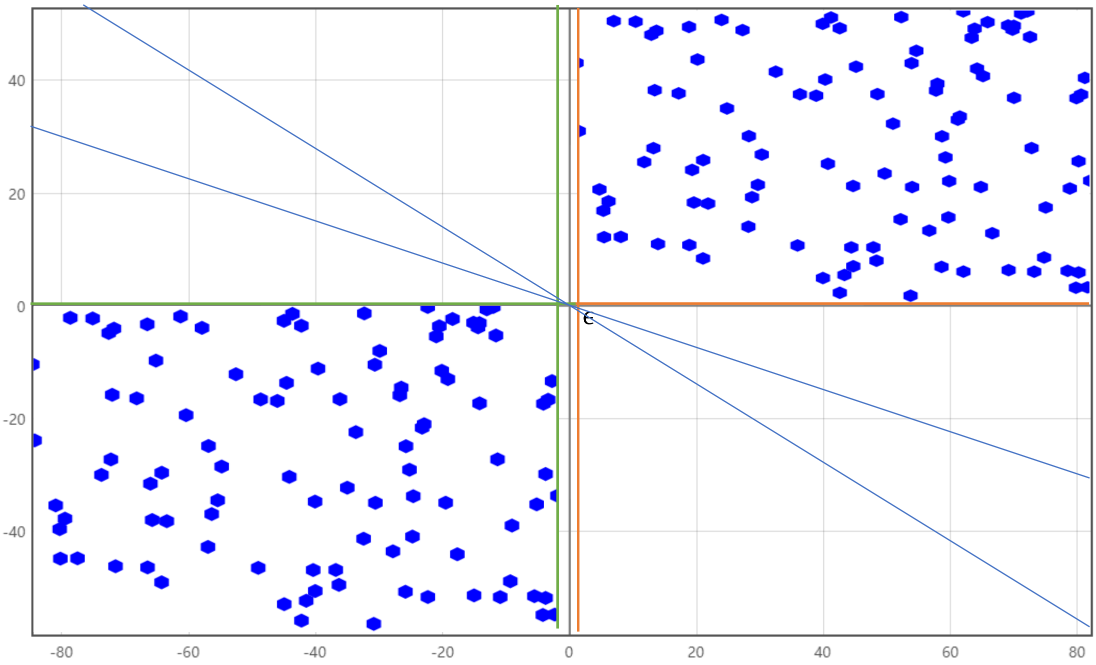
Decision Tree Problems Solutions:

1)

From the question we can know Y is independent of $X_i, i = 1, \dots, k-1$, and $Y = \begin{cases} +1, & X_k > 0 \\ -1, & X_k < 0 \end{cases}$, so we have a relationship of X_i and Y that X_i and Y have same sign because $D \geq 0$.
So we can have a simple graph describing this relationship.



So we know that if

$$(x_1, x_2, \dots, x_k) \mapsto (1, x_1, x_2, \dots, x_k).$$

$$b, (w_1, \dots, w_k) \mapsto (b, w_1, w_2, \dots, w_k).$$

We have $b = w_1 = \cdots = w_{k-1} = 0$, we will definitely get the above graph and using the orange and green line we can perfect separate the data. So for all blue lines (planes) we have showed between the orange and green lines, the normal vectors of those blue planes can be perceptron that correctly classifies this data. For example, considering the plane $Y = -X_k$, the normal vector would be $w = (0, 0, 0, \dots, 0, 1)$ and it is easy to prove that this weight vector is a normalize perceptron.
So we can conclude that there is a perceptron that correctly classifies this data.
And this perceptron is not unique because we can have another example $w = (0, 0, 0, \dots, 0, 2)$. Though these two vectors are parallel vector, and another $w = (1, 0, 0, \dots, 0, 1)$ but in my opinion, they are different perceptron.
I think there are many parallel vectors so those vectors are all 'best' perceptron, if we must give a 'best', I think the best is the normalized $w = (0, 0, 0, \dots, 0, 1)$.

2)

The theoretical answer in the question 1 is $b = 0, \underline{w} = (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1)$ and we put b into w then get a simplified weight vector:

$$w = (t,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1)$$

where $t \geq 0$ such that $t \times b \geq 0$.

But in the practice, I get a

$$\begin{aligned} w = (&-1, 0.44331145, 1.57311376, 0.67828259, -1.00218947, \quad 0.34768063, \\ &-0.84148654, -0.18797422, 0.72818798, 1.44577673, -0.03722489, \\ &1.12597341, 1.30711947, 0.02868002, -2.02766875, \quad 0.56759471, \\ &0.96812664, -2.0764666, \quad -0.7350216, 1.2197309, 10.23670977) \end{aligned}$$

That is different from the theoretical one, so I think because the m is too small so we can just predict the overall perceptron and the w we get is one of the perceptron for these 100 data. For these 100 data, either of the two perceptron can classify the data. But in general I think the theoretical one is better because it will ignore the noise in the weight vector.

3)

For this question, I use 500 different data set to get the average result. The code is in question3.py And I choose epsilon from the following set.

$$\epsilon = [0, 0.02, 0.04, 0.06, 0.08, 0.1, 0.12, \ldots \ldots, 0.9, 0.92, 0.94, 0.96, 0.98, 1]$$

The result is:

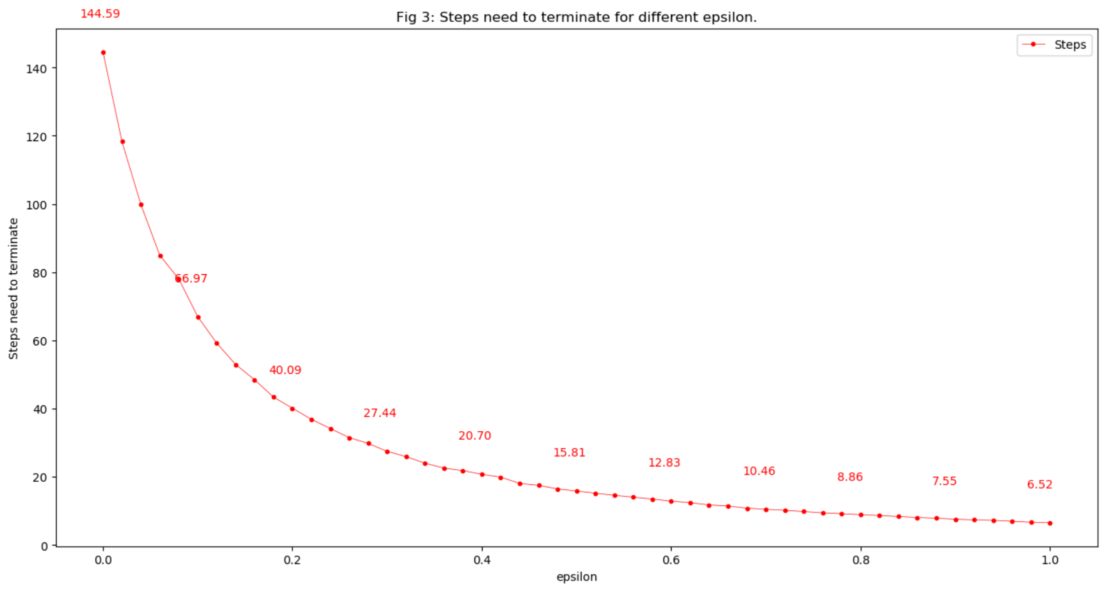
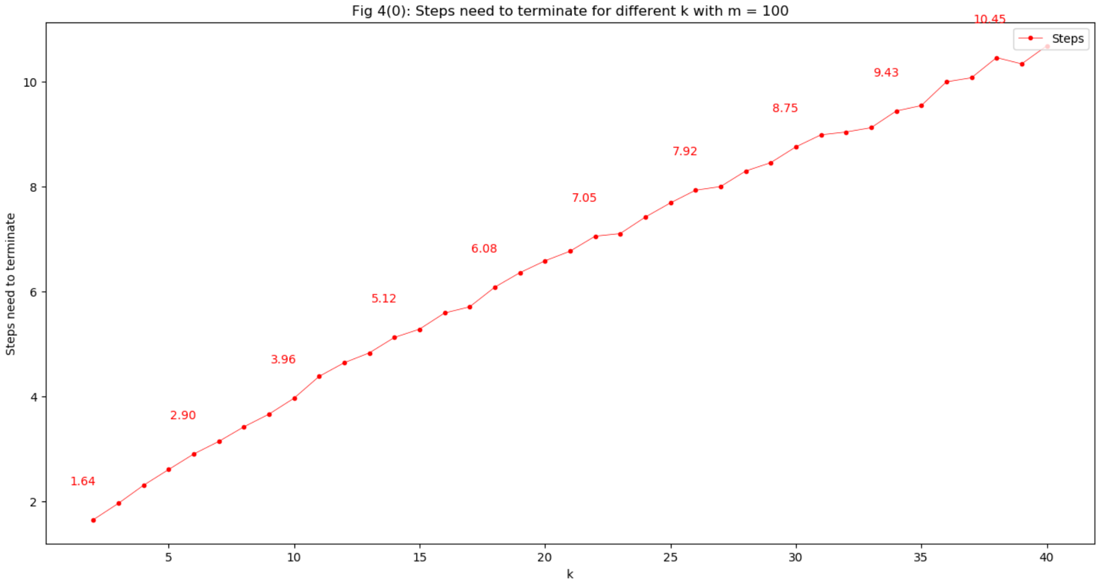


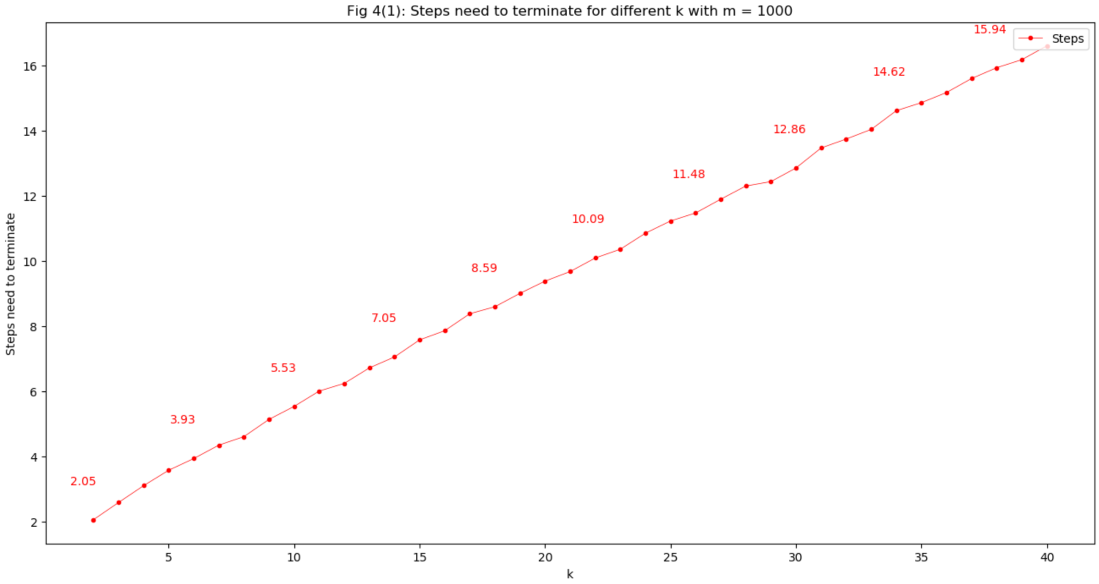
Fig 3: Steps need to terminate for different epsilon.

So we can see that when ϵ is getting bigger, the steps we need to terminate will be smaller. I think it is because the gap between two classes is getting bigger when ϵ becomes larger. We can connect with the graph in the question1 and draw this conclusion.

4)

Ideally, in the description of this question, the number of steps should be independent of k.
But my results show that the number of steps will increase when k increases. I think maybe it is because the complexity of data is increasing with k, so in a small m we can’t get the result of independence. It suppose to show independence when m is large enough.
Below is the result with 500 iterations.





5)

From the formula

$$Y = \begin{cases} +1 & \text{if } \sum_{i=1}^{k} X_i^2 \geq k \\ -1 & \text{else.} \end{cases}$$

We can get that:

$$Y = sign\left(\sum\nolimits_{i=1}^{k} X_i^2 - k\right) = sign\left(\sum\nolimits_{i=1}^{k} (X_i^2 - 1)\right)$$

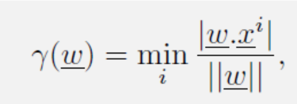
And we already know that if Y is linear separable that Y and $\underline{X}$ have linear relation:

$$Y = sign\left(\left(\sum\nolimits_{i=1}^{k} w_i \cdot X_i\right) + b\right) = sign\left(\sum\nolimits_{i=1}^{k} w_i \cdot X_i + b/k\right)$$

From the above two formula, we can easily tell that if $Y = sign\left(\sum_{i=1}^{k}(X_i^2 - 1)\right)$, Y and $\underline{X}$ is definitely not having linear relation, and I think Y and $\underline{X}$ would have a ***Quadratic relationship.***

The code is in question5.py

About how could we find a flag of terminator:
Explanation:
Because:



$$\gamma(\underline{w}) = \min_i \frac{|\underline{w}.\underline{x}^i|}{||\underline{w}||},$$

So we can get a hypothesis that ***if m is big enough, we will finally get two*** $\underline{x}$***:*** $\underline{x_1} \neq \underline{x_2}$ ***such that they have the same projection to the separate plane which means they have the same minimum*** $\boldsymbol{\gamma(\underline{w})}$***.***
And the above hypothesis is easy to proof. So we have a method based on that hypothesis:

If we have $\underline{x_1} \neq \underline{x_2}$ *such that* $\gamma_{x_1}(\underline{w}) = \gamma_{x_2}(\underline{w})$*, we can calculate the* $\gamma_{x_1}(\underline{w})$ *by using the property of symmetry. So, the* $\gamma_{x_1}(\underline{w}) = \gamma_{x_2}(\underline{w}) = \frac{1}{2}\left(\min_i \frac{|\underline{w*x_1}|}{\|\underline{w}\|} + \min_i \frac{|\underline{w*x_2}|}{\|\underline{w}\|}\right)$*, which is the distance between* $\underline{x_1}$ *and* $\underline{x_2}$ *times the normalized* $\underline{w}$*. So we can just find this value and this value would be less than or equal to* $\frac{1}{T}$*.*

So, T can be as max as 1/2(*distance between* $\underline{x_1}$ *and* $\underline{x_2}$ *times the normalized* $\underline{w}$), as I think.