ECE M146
Machine Learning
Instructor: Lara Dolecek

Spring 2019 Midterm
Monday, May 6th, 2019
Maximum score: 100 points

Maximum score is 100 points. You have 110 minutes to complete the quiz. Please show your work.

Instructions

- You may find the following useful.
  - $H_b(\frac{3}{8}) = 0.95$, $H_b(\frac{1}{3}) = 0.92$, $H_b(\frac{1}{4}) = 0.81$, $H_b(\frac{1}{5}) = 0.72$,
  - The inverse of a 2×2 matrix is given by:
  $$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$
  - The quotient rule:
  $$\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{g(x)f'(x) - f(x)g'(x)}{(g(x))^2}.$$

Your Name:

Your ID Number:

Name of person on your left:

Name of person on your right:

Problem	Score	Possible
1	20	20
2	16	16
3	12	16
4	16	16
5	16	16
6	14	16
Total	94	100

20

1. (20 pts) **True or False.**

Circling the correct answer is worth +4, circling the incorrect answer is worth −2 points. Not circling either is worth 0 points.

(a) The perceptron algorithm does not converge if the training samples are not linearly separable.

TRUE FALSE

(b) k-nearest neighbors classification algorithm will always give a linear decision boundary.

TRUE FALSE

(c) The derivative of the sigmoid function $\sigma(x) = \frac{1}{1+\exp(-x)} = \frac{\exp(x)}{1+\exp(x)}$ satisfies: $\sigma'(x) = \sigma(x)(1 - \sigma(x))$.

$$\sigma' = \frac{\exp(-x)}{(1+\exp(-x))^2} = \frac{1}{1+\exp(-x)} \cdot \frac{\exp(-x)}{1+\exp(-x)}$$

$$= \sigma(1-\sigma)$$

TRUE FALSE

(d) Suppose the data is linearly separable. The hyperplane defined by $w_1^T x + b_1 = 0$ we get by solving the following optimization problem

$$\begin{aligned} \min_{w_1, b_1} \quad & \frac{1}{2}\|w_1\|^2 \\ s.t. \quad & y^{(i)}(w_1^T x^{(i)} + b_1) \geq 1, \quad i = 1, \cdots, m, \end{aligned}$$

is the same as the hyperplane defined by $w_2^T x + b_2 = 0$ we get by solving the following optimization problem

$$\begin{aligned} \min_{w_2, b_2} \quad & \frac{1}{2}\|w_2\|^2 \\ s.t. \quad & y^{(i)}(w_2^T x^{(i)} + b_2) \geq 2, \quad i = 1, \cdots, m. \end{aligned}$$

TRUE FALSE

(e) For $x_1, x_2 \in \mathbb{R}$, $K(x_1, x_2) = (1 + x_1 x_2)^2$ is a valid kernel.

TRUE FALSE

$$(1+x_1x_2)^2 = 1 + 2x_1x_2 + x_1^2x_2^2$$

$$\phi(x) = (1, \sqrt{2}x, x^2)$$

16

2. (16 pts) **Perceptron**

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

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

$w^{t+1} = w^t + y_i x_i$

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

y_i is positive

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

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

y_i is negative

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

Sample #	1	2	3	4
$\boldsymbol{x}$	[10,10]	[0,0]	[3,3]	[4,8]
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 the initial weight $\boldsymbol{w} = [w_0, w_1, w_2] = [0, 1, 1]$. We assume the learning rate $\eta = 1$. (Update when $y\boldsymbol{w}^T\boldsymbol{x} \leq 0$)

1: $y_1 w^T x_1 = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 10 \\ 10 \end{bmatrix} = 20 > 0 \longrightarrow$ correctly classified

$w^{t+1} = w^t = \boxed{\begin{bmatrix} 0 & 1 & 1 \end{bmatrix}^T}$

2: $y_2 w^T x_2 = -\begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = 0 \leq 0 \rightarrow$ incorrectly classified

$w^{t+1} = w^t + y_2 x_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \boxed{\begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}}$

3: $y_3 w^T x_3 = -\begin{bmatrix} -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 3 \end{bmatrix} = -5 \leq 0 \rightarrow$ incorrectly classified

$w^{t+1} = w^t + y_2 x_3 = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 3 \\ 3 \end{bmatrix} = \boxed{\begin{bmatrix} -2 \\ -2 \\ -2 \end{bmatrix}}$

4: $y_4 w^T x_4 = \begin{bmatrix} -2 & -2 & -2 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 8 \end{bmatrix} = -2(13) = -26 \leq 0 \rightarrow$ incorrectly classified

$w^{t+1} = w^t + y_4 x_4 = \begin{bmatrix} -2 \\ -2 \\ -2 \end{bmatrix} + \begin{bmatrix} 1 \\ 4 \\ 8 \end{bmatrix} = \boxed{\begin{bmatrix} -1 \\ 2 \\ 6 \end{bmatrix}}$

12

3. (16 pts) **k-Nearest Neighbors**

In the following questions, you will consider a k-nearest neighbor classifier using Euclidean distance metric on a binary classification task. We assign the class of the test point to be the class of the majority of the k nearest neighbors. To avoid ties, only consider odd k. Consider the following dataset:

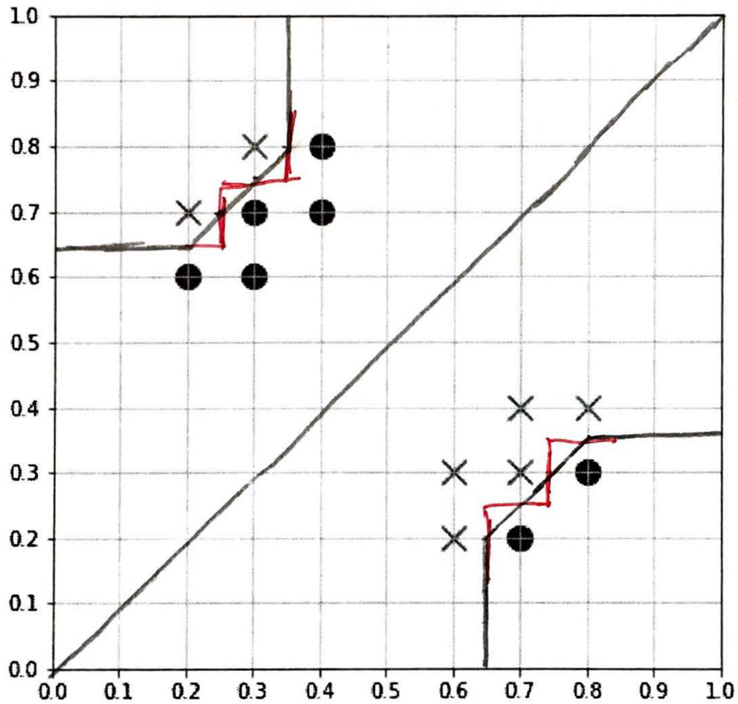


Figure 1: k-Nearest Neighbors

6 (a) In above figure, sketch the 1-nearest neighbor decision boundary for this dataset.

6 (b) What value of k **maximizes** leave-one-out cross-validation error for this dataset? What is the resulting error?

13

$\boxed{k=9 \text{ and } k=14}$ both maximize LOO CV error

both have $\boxed{\text{error}=14}$

k= 3: |||| = 4

k= 5: |||| = 4

k= 7: |||| = 4

k= 9: 14

k= 11: 10

k= 13: 14

4. (16 pts) **Linear Regression**
You are given the following three data points:

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 6 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

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

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

In matrix-vector form, the objective function is

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

for some $\boldsymbol{X}$, $\boldsymbol{y}$ and $\boldsymbol{w} = [w_0, w_1]^T$. What are $\boldsymbol{X}$ and $\boldsymbol{y}$? What is the optimal $\boldsymbol{w}$ that minimizes the objective function? Draw the three data points and the fitted line.

$$\boxed{X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}} \qquad \boxed{Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix}}$$

$$w^* = (X^T X)^{-1} X^T Y = \begin{bmatrix} 5/6 & -1/2 \\ -1/2 & 1/2 \end{bmatrix} \begin{bmatrix} 6 \\ 0 \end{bmatrix}$$

$$\boxed{w^* = \begin{bmatrix} 5 \\ -3 \end{bmatrix}}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix}$$

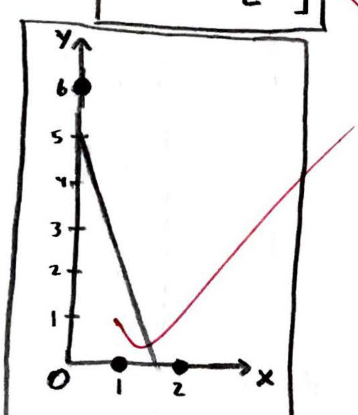
$$X^T Y = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 6 \\ 0 \end{bmatrix}$$

$$\left[\begin{array}{cc|cc} 3 & 3 & 1 & 0 \\ 3 & 5 & 0 & 1 \end{array}\right] \rightarrow \left[\begin{array}{cc|cc} 3 & 3 & 1 & 0 \\ 0 & 2 & -1 & 1 \end{array}\right]$$

$$\downarrow$$

$$\left[\begin{array}{cc|cc} 1 & 0 & 5/6 & -1/2 \\ 0 & 1 & -1/2 & 1/2 \end{array}\right] \leftarrow \left[\begin{array}{cc|cc} 1 & 1 & 1/3 & 0 \\ 0 & 1 & -1/2 & 1/2 \end{array}\right]$$

$$\frac{1}{3} + \frac{1}{2} = \frac{2+3}{6} = \frac{5}{6}$$



5. (16 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 three other classes they had taken previously: ECE102(*Systems and Signals*), ECE131A (*Probability and Statistics*) and MUSC15(*Art of Listening*). The results are summarized in the following table:

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

Calculate the information gain:

$$I(\text{ECE146}; X) = H(\text{ECE146}) - H(\text{ECE146}|X),$$

for

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

Which class among ECE102, ECE131 and MUSC15 would you ask about if you want to infer how he/she did in ECE146?

$H(146) = H_b\left(\frac{5}{8}\right) = H_b\left(\frac{3}{8}\right) = 0.95$

$H(146 \mid 102=0) = H_b\left(\frac{1}{3}\right) = 0.92$ $\quad H(146 \mid 102=1) = H_b\left(\frac{4}{5}\right) = H_b\left(\frac{1}{5}\right) = 0.72$

$H(146 \mid 102) = P(102=0)\,H(146 \mid 102=0) + P(102=1)\,H(146 \mid 102=1)$

$= \frac{3}{8}(0.92) + \frac{5}{8}(0.72) = 3(0.115) + 5(0.09) = 0.345 + 0.45 = 0.795$

$H(146 \mid 131=0) = H_b\left(\frac{1}{4}\right) = 0.81$ $\quad H(146 \mid 131=1) = H_b\left(\frac{4}{4}\right) = 0$

$H(146 \mid 131) = P(131=0)\,H(146 \mid 131=0) = \frac{1}{2}(0.81) = 0.405$

$H(146 \mid 15=0) = H_b\left(\frac{3}{4}\right) = H_b\left(\frac{1}{4}\right) = 0.81$ $\quad H(146 \mid 15=1) = H_b\left(\frac{1}{2}\right) = 1$

$H(146 \mid 15) = P(15=0)\,H(146 \mid 15=0) + P(15=1)\,H(146 \mid 15=1) =$

$= \frac{1}{2}(0.81) + \frac{1}{2}(1) \overset{9}{=} 0.405 + 0.5 = 0.905$

CONTINUED ON NEXT PAGE

This page is intentionally left blank for the students to use.

$I(146;\ 102) = H(146) - H(146 \mid 102)$

$= 0.95 - 0.795$

$\boxed{I(146;102) = 0.155}$

$I(146;\ 131) = H(146) - H(146 \mid 131)$

$= 0.95 - 0.405$

0.55 - 0.005

$\boxed{I(146;131) = 0.545}$

$I(146;\ 15) = H(146) - H(146 \mid 15)$

$= 0.95 - 0.905$

$\boxed{I(146;\ 15) = 0.045}$

Since $I(146;131)$ is the greatest of the three, I would first ask about $\boxed{\text{ECE131}}$ to infer how they did in ECE 146.

6. (16 pts) **Support Vector Machine**
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	1	4	1
2	0	3	1
3	1	1	-1
4	3	0	-1

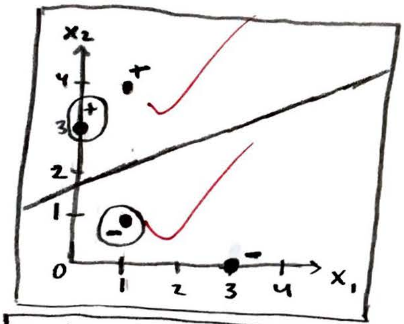
(a) Plot the data. Is the data linearly separable?

(b) Suppose you are asked to find the maximum margin separating hyperplane defined by $[w_1, w_2][x_1, x_2]^T + b = 0$. Write down the (primal) optimization problem explicitly using only w_1, w_2 and b, i.e., plugging in $\boldsymbol{x}^{(i)}$ and $y^{(i)}$.

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

(d) 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.



data is linearly separable

b. $\min_{w_1, w_2, b} \frac{1}{2}(w_1^2 + w_2^2)$ ~~such that~~ $y^{(i)}([w_1 \ w_2]x^{(i)} + b) \geq 1$ -2

c. see plot from part a

point $= \left(\frac{0+1}{2}, \frac{3+1}{2}\right) = \left(\frac{1}{2}, 2\right)$

slope $= -\frac{0-1}{3-1} = \frac{1}{2}$ $\quad x_2 - 2 = \frac{1}{2}\left(x_1 - \frac{1}{2}\right)$

$2x_1 - 4x_2 + 7 = 0$ $\quad x_2 = \frac{1}{2}x_1 + \frac{7}{4}$

$\frac{1}{2}x_1 - x_2 + \frac{7}{4} = 0$

d. $\mathcal{L}(\alpha) = \sum_{i=1}^{4} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{4} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$

$\|x_2\|^2 = 9$

$\|x_3\|^2 = 2$

$\langle x_2, x_3 \rangle = 3$

$\boxed{\alpha_1 = 0 \quad \alpha_4 = 0}$

$= \alpha_2 + \alpha_3 - \frac{1}{2}\left[y_2^2 \alpha_2^2 \|x_2\|^2 + 2y_2 y_3 \alpha_2 \alpha_3 \langle x_2, x_3 \rangle + y_3^2 \alpha_3^2 \|x_3\|^2\right]$

$= \alpha_2 + \alpha_3 - \frac{9}{2}\alpha_2^2 + 3\alpha_2\alpha_3 - \alpha_3^2$

$= 2\alpha_2 - \frac{5}{2}\alpha_2^2$

$\frac{\partial \mathcal{L}}{\partial \alpha_2} = 2 - 5\alpha_2 = 0 \rightarrow \boxed{\alpha_2 = \frac{2}{5}} \rightarrow \boxed{\alpha_3 = \frac{2}{5}}$

$\sum_{i=1}^{4} \alpha_i y_i = 0$

$\alpha_2 y_2 + \alpha_3 y_3 = 0$

$\alpha_2 = \alpha_3$

$w = \sum_{i=1}^{4} \alpha_i y_i x_i$

$= \alpha_2 y_2 x_2 + \alpha_3 y_3 x_3$

11 $= \frac{2}{5}(x_2 - x_3) = \frac{2}{5}\begin{bmatrix}-1\\2\end{bmatrix} \rightarrow \boxed{w = \begin{bmatrix}-2/5\\4/5\end{bmatrix}}$

$b = y^{(i)} - w^T x^{(i)}$

$= 1 - [-2/5 \ \ 4/5]\begin{bmatrix}0\\3\end{bmatrix} = 1 - \frac{12}{5} = -\frac{7}{5} \rightarrow \boxed{b = -\frac{7}{5}}$