ECE M146
Introduction to Machine Learning
Instructor: Lara Dolecek

Spring 2020 Midterm Assessment
Monday, May 4th 11:00 AM, 2019
Maximum score: 100 points

You have until **May 5th 11:00 AM (Pacific Time)** to submit your work **directly on Gradescope**.
Please read and carefully follow all the instructions.

Instructions

- You may type your exam or scan your handwritten version. Please show your work and make sure all the work is discernible.
- Make sure to include your **full name** and **UID** in your submitted file.
- For questions related to the exam, you may deal into the following Zoom Q&A sessions:
  - May 4th, 1:00pm - 1:30pm.
  - May 4th, 3:00pm - 3:30pm.
  - May 4th, 6:00pm - 6:30pm.
  - May 4th, 9:00pm - 9:30pm.
  - May 5th, 7:30am - 8:00am.
  - May 5th, 9:00am - 9:30am.

  Links to these Zoom Meetings are available under Week 6 on CCLE. **Only clarification questions** will be answered. Please do not ask for hints. We will also have a forum under Week 6 that reiterates all answered questions. Make sure to check the forum before dial in.
- **Important:** Throughout this exam, you will find a parameter α in some of the questions. All α refers to the same parameter. This parameter α is dependent on your UID, specifically, $\alpha = (\text{Last digit of UID} \mod 8) + 1$. For example, a person with UID: 123456789 will use $\alpha = 2$ throughout this test. Please clearly indicate what is your α on the first page of your answers. You **will lose points** if the correct α is not used.
- **Academic Integrity**
  During this exam, you are **allowed** to use all course material posted online, including lectures, discussion, and homeworks, and your own textbooks. You are **disallowed** to contact with a fellow student or with anyone outside the class who can offer a solution e.g., web forum.
  Please write the following statement on the first page of your answer sheet. You will **lose 20 points** if we can not find this statement. The policy on academic dishonesty can be found at the same place with this exam.

  I __*YourName*__ with UID ____ have read and understood the policy on academic dishonesty available on the course website.

1. (20 pts) **Perceptron** (Recall: $\alpha =$ (Last digit of UID mod 8)+1)

   (a) (4 pts) Write down the perceptron learning rule by filling in the blank below with a proper sign (+ or -). Note that η is a small constant learning rate factor.

   i. Input $\boldsymbol{x}$ is falsely classified as positive:

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^{t} \underline{\quad - \quad} \eta \boldsymbol{x}$$

   ii. Input $\boldsymbol{x}$ is falsely classified as negative:

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^{t} \underline{\quad + \quad} \eta \boldsymbol{x}$$

   (b) (16 pts) Consider a perceptron algorithm to learn a 3-dimensional weight vector $\boldsymbol{w} = [w_0, w_1, w_2]^T$ with w_0 the bias term. Suppose we have training set as following:

Sample #	1	2	3	4
$\boldsymbol{x}$	$[\alpha, \alpha]$	$[-\alpha, -2\alpha]$	[-8,-16]	[3,1]
y	+1	+1	-1	-1

Show the weights at each step of the perceptron learning algorithm. Loop through the training set once (i.e. MaxIter = 1) with the same order presented in the above table. Start the algorithm with initial weight $\boldsymbol{w} = [w_0, w_1, w_2]^T = [0, 1, 1]^T$. And we assume the learning rate $\eta = 1$.(Update when $y\boldsymbol{w}^T\boldsymbol{x} \leq 0$)
Solution:
Starting weights: $\boldsymbol{w} = [0, 1, 1]$.
Update weights based on $[\alpha, \alpha]^T$: no update.
Update weights based on $[-\alpha, -2\alpha]^T$: $\boldsymbol{w} \leftarrow \boldsymbol{w} + [1, -\alpha, -2\alpha] = [1, 1-\alpha, 1-2\alpha]$.
Update weights based on $[-8, -16]^T$: $\boldsymbol{w} \leftarrow \boldsymbol{w} - [1, -8, -16] = [0, 9-\alpha, 17-2\alpha]$.
Update weights based on $[3, 1]^T$: $\boldsymbol{w} \leftarrow \boldsymbol{w} - [1, 3, 1] = [-1, 6-\alpha, 16-2\alpha]$.

2. (20 points) **K-NN classifier**

This is a programming question. **Please attach a printout of your code at the end of your answer. You will loss points if you don't attach you code.** You will be asked to build a k-NN classifier from first principles. You may **not** use **fitcknn** (**sklearn.neighbors.KNeighborsClassifier** for python) in this problem as you will get incorrect answer by using those built-in functions.

The data is provided in *Q2data.csv*. The first two columns contain the two-dimensional features for each data point and the last column contains the label (0 or 1) for each data point. There are 80 data points in *Q2data.csv* and you need to separate it into the training data and testing data based on α. The rule is as follows: use the $(10(\alpha-1)+1)$-th to (10α)-th rows from *Q2data.csv* as the testing data and the rest as the training data. For example, a person with $\alpha = 1$ will use the first 10 rows as the testing data.

The k-NN classifier classifies a data point with feature x_{test} based on a training set by performing the following procedures:

- Compute the distance from x_{test} to the feature of all training points. We will use the L1 distance in this problem. The definition of L1 distance between two vectors $x, y \in \mathbf{R}^N$ is $L1(x, y) = \sum_{i=1}^{N} |x_i - y_i|$.
- Find the k nearest neighbors of this point.
- Classify this points as the majority class of its k nearest neighbors.

We use the following two rules to handle ties:

(a) Let d_k be the distance from x_{test} to the k-th nearest neighbor of x_{test}. If there are multiple training points that have distance d_k from x_{test}. Choose those points with the smallest indexes to be included in the k nearest neighbors. For example, let $k = 3$, if there is x_9 that is distance 1 away from x_{test}; x_1, x_3 and x_4 that is distance 2 away from x_{test}, then the 3 nearest neighbor of x_{test} are x_1, x_3 and x_9. Note that $d_k = 2$ in this example.

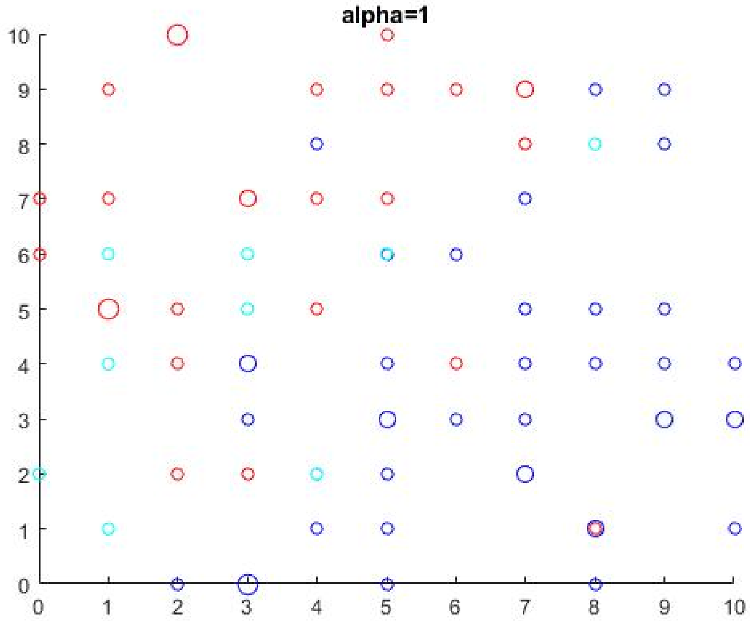
(b) For even k, among all k nearest neighbors of a data point, if the number of points from class 0 is the same as the number of points from class 1, classify this data point as class 0 deterministically.

Plot the training data with red points denoting those data points with label 1 and blue points denoting those data points with label 0. In the same plot, also plot the testing data with color cyan. Is the data linearly separable? Find and plot (in another figure) the testing accuracy for $k = 1, 2, \cdots, 9$.

Solution: The testing accuracy for each k and α are summarized in the following table:

k / α	1	2	3	4	5	6	7	8	9
1	0.9	0.7	0.8	0.7	0.9	0.8	0.7	0.8	0.8
2	0.9	0.5	0.9	0.6	0.8	0.6	0.6	0.6	0.6
3	1.0	1.0	1.0	1.0	0.9	0.9	0.9	0.9	0.9
4	0.7	0.8	0.7	0.8	0.7	0.7	0.7	0.8	0.8
5	0.9	0.8	0.9	0.8	0.8	0.8	0.8	0.8	0.8
6	0.7	0.9	0.8	0.8	0.8	0.8	0.8	0.8	0.8
7	0.8	0.8	0.7	0.8	0.7	0.7	0.7	0.7	0.7
8	0.9	0.9	0.9	0.9	0.9	0.9	0.8	0.8	0.8

The first plot for $\alpha = 1$ is shown below. Note that the radius of the circle represent the number of data at this point. All other plots are omitted.



3. (20 pts) **Decision Tree**

There are 8 students who have taken the course *Introduction to Machine Learning* in the previous quarter. At the end of the quarter, we did a survey trying to learn how their background affects their performance in this class. Each student reports whether he/she did well (binary feature 1) or not well (binary feature 0) in ECE146(*Introduction to Machine Learning*) and four other classes: ECE102(*System and Signals*), ECE131A(*Probability and Statistics*), MATH61(*Introduction to Discrete Structures*) and MUSC15(*Art of Listening*). The results are summarized in the following table:

Student #	ECE102	ECE131	MATH61	MUSC15	ECE146
1	1	1	1	1	1
2	0	1	1	0	1
3	1	1	0	0	1
4	0	1	0	1	1
5	1	0	0	1	0
6	0	0	0	0	0
7	1	0	1	1	1
8	0	0	0	1	0

(a) (1 pt) What is the binary entropy of this data set, i.e., $H(ECE146)$?

Solution: Define the binary entropy function as follows:

$$H_b(p) = -p\log(p) - (1-p)\log(1-p).$$

$$H(ECE146) = H_b(\frac{5}{8}) = -(\frac{5}{8}\log(\frac{5}{8}) + \frac{3}{8}\log(\frac{3}{8})) \approx 0.9544$$

(b) (4 pts) Calculate the conditional entropy of

$$H(ECE146|X), \text{for } X \in \{ECE102, ECE131, MATH61, MUSC15\},$$

i.e., the conditional entropy of ECE146 conditioning on the features.

Solution:

$$H(ECE146|ECE102) = \frac{1}{2}H_b(\frac{3}{4}) + \frac{1}{2}H_b(\frac{1}{2}) \approx 0.9056.$$

$$H(ECE146|ECE131) = \frac{1}{2}H_b(1) + \frac{1}{2}H_b(\frac{3}{4}) \approx 0.4056.$$

$$H(ECE146|MATH61) = \frac{3}{8}H_b(1) + \frac{5}{8}H_b(\frac{3}{5}) \approx 0.6068.$$

$$H(ECE146|MUSC15) = \frac{5}{8}H_b(\frac{3}{5}) + \frac{3}{8}H_b(\frac{1}{3}) \approx 0.9512.$$

(c) (4 pts) Calculate the information gain:

$$I(ECE146; X) = H(ECE146) - H(ECE146|X),$$

for

$$X \in \{ECE102, ECE131, MATH61, MUSC15\}.$$

Solution:

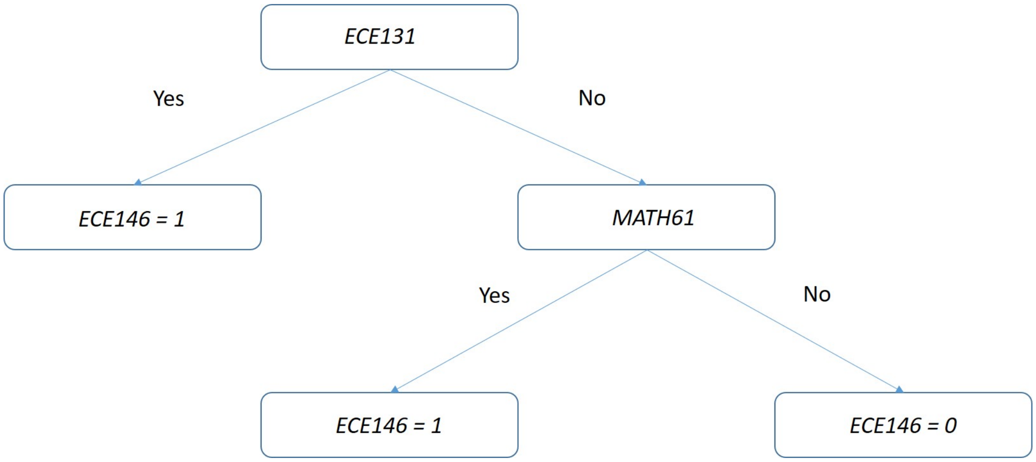
$$\begin{aligned} I(ECE146|ECE102) &= 0.9544 - 0.9056 = 0.0488; \\ I(ECE146|ECE131) &= 0.9544 - 0.4056 = 0.5488; \\ I(ECE146|MATH61) &= 0.9544 - 0.6068 = 0.3476; \\ I(ECE146|MUSC15) &= 0.9544 - 0.9512 = 0.0032. \end{aligned}$$

(d) (1 pt) Based on the information gain, determine the first feature to split on.

Solution: We choose $ECE131$ which has the largest information gain.

(e) (8 pts) Make the full decision tree. Make sure to show all your work. After each split, treat the sets of samples with $X = 0$ and $X = 1$ as two separate sets and redo (b), (c) and (d) on each of them. X is the feature for previous split and is thus excluded from the available features which can be split on next. Terminate splitting if after the previous split, the entropy of ECE146 in the current set is 0.

Solution: Below show the decision tree if we choose $ECE131$ as the first splitting feature.



After the first split, $H(ECE146|ECE146 = 1) = 0$ so the tree stops growing on that branch. We are left with the samples that have $ECE131 = 0$ which is summarized in the following table.

Sample #	ECE102	MATH61	MUSC15	ECE146
5	1	0	1	0
6	0	0	0	0
7	1	1	1	1
8	0	0	1	0

By observation, Feature $MATH61$ has the highest information gain. Then the next split should be $MATH61$. After this split, every leaf is pure, i.e., $ECE146$ is either 0 or 1. Therefore, we stop growing the tree.

(f) (2 pts) Now, determine if students 9 and 10 are good at $ECE146$ or not based on the decision tree you made.

Student #	ECE102	ECE131	MATH61	MUSC15	ECE146
9	1	0	1	0	?
10	1	0	0	0	?

Solution:

Student 9: Good

Student 10: Not good

4. (20 points)**Linear Regression** (Recall: α = (Last digit of UID mod 8)+1)
Please show intermediate steps for this question, the problem is designed to be done by hand calculation.
You are given the following three data points:

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 6 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}, \begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} \alpha+1 \\ 0 \end{bmatrix}.$$

You want to fit a line, i.e., $\hat{y} = w_1 x + w_0$, that minimize the following sum of square error:

$$J(\boldsymbol{w}) = \sum_{i=1}^{3} (w_1 x_i + w_0 - y_i)^2.$$

In matrix-vector form, the objective function is

$$J(\boldsymbol{w}) = \|\boldsymbol{X}\boldsymbol{w} - y\|^2,$$

for some $\boldsymbol{X}$, y and $\boldsymbol{w} = [w_0, w_1]^T$. What are $\boldsymbol{X}$ and y (3 pts)? What is the optimal $\boldsymbol{w}$ that minimize the objective function (13 pts)? Draw the three data points and the fitted line (4 pts).
Solution:

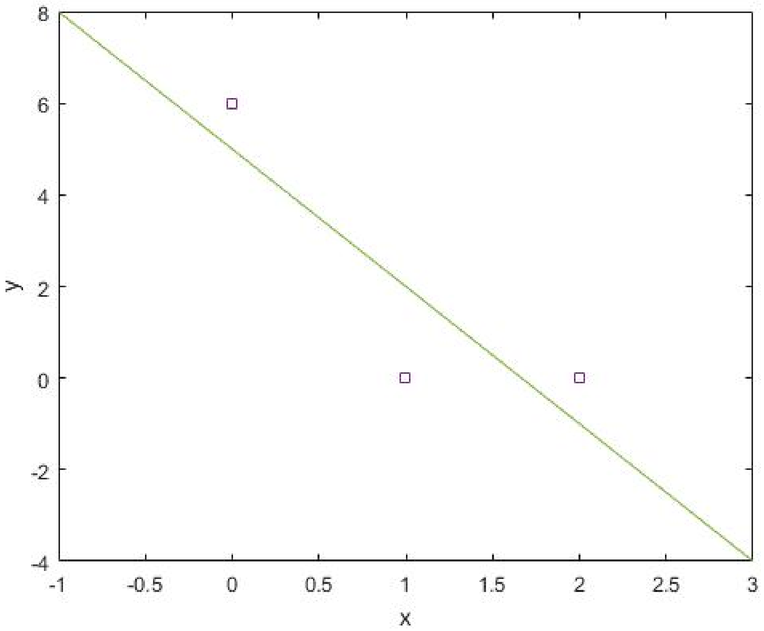
$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & \alpha \\ 1 & \alpha+1 \end{bmatrix}, y = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix}.$$

$$\begin{aligned} \boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y &= \begin{bmatrix} 3 & 2\alpha+1 \\ 2\alpha+1 & 2\alpha^2+2\alpha+1 \end{bmatrix}^{-1} \times \begin{bmatrix} 6 \\ 0 \end{bmatrix} \\ &= \frac{1}{2\alpha^2+2\alpha+2} \begin{bmatrix} 2\alpha^2+2\alpha+1 & -2\alpha-1 \\ -2\alpha-1 & 3 \end{bmatrix} \times \begin{bmatrix} 6 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{6\alpha^2+6\alpha+3}{\alpha^2+\alpha+1} \\ \frac{-6\alpha-3}{\alpha^2+\alpha+1} \end{bmatrix}. \end{aligned}$$

Numerical result is shown in the table below.

α	1	2	3	4	5	6	7	8
w_0	5	5.57	5.77	5.86	5.9	5.93	5.95	5.96
w_1	-3	-2.14	-1.62	-1.29	-1.06	-0.91	-0.79	-0.70

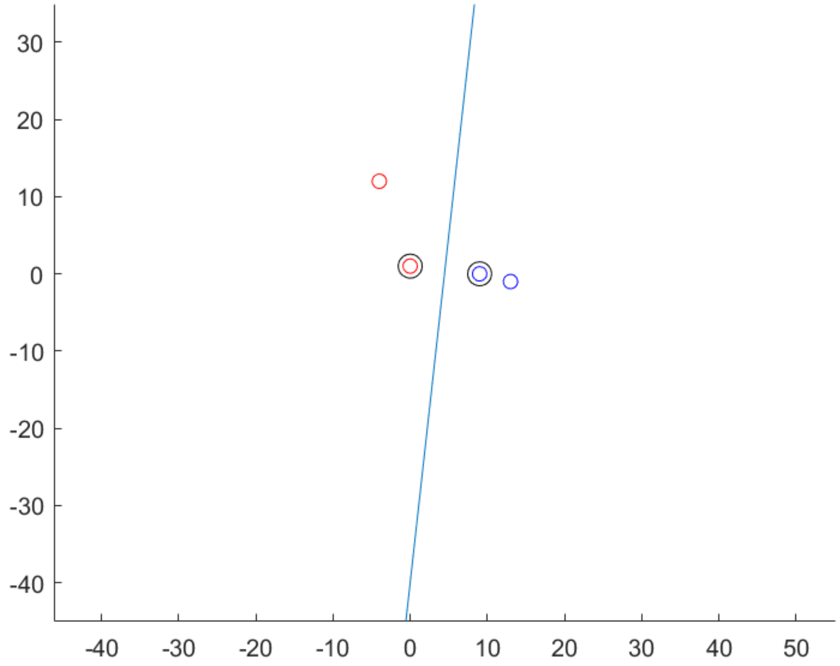
The plot for $\alpha = 1$ is shown below. Other plots are tilted version of this and are omitted.


8
6
4
2
0
-2
-4
y
-1
-0.5
0
0.5
1
1.5
2
2.5
3
x

5. (20 pts) **Support Vector Machine** (Recall: $\alpha =$ (Last digit of UID mod 8)+1)
You are given the following data set which is comprised of $\boldsymbol{x}^{(i)} \in \mathbb{R}^2$ and $y^{(i)} \in \{-1, 1\}$.

i	$x_1^{(i)}$	$x_2^{(i)}$	y_i
1	-4	12	1
2	0	α	1
3	10-α	0	-1
4	13	-1	-1

(a) (4 pts) Plot the data. Is the data linearly separable?
Solution: Yes, data is linearly separable. Plot for $\alpha = 1$ is shown below. Plots for other α's are tilted version of this and are omitted.



(b) (5 pts) Suppose you are asked to find the maximum margin separating hyperplane of the form $[w_1, w_2][x_1, x_2]^T + b = 0$. Write down the (primal) optimization problem **explicitly** using only w_1, w_2 and b.
Solution:
The optimization problem is as follows:

$$\begin{aligned} \min_{w_1, w_2, b} \quad & w_1^2 + w_2^2 \\ s.t. \quad & -4w_1 + 12w_2 + b \geq 1, \\ & \alpha w_2 + b \geq 1, \\ & -(10-\alpha)w_1 - b \geq 1, \\ & -13w_1 + w_2 - b \geq 1. \end{aligned}$$

(c) (6 pts) Look at the data and circle the support vectors by inspection. Find and plot the maximum margin separating hyperplane.

Solution:

The two support vectors are $[0, \alpha]^T$ and $[10-\alpha, 0]^T$. The line that has normal vector $[\alpha - 10, \alpha]$ and also pass through the midpoint of support vectors ($\left[\frac{10-\alpha}{2}, \frac{\alpha}{2}\right]^T$) is $(\alpha - 10)x_1 + \alpha x_2 - 10\alpha + 50 = 0$.

(d) (5 pts) Solve the dual problem for the Lagrange multipliers α_is and use your dual solution to find the $\boldsymbol{w}$ and b of the primal problem.

Solution:

Since we only have two support vectors, only the Lagrange multiplier corresponding to the support vectors are non-zero. Let α_2 denote the Lagrange multiplier for $x^{(2)}$ and similarly α_3 for $x^{(3)}$. From the condition $\sum_{i=1}^{4} \alpha_i y_i = 0$, we get $\alpha_2 = \alpha_3 = \alpha_0$. Write down the objective of the dual problem of SVM

$$\begin{aligned} W(\boldsymbol{\alpha}) &= \sum_{i=1}^{4} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{4} y_i y_j a_i a_j \boldsymbol{x}^{(i)T} \boldsymbol{x}^{(j)} \\ &= 2\alpha_0 - \frac{1}{2}\alpha_0^2 \boldsymbol{x}^{(2)T}\boldsymbol{x}^{(2)} + \alpha_0^2 \boldsymbol{x}^{(2)T}\boldsymbol{x}^{(3)} - \frac{1}{2}\alpha_0^2 \boldsymbol{x}^{(3)T}\boldsymbol{x}^{(3)} \\ &= 2\alpha_0 - \frac{\alpha_0^2}{2}(2\alpha^2 - 20\alpha + 100). \end{aligned}$$

Maximizing $W(\boldsymbol{\alpha})$ over α_0, we get $\alpha_3 = \alpha_2 = \alpha_0 = \frac{1}{\alpha^2 - 10\alpha + 50}$. Using $\boldsymbol{w} = \sum_{m \in \mathcal{S}} \alpha_m y^{(m)} \boldsymbol{x}^{(m)}$, we get $\boldsymbol{w} = \frac{1}{\alpha^2 - 10\alpha + 50}[\alpha - 10, \alpha]^T$. To find b, recall that

$$y^{(i)} \left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b \right) = 1$$

for any support vectors $x^{(i)}$. Use any support vector, we can get $b = \frac{50 - 10\alpha}{\alpha^2 - 10\alpha + 50}$. The $\boldsymbol{w}$ and b we find by solving the dual problem is a scaled version of $[w_1, w_2]^T$ and w_0 in part (c). These solutions therefore give the same separating hyperplane.