ECE 2372 - Homework 3

Jiyang Liu 4731134

Problem 1: [Convergence of PLA] Prove that the perceptron learning algorithm will converge (eventually) to a linear separator for linearly separable data set. For mathematical simplicity assume that the starting point is $\theta^0 = 0$ and for iterations $j \geq 1$, the algorithm proceeds by setting

$$\theta^j = \theta^{j-1} + y_{i_j}\tilde{x}_{i_j}$$

Here $(\tilde{x}_{i_j}, y_{i_j}$ represents the input/output pair that is misclassified by θ^{j-1}. We hope that as we proceed with the iterations $j \geq 1$, the estimate θ^j converges to θ^* which separates the data. We want to find a finite j where this eventually happens which we can record as the upper bound for the algorithm's converge.

(a). Suppose that θ^* is normalized so that $\rho = min_i \mid \langle\theta^*, \tilde{x}_i\rangle \mid$ calculates the distance between the closest x_i in the training data to the hyperplane defined by θ^*. Please argue that

$$min_i\ y_i\langle\theta^*, \tilde{x}_i\rangle = \rho > 0$$

Sol.

We have $y_i = \{-1, 1\}$.

$$If\ y_i = 1, \omega^T x_i + b \geq \rho$$
$$If\ y_i = -1, \omega^T x_i + b \leq -\rho$$

And also we know that,

$$\langle\theta^*, \widetilde{x_i}\rangle = \omega^T x_i + b$$

From above,

$$min_i y_i\langle\theta^*, \widetilde{x_i}\rangle = \rho > 0$$

(b). Show by induction that

$$\langle\theta^j, \theta^*\rangle \geq \langle\theta^{j-1}, \theta^*\rangle + \rho$$

and conclude that $\langle\theta^j, \theta^*\rangle \geq j\rho$

Sol.

If $j = 1$,

$$\langle\theta^1, \theta^*\rangle = \langle\theta^0 + y_{i_1}\widetilde{x_{i_1}}, \theta^*\rangle$$

For inner products, we can use the additive distributive law,

$$\langle\theta^1, \theta^*\rangle = \langle\theta^0, \theta^*\rangle + \langle y_{i_1}\widetilde{x_{i_1}}, \theta^*\rangle$$
$$= 0 + y_{i_1}\langle\widetilde{x_{i_1}}, \theta^*\rangle$$
$$= y_{i_1}\langle\widetilde{x_{i_1}}, \theta^*\rangle$$

From (a),

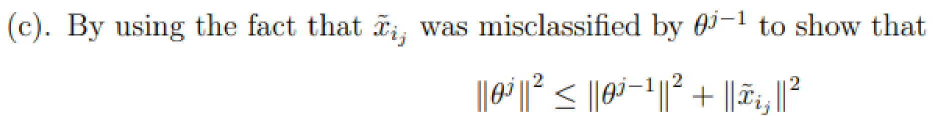
$$\langle\theta^1, \theta^*\rangle = y_{i_1}\langle\widetilde{x_{i_1}}, \theta^*\rangle \geq 1 * \rho$$

If $j = j + 1$, *assume that* $\langle\theta^j, \theta^*\rangle \geq j\rho$,

$$\langle\theta^{j+1}, \theta^*\rangle = \langle\theta^j + y_{i_{j+1}}\widetilde{x_{i_{j+1}}}, \theta^*\rangle = \langle\theta^j, \theta^*\rangle + y_{i_{j+1}}\langle\widetilde{x_{i_{j+1}}}, \theta^*\rangle$$

$$\geq j\rho + \rho = (j + 1)\rho$$

Therefore we proved $\langle \theta^j, \theta^* \rangle \geq j\rho$.



(c). By using the fact that $\tilde{x}_{i_j}$ was misclassified by θ^{j-1} to show that

$$\|\theta^j\|^2 \leq \|\theta^{j-1}\|^2 + \|\tilde{x}_{i_j}\|^2$$

Sol.

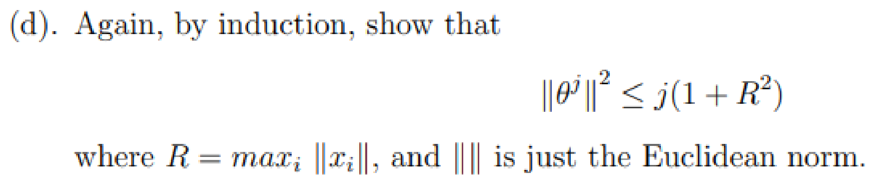
$$\|\theta^j\|^2 = \langle \theta^j, \theta^j \rangle$$

$$= \langle \theta^{j-1} + y_{i_j}\tilde{x}_{i_j}, \theta^{j-1} + y_{i_j}\tilde{x}_{i_j} \rangle$$

$$= \langle \theta^{j-1}, \theta^{j-1} \rangle + y_{i_j}^2 \langle \tilde{x}_{i_j}, \tilde{x}_{i_j} \rangle + 2\langle \theta^{j-1}, y_{i_j}\tilde{x}_{i_j} \rangle$$

$$= \|\theta^{j-1}\|^2 + \left\|\tilde{x}_{i_j}\right\|^2 + 2\langle \theta^{j-1}, y_{i_j}\tilde{x}_{i_j} \rangle$$

$$\leq \|\theta^{j-1}\|^2 + \left\|\tilde{x}_{i_j}\right\|^2$$



(d). Again, by induction, show that

$$\|\theta^j\|^2 \leq j(1+R^2)$$

where $R = max_i \|x_i\|$, and $\|\|$ is just the Euclidean norm.

Sol.

If $j = 1$, *from* (c),

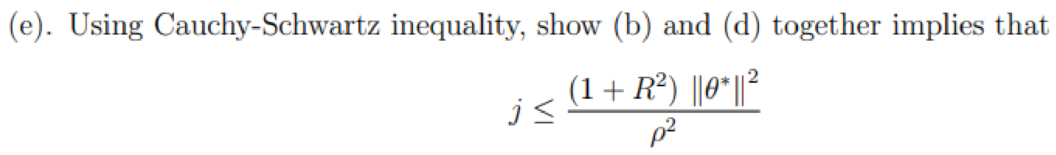
$$\|\theta^1\|^2 \leq \|\theta^0\|^2 + \left\|\tilde{x}_{i_1}\right\|^2 = 0 + R^2 \leq 1(1+R^2)$$

If $j = j+1$, *assume that* $\left\|\theta^j\right\|^2 \leq j(1+R^2)$,

$$\left\|\theta^{j+1}\right\|^2 \leq \left\|\theta^j\right\|^2 + \left\|\tilde{x}_{i_{j+1}}\right\|^2 \leq j(1+R^2) + R^2$$

$$\leq j(1+R^2) + 1 + R^2 \leq (j+1)(1+R^2)$$

Therefore we proved $\left\|\theta^j\right\|^2 \leq j(1+R^2)$.



(e). Using Cauchy-Schwartz inequality, show (b) and (d) together implies that

$$j \leq \frac{(1+R^2)\ \|\theta^*\|^2}{\rho^2}$$

Sol.

Cauchy Schwartz inequality:

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \| \mathbf{u} \| \| \mathbf{v} \|$$

From (b) *and* (d),

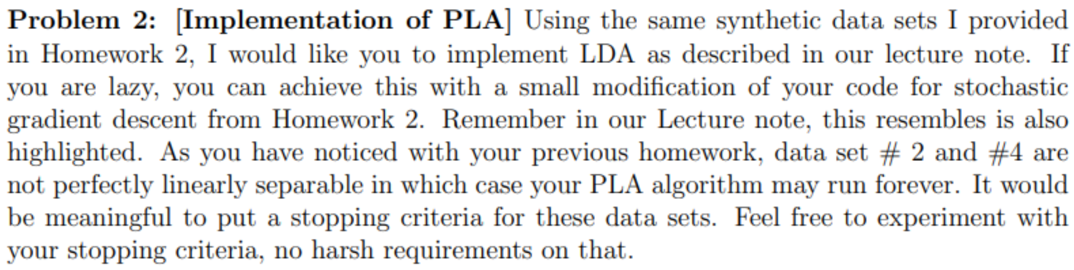
$$j\rho \leq \langle \theta^j, \theta^* \rangle$$
$$= \langle \theta^j, \theta^* \rangle \leq \left\|\theta^j\right\| \|\theta^*\|$$

$$j\rho \leq \sqrt{j(1+R^2)}\,\|\theta^*\|$$

Square both sides of the equation,

$$j^2\rho^2 \leq j(1+R^2)\|\theta^*\|^2$$

$$j \leq \frac{(1+R^2)\|\theta^*\|^2}{\rho^2}$$



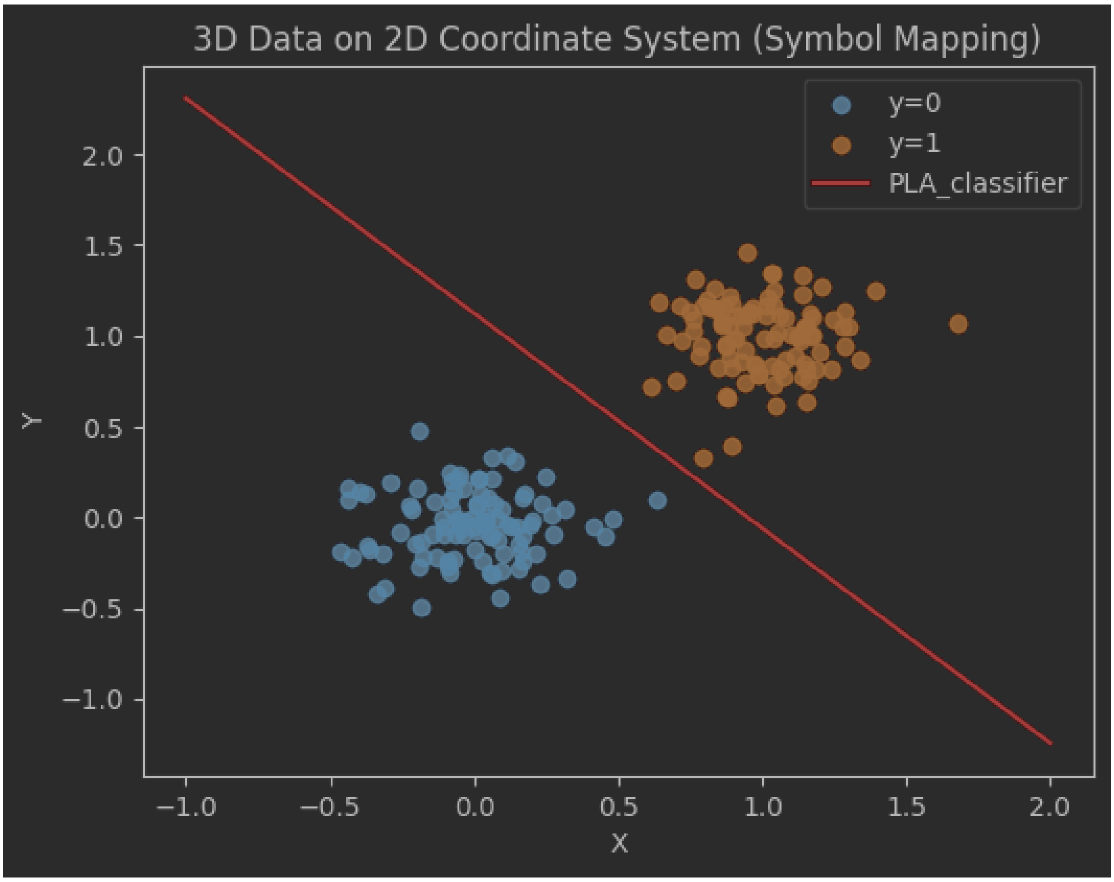
Problem 2: **[Implementation of PLA]** Using the same synthetic data sets I provided in Homework 2, I would like you to implement LDA as described in our lecture note. If you are lazy, you can achieve this with a small modification of your code for stochastic gradient descent from Homework 2. Remember in our Lecture note, this resembles is also highlighted. As you have noticed with your previous homework, data set # 2 and #4 are not perfectly linearly separable in which case your PLA algorithm may run forever. It would be meaningful to put a stopping criteria for these data sets. Feel free to experiment with your stopping criteria, no harsh requirements on that.

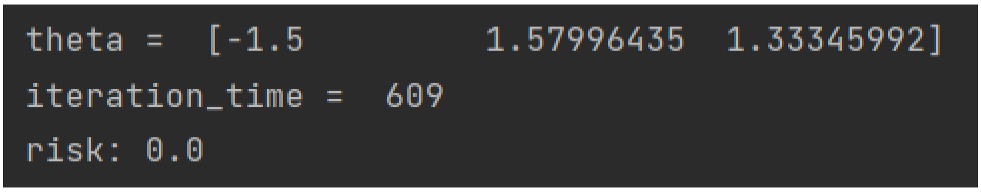
Sol.

The code is in HW3_2.ipynb.

For the stopping criterion, I set a variable called "continuousCount", which represents the number of times the classifier has successfully classified the data continuously so far. When "continuousCount" reaches 200, it means that the classifier has successfully classified all data, and we will stop training at this time.

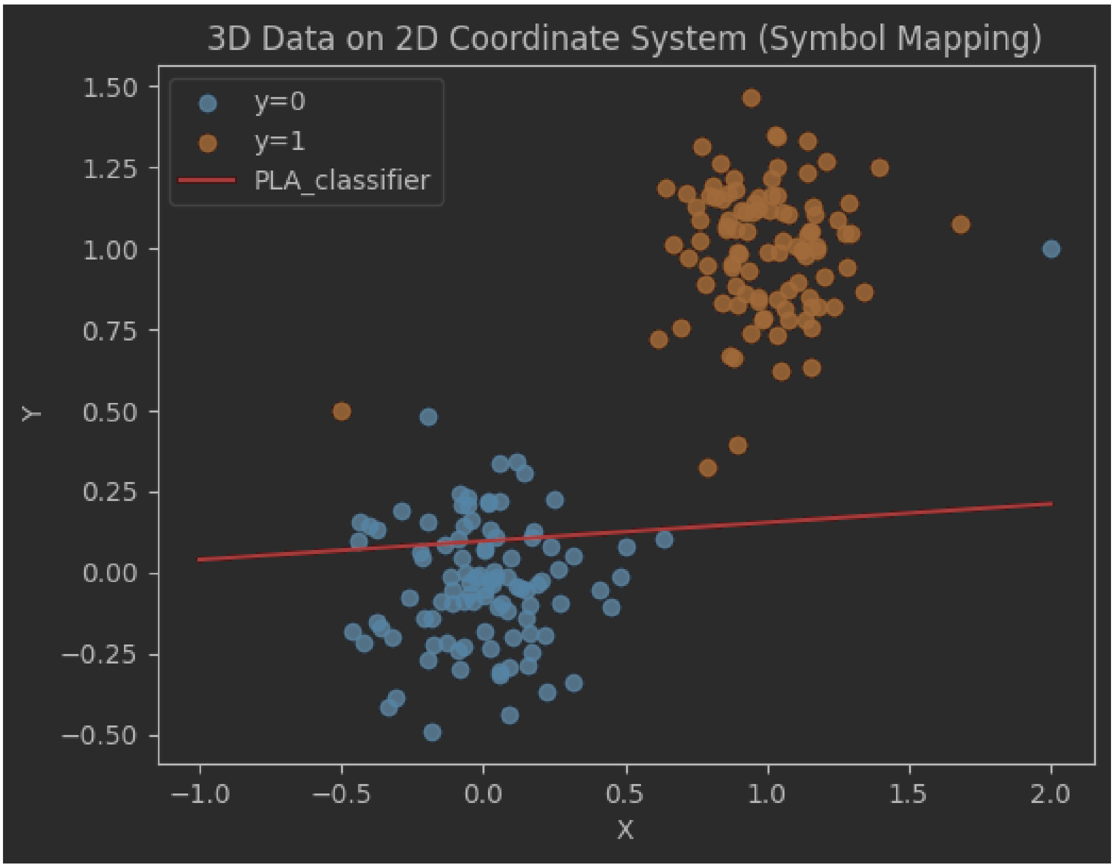
Synthetic1

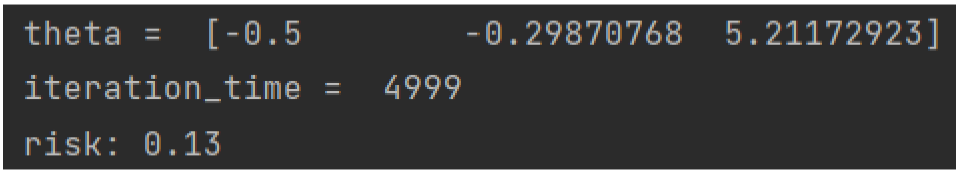




```
theta =  [-1.5         1.57996435  1.33345992]
iteration_time =  609
risk: 0.0
```

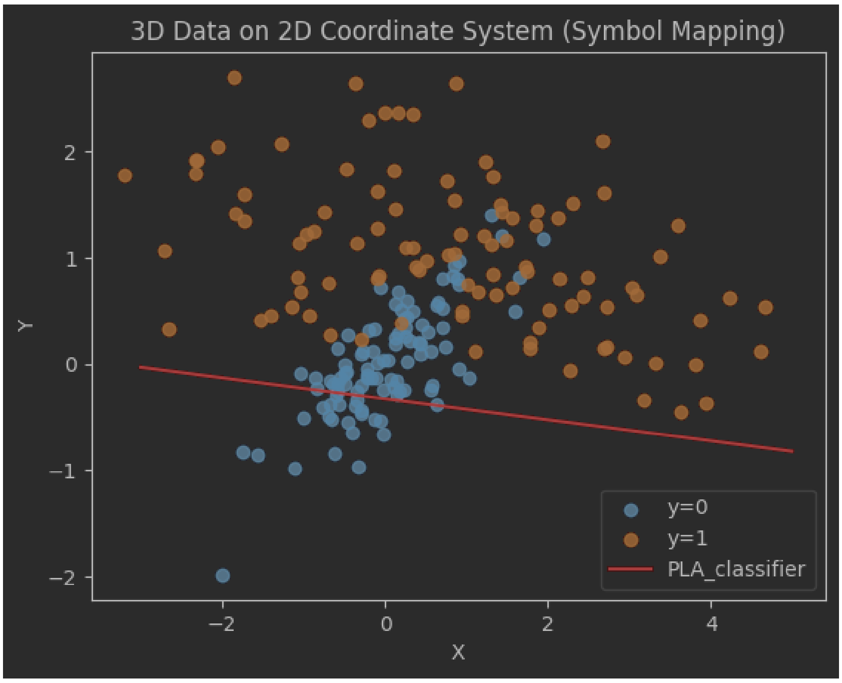
Synthetic2

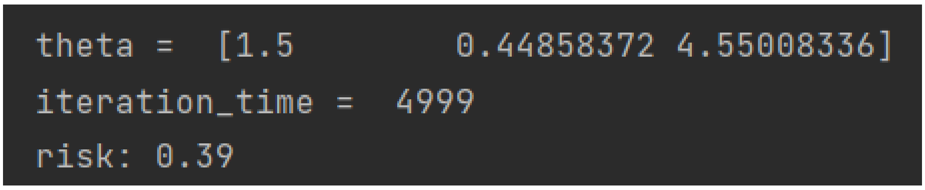




```
theta =  [-0.5         -0.29870768  5.21172923]
iteration_time =  4999
risk: 0.13
```

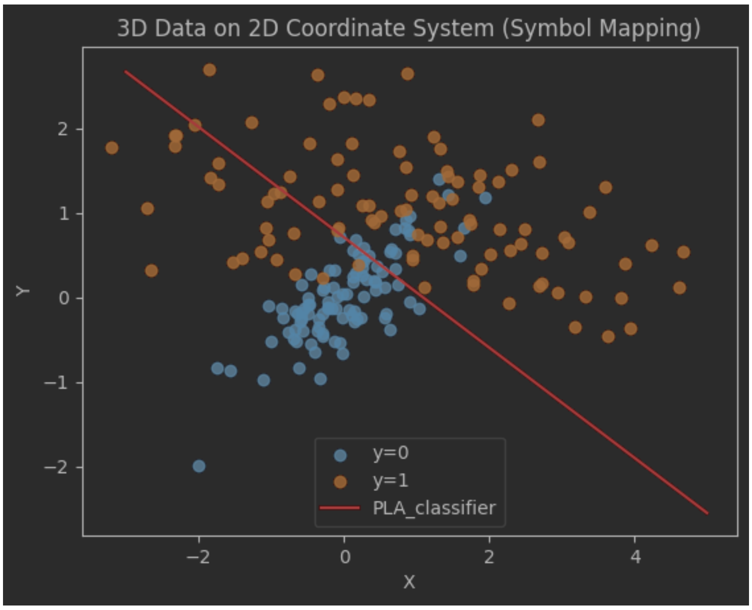
Synthetic3

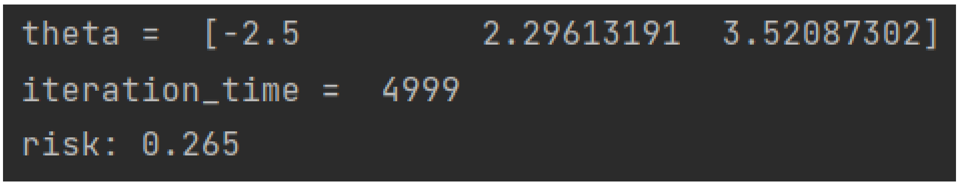




```
theta =  [1.5        0.44858372 4.55008336]
iteration_time =  4999
risk: 0.39
```

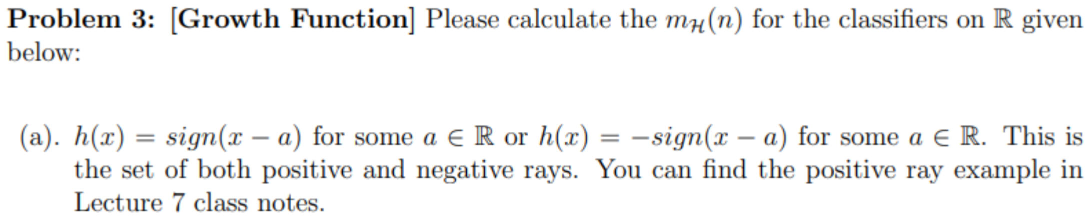
Synthetic4





```
theta =  [-2.5        2.29613191  3.52087302]
iteration_time =  4999
risk: 0.265
```

Only Synthetic1 successfully completed the classification, indicating that PLA is only suitable for linearly separable data.



Problem 3: [Growth Function] Please calculate the $m_{\mathcal{H}}(n)$ for the classifiers on $\mathbb{R}$ given below:

(a). $h(x) = sign(x - a)$ for some $a \in \mathbb{R}$ or $h(x) = -sign(x - a)$ for some $a \in \mathbb{R}$. This is the set of both positive and negative rays. You can find the positive ray example in Lecture 7 class notes.

Sol.

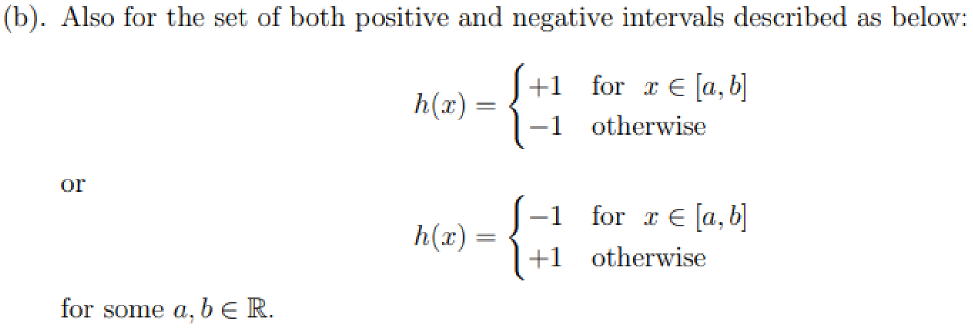
For positive ray with n points, there will be $n + 1$ *intervals for classifying.*

For negative ray with n points, there will be $n + 1$ *intervals for classifying.*

We repeated the calculation twice for the leftmost interval and the rightmost interval.

then

$$m_H(n) = 2(n + 1) - 2 = 2n$$



(b). Also for the set of both positive and negative intervals described as below:

$$h(x) = \begin{cases} +1 & \text{for} \quad x \in [a, b] \\ -1 & \text{otherwise} \end{cases}$$

or

$$h(x) = \begin{cases} -1 & \text{for} \quad x \in [a, b] \\ +1 & \text{otherwise} \end{cases}$$

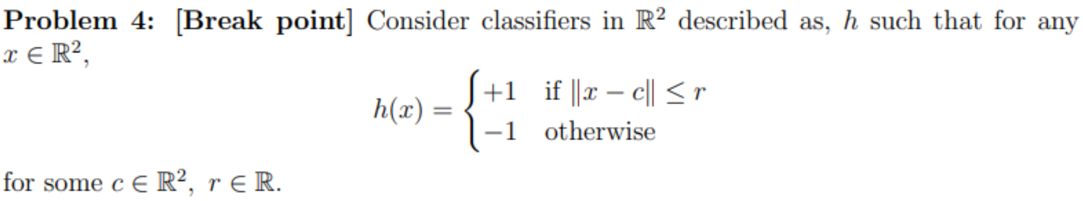
for some $a, b \in \mathbb{R}$.

For positive intervals with n points, there will be $n + 1$ *intervals for classifying.*

We select 2 *intervals for the* $[a, b]$*, and we can also put* $[a, b]$ *in one interval. Therefore*

$$m_H(n) = \binom{n+1}{2} + 1$$

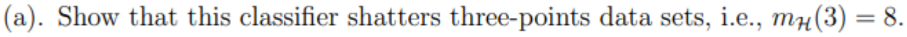
$$= \frac{n^2}{2} + \frac{n}{2} + 1$$



Problem 4: [Break point] Consider classifiers in $\mathbb{R}^2$ described as, h such that for any $x \in \mathbb{R}^2$,

$$h(x) = \begin{cases} +1 & \text{if } \|x - c\| \leq r \\ -1 & \text{otherwise} \end{cases}$$

for some $c \in \mathbb{R}^2$, $r \in \mathbb{R}$.



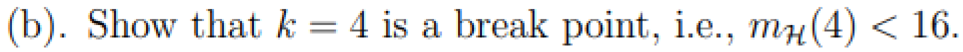
(a). Show that this classifier shatters three-points data sets, i.e., $m_{\mathcal{H}}(3) = 8$.

Sol.

From the question, the classifier is a circle where the distance between x and center c should be less than radius r. If the point is in the circle, we denote $+1$*, otherwise denote* -1.

To calculate the growth function of 3 *points, we are free to choose each point in or not in classifier, which means,*

$$m_H(3) = 2^3 = 8$$

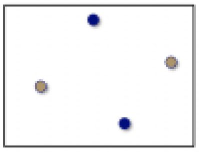


(b). Show that $k = 4$ is a break point, i.e., $m_{\mathcal{H}}(4) < 16$.

Sol.

For $k = 4$*, in the same way, the Dichotomies*$(4) = 2^4 = 16$.

However, there are 2 *exceptions*:



We can't classify this case by using convex sets, and the another case is change blue point to yellow, and change yellow pont to blue, which works exactly in the same way.

Therefore,

$$m_H(4) = 2^4 - 2 = 14$$

Since $m_H(4) <$ *Dichotomies*(4)*, we conclude that* $k = 4$ *is the break point.*