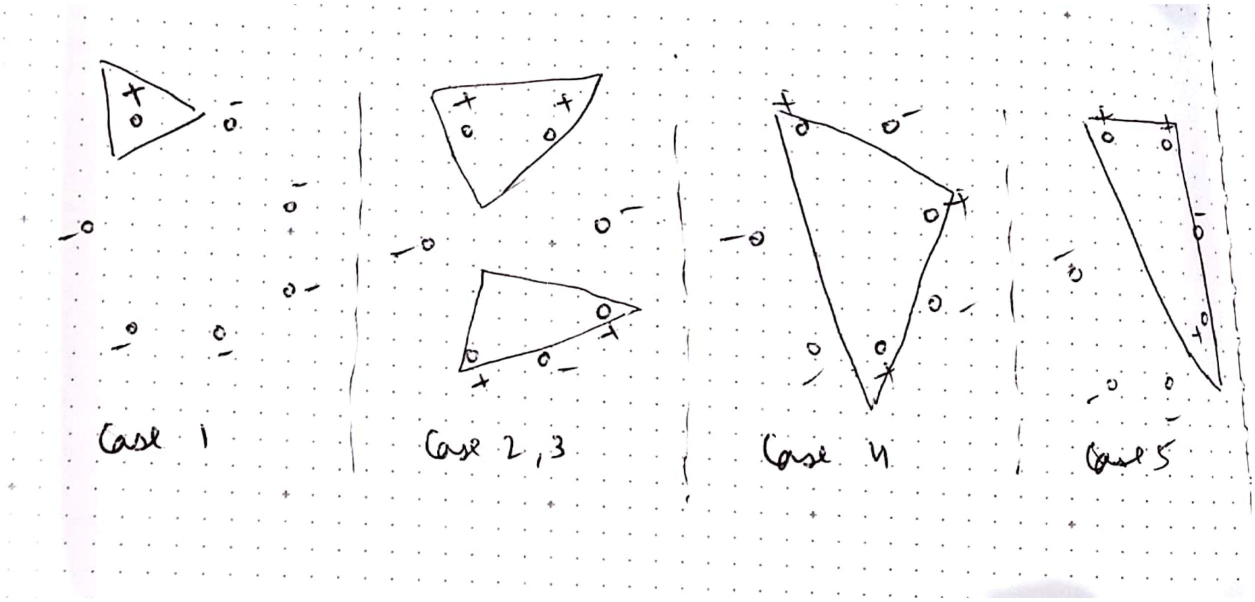
As shown below, if it lies inside the triangle, then the circle can't shatter the below mentioned labelling.



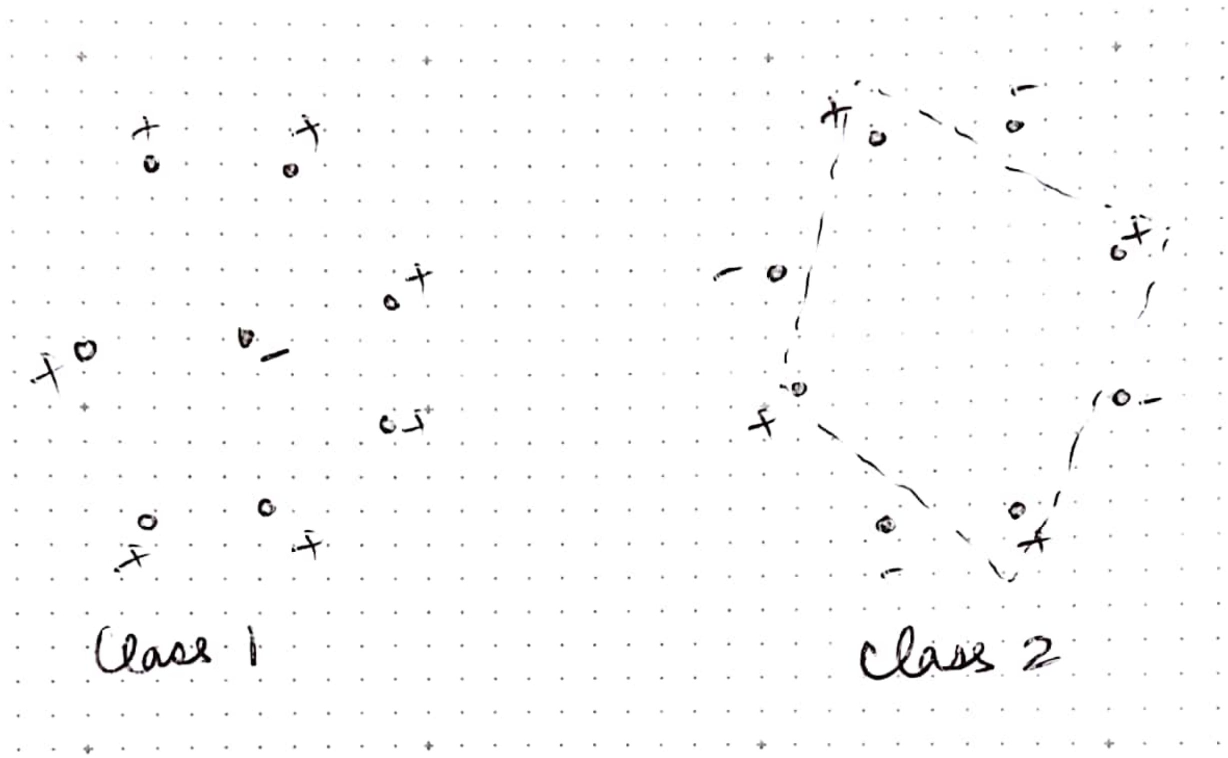
Also as shown above, it it forms a convex polygon, then it wont be able to shatter a labelling whenever the 2 diagonally opposite points which have larger or equal length diagonal are labelled + since it would require an eclipse shape to have them inside which is not possible using a circle.

2.2 Triangles

We can see that Triangles can correctly classify 7 points which form a convex polygon with 7 sides. As we can see from figure below, there can be cases such that there is 1 positive label (Case 1), 2 positive labels which are neighbours or not (Case 2,3) , 3 positive labels where the points are adjacent or have negative points between them (Case 4,5), 4 positive labels where there must be atleast 2 points which are adjacent to each other and so on. Now, the main observation is that we can always draw a triangle which separates these positive points by keeping them inside the triangle and thus its able to always shatter the points. The main reason why its able to shatter is that there are no three points which form a straight line so its always able to avoid the negative labels which come in between the positive labels while drawing the triangle.



Now, we show that the VC Dimension of triangles is 7 since its not able to shatter any set of 8 points. Now, extending from set of 7 points, the 8th point could be either inside the convex polygon formed by 7 points or it could help form a convex polygon with 8 sides. As we can see from the figure shown below, its not able to shatter if the 8th point is inside.



Coming to the second case, the triangle will not be able to shatter the labelling which has none of the positive labels adjacent to each other i.e. the positive labels cannot be enclosed in a triangle without enclosing a negative label. Hence, VC Dimension of triangles is 7.

2.3 Trees

- The number of syntactically different trees are

  $$n * (n-1) * 2^4$$

  i.e.

  $$n * (n-1) * 16$$

  . The reason being that we can have n possible values at the root since there are n variables. Now, once we choose a variable at the root, we cannot use it at child nodes. Also, since both the child nodes need to be split using the same variable therefore we have $n-1$ possible values. Finally, for the leaf nodes, at each leaf node, we have two possible labels $+$ or $-$. Since, there are 4 leaf nodes, therefore total ways is 2^4. Thus, our answer is $n * (n-1) * 2^4$.

- Let, the probability of finding a bad hypothesis be represented as δ. Now, we know that the minimum number of examples represented by m needed for a consistent PAC learnable model is:

  $$m > 1/\epsilon\{ln(|H|) + ln(1/\delta)\}$$

  i.e.

  $$m > 1/\epsilon\{ln(n * (n-1) * 16)) + ln(1/\delta)\}$$

  Finally, since δ represents the probability of finding a bad consistent hypothesis, therefore, confidence $\sigma = 1 - \delta$. Substituting δ by $1 - \sigma$, in the above equation, we get,

  $$m > 1/\epsilon\{ln(n * (n-1) * 16)) + ln(1/(1-\sigma))\}$$

  Now, suppose we have two variables i.e. $n = 2$, we get,

  $$m > 1/\epsilon\{ln(32)) + ln(1/(1-\sigma))\}$$