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

ECE M146 Midterm Spring 2020
Melody Chen
#705120273

I Melody Chen with UID 705120273 have read and understood the policy on academic dishonesty available on the course website.

7051202 73 $3\%8 = 3$

$3 + 1 = 4$

1. (20 pts) **Perceptron** (Recall: α = (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:

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

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

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

   (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 as the bias term. Suppose we have training set as following:

$\alpha = 4$

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 as presented in the above table. Start the algorithm with initial weight $\boldsymbol{w} = [w_0, w_1, w_2]^T = [0, 1, 1]^T$. Assume the learning rate $\eta = 1$. (Update when $y\boldsymbol{w}^T\boldsymbol{x} \leq 0$.)

$k=1$ $w_1 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$

Sample #1: $w_1^T x_1 = [0\ 1\ 1] \cdot \begin{bmatrix} 1 \\ 4 \\ 4 \end{bmatrix} = 8$ sign(8) = +1, $y_1 = +1$ ✓ $w_1 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$

No update.

Sample #2: $w_1^T x_2 = [0\ 1\ 1] \cdot \begin{bmatrix} 1 \\ -4 \\ -8 \end{bmatrix}$ $k=2$

$= -4 - 8 = -12$ sign(-12) = -1, $y_2 = +1$ ✗

Sample #2 is misclassified.

$w_2 = w_1 + y_2 x_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} + (1)\begin{bmatrix} 1 \\ -4 \\ -8 \end{bmatrix} = \begin{bmatrix} 1 \\ -3 \\ -7 \end{bmatrix}$ $w_2 = \begin{bmatrix} 1 \\ -3 \\ -7 \end{bmatrix}$

Sample #3: $w_2^T x_3 = [1\ -3\ -7] \cdot \begin{bmatrix} 1 \\ -8 \\ -16 \end{bmatrix} = 137$ $k=3$ $\text{sign}(137) = +1 \neq y_3$

Sample #3 is misclassified.

$w_3 = w_2 + y_3 x_3 = \begin{bmatrix} 1 \\ -3 \\ -7 \end{bmatrix} - \begin{bmatrix} 1 \\ -8 \\ -16 \end{bmatrix} = \begin{bmatrix} 0 \\ 5 \\ 9 \end{bmatrix}$ $w_3 = \begin{bmatrix} 0 \\ 5 \\ 9 \end{bmatrix}$

Sample #4: $w_3^T x_4 = [0\ 5\ 9] \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix} = 24$ $k=4$ $\text{sign}(24) = +1 \neq y_4$

Sample #4 is misclassified.

$w_4 = w_3 + y_4 x_4 = \begin{bmatrix} 0 \\ 5 \\ 9 \end{bmatrix} - \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \\ 8 \end{bmatrix}$ $w_4 = \begin{bmatrix} -1 \\ 2 \\ 8 \end{bmatrix}$

2. (20 points) **K-NN classifier** $\alpha = 4$ 31 - 40th

This is a programming question. **Please attach a (scanned) printout of your code at the end of your answer. You will lose 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 may 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 test data point x_{test} based on a training set by performing the following procedure:

- Compute the distance from x_{test} to each of the training points. We will use the L_1 distance in this problem. The definition of L_1 distance between two vectors $x, y \in \mathbf{R}^N$ is $L_1(x, y) = \sum_{i=1}^{N} |x_i - y_i|$.
- Find the k nearest neighbors of test data point x_{test}.
- Declare the label of test data point x_{test} as being the majority class of its k nearest neighbors.

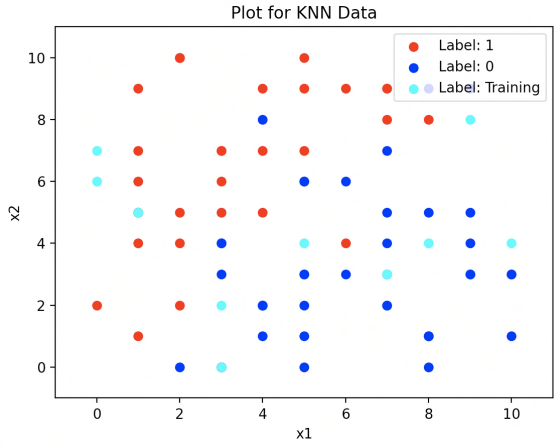
We use the following two rules to handle ties:

- 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 the same distance d_k from x_{test}, choose those points with the smallest indices to be included in the k nearest neighbors. As an example, consider $k = 3$. Suppose that the point x_9 is at distance 1 from x_{test}, the points x_1, x_3, and x_4 are at distance 2 from x_{test}, and all other training points are strictly more than distance 2 away from x_{test}. Then, the 3 nearest neighbors of x_{test} are x_1, x_3 and x_9.
- 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.

(a) (2 pts) 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?

(b) (18 pts) Find and plot (in another figure) the testing accuracy for $k = 1, 2, \cdots, 9$.

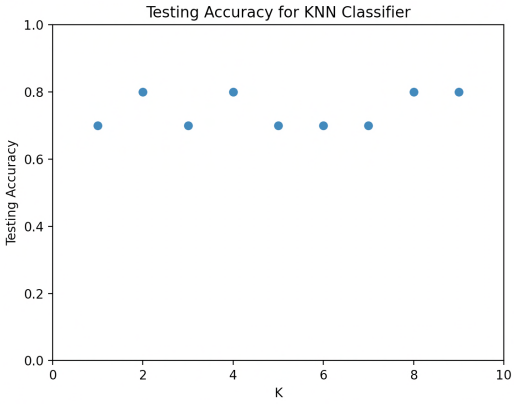
a)



$\alpha = 4$ for my ID #.
Testing Data is 31 to 40 row.

Data is not linearly separable.

b)



Testing Accuracy:

```
K = 1 Accuracy = 0.7
K = 2 Accuracy = 0.8
K = 3 Accuracy = 0.7
K = 4 Accuracy = 0.8
K = 5 Accuracy = 0.7
K = 6 Accuracy = 0.7
K = 7 Accuracy = 0.7
K = 8 Accuracy = 0.8
K = 9 Accuracy = 0.8
```

3. (20 pts) **Decision Tree**
There are 8 students who have taken the course ECE146 *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(*Systems 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., $H(ECE146)$?

(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.

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

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

for each

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

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

(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.

(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	?

* Formula used below to calculate entropy derived in class: $H(Z) = -p\log_2 p - (1-p)\log_2(1-p)$ occasionally omitted writing formula by plugging directly into calculator, but calculation for H is based on above formula.

a) $p(ECE146) = \frac{5}{8}$

$H(ECE146) = -\frac{5}{8}\log_2\left(\frac{5}{8}\right) - \left(1-\frac{5}{8}\right)\log_2\left(1-\frac{5}{8}\right)$

$= 0.9544$

b) 102: $H(ECE146 \mid ECE102) = H(ECE146 \mid ECE102=T)\,p(ECE102=T) + H(ECE146 \mid ECE102=F)\,p(ECE102=F)$

$p(ECE146=T \mid ECE102=T) = \frac{3}{4}$

$H(ECE146 \mid ECE102=T) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.8113$

$p(ECE102=T) = \frac{1}{2}$

$p(ECE146=T \mid ECE102=F) = \frac{1}{2}$

$H(ECE146 \mid ECE102=F) = 1$

* written as rounded but not actually rounded, used ↓ actual value from calculator

$H(ECE146 \mid ECE102) = 0.8113\left(\frac{1}{2}\right) + 1\left(\frac{1}{2}\right) = 0.9056$

131: $H(ECE146 \mid ECE131) = H(ECE146 \mid ECE131=T)\,p(ECE131=T) + H(ECE146 \mid ECE131=F)\,p(ECE131=F)$

$p(ECE146=T \mid ECE131=T) = 1$

$H(ECE146 \mid ECE131=T) = 0$

$p(ECE131=T) = \frac{1}{2}$

$p(ECE146=T \mid ECE131=F) = \frac{1}{4}$

$H(ECE146 \mid ECE131=F) = 0.8113$

* written as rounded but not actually rounded, used ↓ actual value from calculator

$H(ECE146 \mid ECE131) = 0.8113\left(\frac{1}{2}\right) = 0.4056$

Math 61: $H(ECE146 \mid Math61) = H(ECE146 \mid Math61=T)\,p(Math61=T) + H(ECE146 \mid Math61=F)\,p(Math61=F)$

$p(ECE146=T \mid Math61=T) = 1$

$H(ECE146 \mid Math61=T) = 0$

$p(Math61=T) = 3/8$

$p(ECE146=T \mid Math61=F) = \frac{2}{5}$

$H(ECE146 \mid Math61=F) = 0.9710$

* written as rounded but not actually rounded, used ↓ actual value from calculator

$H(ECE146 \mid Math61) = 0.9710\ (5/8) = 0.6068$

MUSC 15: $H(ECE146 \mid MUSC15) = H(ECE146 \mid MUSC15=T)\,p(MUSC15=T) + H(ECE146 \mid MUSC15=F)\,p(MUSC15=F)$

$p(ECE146=T \mid MUSC15=T) = 3/5$

$H(ECE146 \mid MUSC15=T) = 0.9710$

$p(MUSC15=T) = 5/8$

$p(ECE146=T \mid MUSC15=F) = 2/3$

$H(ECE146 \mid MUSC15=F) = 0.9183$

* written as rounded but not actually rounded, used actual value from calculator

$H(ECE146 \mid MUSC15) = 0.9710(5/8) + (0.9183)(3/8) = 0.9512$

c) $IG(ECE146; ECE102) = 0.9544 - 0.9056 = 0.0487$

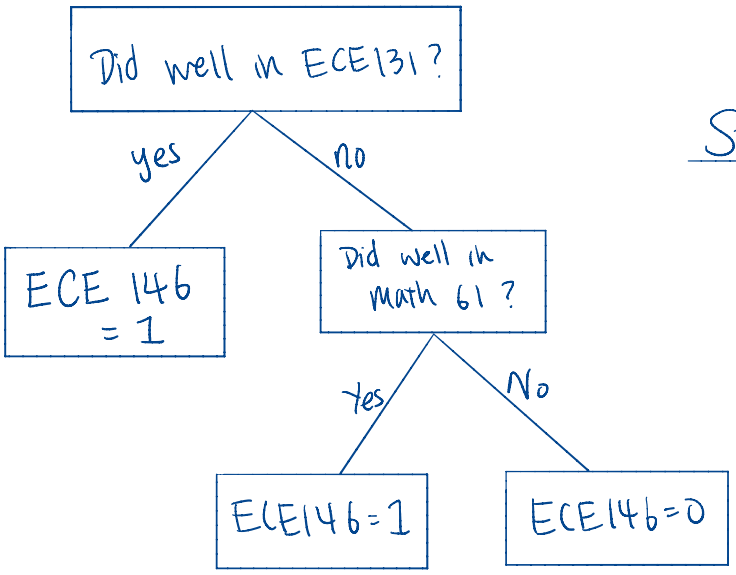
$IG(ECE146; ECE131) = 0.9544 - 0.4056 = 0.5488$

$IG(ECE146; Math61) = 0.9544 - 0.6068 = 0.3475$

$IG(ECE146; MUSC15) = 0.9544 - 0.9512 = 0.0032$

d) Split on ECE 131 first with highest IG.

e)



Sample with ECE131 = 0

Student #	ECE102	Math61	MUSC15	ECE146
5	1	0	1	0
6	0	0	0	0
7	1	1	1	1
8	0	0	1	0

$H(ECE146) = 0.8113$

$H(ECE146 \mid MUSC15=T) = H(\frac{1}{3}) = 0.9183$

$H(ECE146 \mid MUSC15=F) = H(0) = 0$

$H(ECE146 \mid MUSC15) = \frac{3}{4}(0.9183) = 0.6887$

$IG(ECE146; MUSC15) = 0.8113 - 0.6887$

$H(ECE146 \mid ECE102=T) = H(\frac{1}{2}) = 1$

$H(ECE146 \mid ECE102=F) = H(0) = 0$

$H(ECE146 \mid ECE102) = 1(\frac{1}{2}) = \frac{1}{2}$

$IG(ECE146; ECE102) = 0.8113 - \frac{1}{2}$

$H(ECE146 \mid Math61=T) = 0$

$H(ECE146 \mid Math61=F) = 0$

$H(ECE146 \mid Math61) = 0$

$IG(ECE146; Math61) = 0.8113$

↑ highest IG possible out of 3 choices

So we split next on Math 61.

f) Student #9 did well in ECE146.

Student #10 did not do well in ECE 146.

$\alpha = 4$

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:

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 6 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \end{bmatrix}, \begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} 4+1 \\ 0 \end{bmatrix}.$$

You want to fit a line, i.e., $\hat{y} = w_1 x + w_0$, that minimizes 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 $\boldsymbol{w} = [w_0, w_1]^T$.

(a) (3 pts) What are $\boldsymbol{X}$ and y?

(b) (13 pts) What is the optimal $\boldsymbol{w}$ that minimizes the objective function?

(c) (4 pts) Draw the three data points and the fitted line.

a) $X = \begin{bmatrix} 1 & 0 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}$ $y = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix}$

b) $W = (X^T X)^{-1} \cdot X^T y$ ← Derived in class

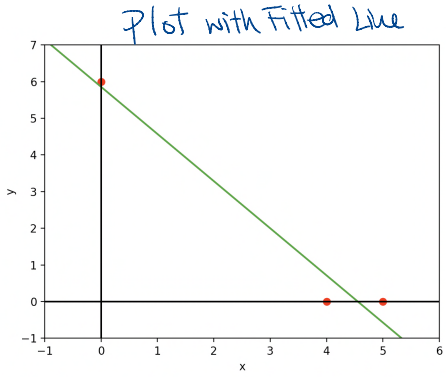
$X^T = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 4 & 5 \end{bmatrix}$ $X^T \cdot X = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 4 & 5 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} = \begin{bmatrix} 3 & 9 \\ 9 & 41 \end{bmatrix}$

$(X^T X)^{-1} = \frac{1}{42} \begin{bmatrix} 41 & -9 \\ -9 & 3 \end{bmatrix}$ $(X^T X)^{-1} \cdot X^T = \frac{1}{42} \begin{bmatrix} 41 & -9 \\ -9 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 4 & 5 \end{bmatrix} = \begin{bmatrix} \frac{41}{42} & \frac{5}{42} & -\frac{4}{42} \\ -\frac{9}{42} & \frac{3}{42} & \frac{6}{42} \end{bmatrix}$

$W = (X^T X)^{-1} X^T y = \frac{1}{42} \begin{bmatrix} 41 & 5 & -4 \\ -9 & 3 & 6 \end{bmatrix} \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix} = \frac{1}{42} \begin{bmatrix} 246 \\ -54 \end{bmatrix} = \begin{bmatrix} 5.8571 \\ -1.2857 \end{bmatrix}$

c) Fitted line: $w^T x = y$ $w_1 x_1 + w_0 x_0 = y$ $w_1 x_1 + w_0 = y$

$y = -\frac{9}{7} x_1 + \frac{41}{7}$



$\alpha = 4$

5. (20 pts) **Support Vector Machine** (Recall: α = (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	4	1
3	10 (6) α	0	-1
4	13	-1	-1

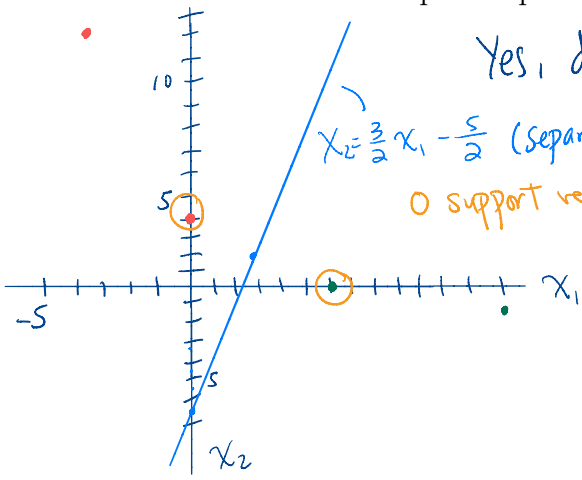
(a) (4 pts) Plot the data. Is the data linearly separable?

(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.

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

(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.

a)



Yes, data is linearly separable.

$x_2 = \frac{3}{2}x_1 - \frac{5}{2}$ (separating hyperplane)

0 support vectors

b) primal optimization:

$$\underset{w_1, w_2, b}{\arg\min} \|w\|^2 = \underset{w_1, w_2, b}{\arg\min} \left(w_1^2 + w_2^2\right)$$

subject to $(-4w_1 + 12w_2 + b) \geq 1$,

$4w_2 + b \geq 1$,

$-1(6w_1 + b) \geq 1$,

$-1(13w_1 - w_2 + b) \geq 1$

c) Support vector points: $(0,4)$, $(6,0)$

The two support vectors are $\begin{bmatrix}0\\4\end{bmatrix}$ and $\begin{bmatrix}6\\0\end{bmatrix}$.

$m = \frac{4}{-6} = -\frac{2}{3}$ $\quad m_\perp = \frac{3}{2}$

Midpoint: $(3,2)$

Separating hyperplane: $x_2 - 2 = \frac{3}{2}(x_1 - 3)$

$x_2 = \frac{3}{2}x_1 - \frac{9}{2} + 2$

equation of line → $x_2 = \frac{3}{2}x_1 - \frac{5}{2}$ or $-3x_1 + 2x_2 + 5 = 0$

Line has normal vector $[-3,2]^T$, and passes through midpoint $[3,2]^T$. So, $w = \begin{bmatrix}-3\\2\end{bmatrix}$, $b = 5$.

Line shown in plot of part A.

d) * we represent α_i here with a_i.

We use dual form of quadratic program

$$L = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} a_n t_n a_k t_k x_n^T x_k$$

want to maximize objective of dual →

$$= 2a_2 - \frac{1}{2}(a_2^2\|x_2\|^2 + a_2^2\|x_3\|^2 - 2a_2^2 x_2^T x_3)$$

$$= 2a_2 - \frac{1}{2}a_2^2\|x_2\|^2 - \frac{1}{2}a_2^2\|x_3\|^2 + a_2^2 x_2^T x_3$$

$$\frac{\partial}{\partial a_2}\left(2a_2 - \frac{1}{2}a_2^2\|x_2\|^2 - \frac{1}{2}a_2^2\|x_3\|^2 + a_2^2 x_2^T x_3\right)$$

$$= 2 - a_2\|x_2\|^2 - a_2\|x_3\|^2 + 2a_2 x_2^T x_3 = 0$$

$x_2 = \begin{bmatrix}0\\4\end{bmatrix}$ $\quad x_3 = \begin{bmatrix}6\\0\end{bmatrix}$

$$2 - a_2(4^2) - a_2(6^2) = 0$$

$$52a_2 = 2$$

$$a_3 = a_2 = \frac{1}{26}$$

Proved in class:

$a_n \geq 0$ $\quad \sum_{n=1}^{N} a_n t_n = 0$

So, $a_2 = a_3$.

We solve for w:

$$w = \sum_{n=1}^{N} a_n t_n x_n = \frac{1}{26}(1)\begin{bmatrix}0\\4\end{bmatrix} - \frac{1}{26}\begin{bmatrix}6\\0\end{bmatrix} = \frac{1}{26}\begin{bmatrix}-6\\4\end{bmatrix} = \begin{bmatrix}\frac{-3}{13}\\\frac{2}{13}\end{bmatrix}$$

We solve for b:

$t_n(w^T x_n + b) = 1$ $\quad x_n$ is point of support vector.

$1\left(\left[\frac{-3}{13}, \frac{2}{13}\right]\begin{bmatrix}0\\4\end{bmatrix} + b\right) = 1$ $\quad \frac{8}{13} + b = 1$ $\quad b = \frac{5}{13}$

w and b found using the dual form matches w (normal vector) and b found in part c. w and b found here is a scaled version, produce same hyperplane.