50.007 – Machine Learning

May–August Term, 2022

Homework 1 (21 May 2022)

Question 1. (Subtotal: 25 points)
(not a coding question)

Consider the following data set for a binary classification learning:

Instance	x_1^i	x_2^i	x_3^i	x_4^i	y^i
x^1	0	0	1	2	1
x^2	0	0	3	0	1
x^3	4	1	1	1	-1
x^4	1	2	0	1	-1
x^5	1	0	0	2	-1

1.1 Please try to
1) find a linear classifier with offset using the perceptron update rule and (12 points)
2) conclude whether this data set is linearly separable or not (3 points).

Further instructions for 1.1:
Your update process should start with $\theta^0 = 0$ and the initial offset also equals to 0, and terminate when (the training error is 0) or (over 5 updates of θ values are completed) whichever condition is satisfied first. During the updating process, please:
1) use the rule: $sign(\theta, x) = 1$ while $\theta x \geq 0$; $sign(\theta, x) = -1$ while $\theta x < 0$.
2) iterate from instance 1 to 5 sequentially (after finishing instance 5, start with instance 1 again) until you are confident about whether this data set is linearly separable or not.

1.2 Is it possible to tell whether this data set is linearly separable without doing any algorithmic updates with it? If you think it is possible, show your reasons. If you think it is not possible, please also explain why. (10 points)

Further instructions for 1.2:
Only showing one example that satisfies the "linearly separable" condition without explaining why you choose this example (or how you come up with your example) will make you lose partial credits.

Answers:
1.1
Perceptron updating process:
- Update 0: $\theta^0 = (0, 0, 0, 0), \theta_0^0 = 0$

$\theta x^1 = (0, 0, 0, 0) \cdot (0, 0, 1, 2) = 0$, Thus, $sign(\theta^0 x^1) = 1$ and our $y^1 = 1$, so no need to update.
$\theta x^2 = (0, 0, 0, 0) \cdot (0, 0, 3, 0) = 0$, Thus, $sign(\theta^0 x^2) = 1$ and our $y^2 = 1$, so no need to update.
$\theta x^3 = (0, 0, 0, 0) \cdot (4, 1, 1, 1) = 0$, Thus, $sign(\theta^0 x^3) = 1$, but our $y^3 = -1$, so we need to update: $\theta^1 = \theta^0 + (-1) \cdot (4, 1, 1, 1) = (-4, -1, -1, -1)$, and $\theta_0^1 = \theta_0^0 + (-1) = -1$

- Update 1: $\theta^1 = (-4,-1,-1,-1), \theta_0^1 = -1$

$\theta^1 x^4 = (-4,-1,-1,-1)\cdot(1,2,0,1) - 1 = -8$, Thus, $sign(\theta^1 x^4) = -1$ and our $y^4 = -1$, so no need to update.
$\theta^1 x^5 = (-4,-1,-1,-1)\cdot(1,0,0,2) - 1 = -7$, Thus, $sign(\theta^1 x^5) = -1$ and our $y^5 = -1$, so no need to update.
$\theta^1 x^1 = (-4,-1,-1,-1)\cdot(0,0,1,2) - 1 = -4$, Thus, $sign(\theta^1 x^1) = -1$ but our $y^1 = 1$, so we need to update: $\theta^2 = \theta^1 + (1)\cdot(0,0,1,2) = (-4,-1,0,1)$, and $\theta_0^2 = \theta_0^1 + (1) = 0$

- Update 2: $\theta^2 = (-4,-1,0,1), \theta_0^2 = 0$

$\theta^2 x^2 = (-4,-1,0,1)\cdot(0,0,3,0) = 0$, Thus, $sign(\theta^2 x^2) = 1$ and our $y^2 = 1$, so no need to update.
$\theta^2 x^3 = (-4,-1,0,1)\cdot(4,1,1,1) = -16$, Thus, $sign(\theta^2 x^3) = -1$ and our $y^3 = -1$, so no need to update.
$\theta^2 x^4 = (-4,-1,0,1)\cdot(1,2,0,1) = -5$, Thus, $sign(\theta^2 x^4) = -1$ and our $y^4 = -1$, so no need to update.
$\theta^2 x^5 = (-4,-1,0,1)\cdot(1,0,0,2) = -2$, Thus, $sign(\theta^2 x^5) = -1$ and our $y^5 = -1$, so no need to update.
$\theta^2 x^1 = (-4,-1,0,1)\cdot(0,0,1,2) = 2$, Thus, $sign(\theta^2 x^1) = 1$ and our $y^1 = 1$, so no need to update.

- Our Update 2 has correctly classified all the data points in this data set, so we have met our termination criterion.
- Thus, we have found $\theta^2 = (-4,-1,0,1), \theta_0^2 = 0$ as our linear classifier that correctly classify all the data points. Thus, our training data set is linearly separable.

1.2

Method 1:

- Yes, it is possible.
- With the offset taking into consideration, we could add a constant 1 as one more dimension for our input data set:

Instance	x_0^i	x_1^i	x_2^i	x_3^i	x_4^i	y^i
x^1	1	0	0	1	2	1
x^2	1	0	0	3	0	1
x^3	1	4	1	1	1	-1
x^4	1	1	2	0	1	-1
x^5	1	1	0	0	2	-1

- Our question now becomes: given the feature vectors x^1 to x^5, could we find vector θ so that $\theta x = z$ ($z = (z_1, z_2, z_3, z_4, z_5)$) where $z_1 \geq 0$, $z_2 \geq 0$, $z_3 < 0$, $z_4 < 0$, and $z_5 < 0$.
- Meanwhile, we can see that our x^i vectors are all linearly independent, which can be proved by solving $\theta x = 0$ (θ are the variables) where the only possible solutions are $\theta = 0$.
- According to linear algebra basics, this means our x vectors (x^1 to x^5) spans R^5, which more specifically means for any five dimensional vector whose elements are all real numbers, it can be expressed as a linear combination of x^1, x^2, x^3, x^4,and x^5.
- Thus, we could find a vector θ so that $\theta x = z$ where $z_1 \geq 0$, $z_2 \geq 0$, $z_3 < 0$, $z_4 < 0$, and $z_5 < 0$. For instance, while $z = (1,2,-1,-1,-2)$, our $\theta = (\frac{11}{8}, -\frac{3}{8}, -\frac{1}{2}, \frac{5}{8}, -1)$

Method 2:

• Yes, it is possible.

• With the offset taking into consideration, we could add a constant 1 as one more dimension for our input data set:

Instance	x_0^i	x_1^i	x_2^i	x_3^i	x_4^i	y^i
x^1	1	0	0	1	2	1
x^2	1	0	0	3	0	1
x^3	1	4	1	1	1	-1
x^4	1	1	2	0	1	-1
x^5	1	1	0	0	2	-1

• Our question now becomes: given the feature vectors x^1 to x^5, could we find vector θ so that $\theta x = z$ ($z = (z_1, z_2, z_3, z_4, z_5)$) where $z_1 \geq 0$, $z_2 \geq 0$, $z_3 < 0$, $z_4 < 0$, and $z_5 < 0$.

• If $\theta x = z$, then $\theta x x^{-1} = z x^{-1}$, we have $\theta = z x^{-1}$ requiring x is invertible.

• We know that if $|x| \neq 0$, x will be invertible. So we calculate $|x|$, we have $|x| = -14$. Thus, we could find a vector θ so that $\theta x = z$ where $z_1 \geq 0$, $z_2 \geq 0$, $z_3 < 0$, $z_4 < 0$, and $z_5 < 0$.

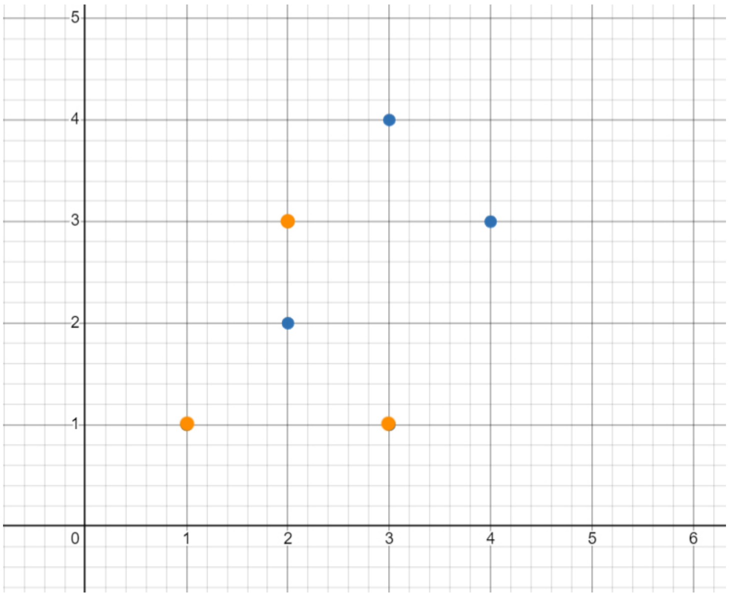
Question 2. (Subtotal: 25 points)
(not a coding question) Consider the following two-dimensional data set for a binary classification learning:

Instance	x_1^i	x_2^i	y^i
x^1	1	1	-1
x^2	3	1	-1
x^3	2	3	-1
x^4	2	2	1
x^5	3	4	1
x^6	4	3	1

2.1 Please draw the 2D graph of this data set (remember to differentiate the data points with different labels). Please find a linear classifier that can correctly classify all the instances in the data set, or explain why no such classifier exists. (7 points)

2.2 Please use the stochastic gradient descent algorithm with hinge-loss function to find a two-dimensional linear classifier for this data set. You should start with $\theta^0 = 0$ with the offset also equals to 0, and terminate when you have updated θ 3 times. Please use the input instance sequentially from instance 1 to instance 6 (after finishing instance 6, starting from instance 1 again) during your update process.Please use $\eta^k = \frac{1}{k+1}$ in your update process. (10 points)
Is θ^3 the best parameter among θ^1, θ^2, and θ^3? If it is, why? If it is not, which one is the best one? Please justify your answers.(8 points)

Answers:
2.1
- Below is the 2D graph: yellow (label -1) and blue (label 1) dots belong to two different labels.



• This data set is not linearly separable because we cannot find a straight line that makes all the blue dots stay on one side, and all the yellow dots stay on the other.

2.2
Below are the updates based on gradient descent with hinge-loss functions:

• Update 0: $\theta^0 = (0, 0), \theta_0^0 = 0$
$y^1\theta^0x^1 = -1 \cdot (0, 0) \cdot (1, 1) = 0 \leq 1$, Thus, we should update: $\theta^1 = \theta^0 + \eta^0y^1x^1 = (0, 0)+1\times(-1)\times(1, 1) = (0, 0)+(-1, -1) = (-1, -1)$,and $\theta_0^1 = \theta_0^0+\eta^0y^1 = 0+(-1) = -1$

• Update 1: $\theta^1 = (-1, -1), \theta_0^1 = -1$
$y^2(\theta^1x^2 + \theta_0^1) = -1(\cdot(-1, -1) \cdot (3, 1) - 1) = 5 > 1$, so no need to update.
$y^3(\theta^1x^3 + \theta_0^1) = -1(\cdot(-1, -1) \cdot (2, 3) - 1) = 6 > 1$, so no need to update.
$y^4(\theta^1x^4 + \theta_0^1) = 1(\cdot(-1, -1) \cdot (2, 2) - 1) = -5 < 1$, Thus, we should update: $\theta^2 = \theta^1+\eta^1y^4x^4 = (-1, -1)+0.5\times(1)\times(2, 2) = (-1, -1)+(1, 1) = (0, 0)$,and $\theta_0^2 = \theta_0^1+\eta^1y^4 = -1 + 0.5 \times 1 = -0.5$

• Update 2: $\theta^2 = (0, 0), \theta_0^2 = -0.5$
$y^5(\theta^2x^5 + \theta_0^2) = 1 \cdot [(0, 0) \cdot (3, 4) - 0.5] = -0.5 \leq 1$, Thus, we should update: $\theta^3 = \theta^2 + \eta^2y^5x^5 = (0, 0) + \frac{1}{3} \times (1) \times (3, 4) = (0, 0) + (1, \frac{4}{3}) = (1, \frac{4}{3})$, and $\theta_0^3 = \theta_0^2 + \eta^2y^5 = -0.5 + \frac{1}{3} \times 1 = -\frac{1}{6}$

• Update 3: $\theta^3 = (1, \frac{4}{3}), \theta_0^3 = -\frac{1}{6}$
We have reached the terminate condition of updating θ 3 times.

Calculating $R_n(\theta)$ for θ^1, θ^2, and θ^3:
• $R_n(\theta^1) = \frac{1}{6}\sum_{i=1}^{6} max[(1 - y^i[(-1, -1) \cdot (x_1^i, x_2^i) - 1]), 0] = \frac{1}{6}(0 + 0 + 0 + 6 + 9 + 9) = 4$
• $R_n(\theta^2) = \frac{1}{6}\sum_{i=1}^{6} max[(1-y^i[(0, 0)\cdot(x_1^i, x_2^i)-0.5]), 0] = \frac{1}{6}(0.5+0.5+0.5+1.5+1.5+1.5) = 1$
• $R_n(\theta^3) = \frac{1}{6}\sum_{i=1}^{6} max[(1-y^i[(1, \frac{4}{3})\cdot(x_1^i, x_2^i)-\frac{1}{6}]), 0] = \frac{1}{6}(3.17+5.17+6.83+0+0+0) = 2.5$

• θ^3 is not the best solution we have.
• We can see that $R_n(\theta^3)$ is bigger than $R_n(\theta^2)$, and $R_n(\theta^2)$ is the smallest. So θ^2 is the best solution we have.

Question 3. (Subtotal: $25 + 2$ points)
(coding question: perceptron)
This exercise requires the student to understand the basics of linear classification using perceptron update rule. To start any machine learning task, it requires data exploration and data cleaning. We will be using a real-life dataset from NBA which includes details about rookie basketball players such as games played, points per game, rebounds, assists and blocks to predict if the player will be in NBA after 5 years. Please open the .ipynb file to see details of this question. We have 2 bonus points at the end of this question.

Answers:
Refer to HW1_notebook_answers.ipynb

Question 4. (Subtotal: 25 points)
(coding question: hinge loss)
This exercise requires the student to understand the basics of linear classification using stochastic gradient with hinge loss. We will be comparing our functions results with sklearn implementation of hinge loss. We will be using a real-life dataset from NBA which includes details about rookie basketball players such as games played, points per game, rebounds, assists and blocks to predict if the player will be in NBA after 5 years. Please open the .ipynb file to see details of this question.

Answers:
Refer to HW1_notebook_answers.ipynb