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.
  - $H_b(\frac{3}{8}) = 0.95443, H_b(\frac{1}{3}) = 0.91830, H_b(\frac{1}{4}) = 0.81128, H_b(\frac{1}{5}) = 0.72193,$

Your Name:

Your ID Number:

Name of person on your left:

Name of person on your right:

Problem	Score	Possible
1		0
2		0
3		0
4		0
5		0
6		0
Total		100

1. (0 pts) **True or False.**

   Circling the correct answer is worth +3 points, circling the incorrect answer is worth −1 points. Not circling either is worth 0 points.

   (a) The perceptron algorithm does not converge if the training samples are not linearly separable.
   Solution: True. The perceptron algorithm's update rule is to adjust the parameters if a point is misclassified, so if it is not possible to correctly classify all training data (e.g. linearly separable) then the algorithm will not reach a stable point.

   (b) k-nearest neighbors will always give a linear decision boundary.
   Solution: False. counter example is given in the sketch the 1-nearest neighbor decision boundary problem.

   (c) The derivative of the sigmoid function $\sigma(x) = \frac{1}{1+\exp(-x)} = \frac{\exp(x)}{1+\exp(x)}$ satisfies: $\sigma'(x) = \sigma(x)(1 - \sigma(x))$.
   Solution: True. Use the second expression and the quotient rule we get

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

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

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

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

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

   Solution: True. We do a change of variable for the second optimization problem with $w_2 = 2w_3$ and $b_2 = 2b_3$. We find the second optimization problem is equivalent to the first one with $w_1 = w_3 = \frac{1}{2}w_2$ and $b_1 = b_3 = \frac{1}{2}b_2$. The two hyperplanes are therefore identical.

   (e) For $x_1, x_2 \in \mathbb{R}$, $K(x_1, x_2) = (1 + x_1x_2)^2$ is a valid kernel.
   Solution: True.

$$K(x_1, x_2) = (1 + x_1x_2)^2 = 1 + 2x_1x_2 + x_1^2x_2^2 = \phi(x_1)^T\phi(x_2),$$

   where

$$\phi(x) = \begin{bmatrix} 1 \\ \sqrt{2}x \\ x^2 \end{bmatrix}.$$

2. (0 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 constant learning rate factor.

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

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

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

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

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

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 initial weight $\boldsymbol{w} = [w_0, w_1, w_2] = [0, 1, 1]$. 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 $[10, 10]^T$: no update.
Update weights based on $[0, 0]^T$: $\boldsymbol{w} \leftarrow \boldsymbol{w} - [1, 0, 0] = [-1, 1, 1]$.
Update weights based on $[3, 3]^T$: $\boldsymbol{w} \leftarrow \boldsymbol{w} - [1, 3, 3] = [-2, -2, -2]$.
Update weights based on $[4, 8]^T$: $\boldsymbol{w} \leftarrow \boldsymbol{w} + [1, 4, 8] = [-1, 2, 6]$.

3. (0 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:

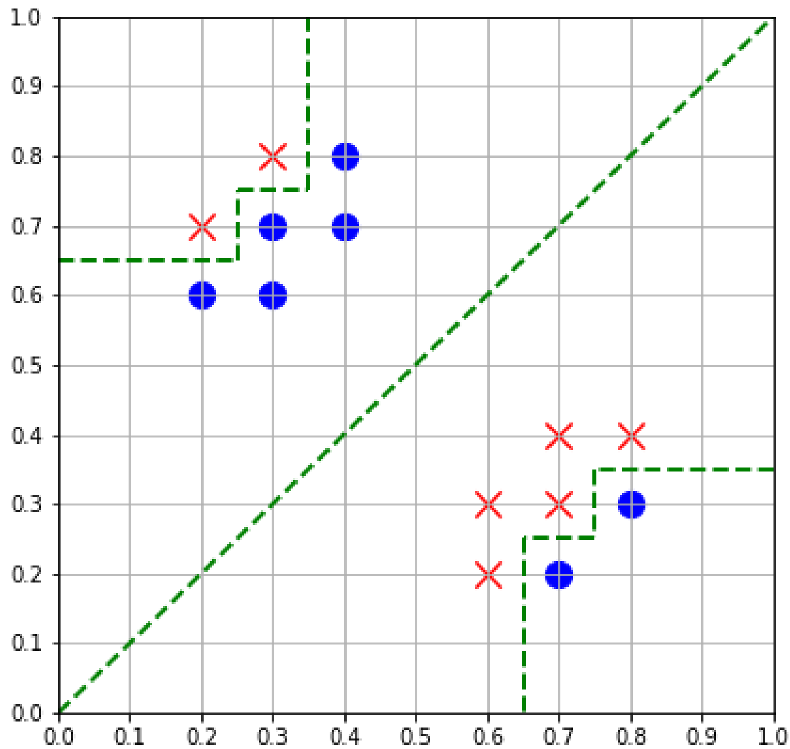


Figure 1: k-Nearest Neighbors

(a) In above figure, sketch the 1-nearest neighbor decision boundary for this dataset.
Solution:
The decision boundaries are shown above.

(b) What value of k maximize leave-one-out cross-validation error for this dataset? What is the resulting error?
Solution:
$k = 9$ or 13 maximize leave-one-out cross-validation error for this dataset, and the resulting error is 1.

4. (0 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 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

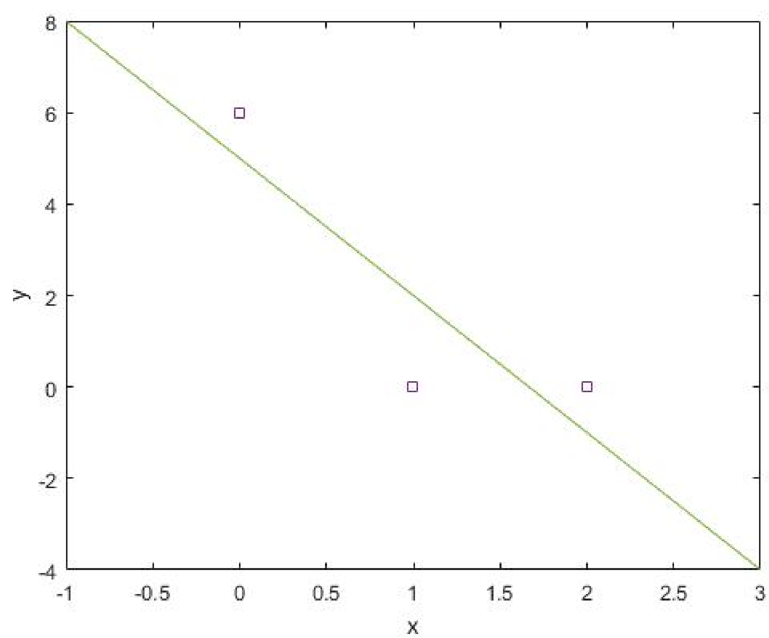
$$J(\boldsymbol{w}) = \|\boldsymbol{X}\boldsymbol{w} - y\|^2,$$

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

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}, y = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix}.$$

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y = \begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix}^{-1} \times \begin{bmatrix} 6 \\ 0 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} 5 & -3 \\ -3 & 3 \end{bmatrix} \times \begin{bmatrix} 6 \\ 0 \end{bmatrix} = \begin{bmatrix} 5 \\ -3 \end{bmatrix}.$$

The plot:



5. (13 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 affect 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: ECE102(*System 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?
Solution:

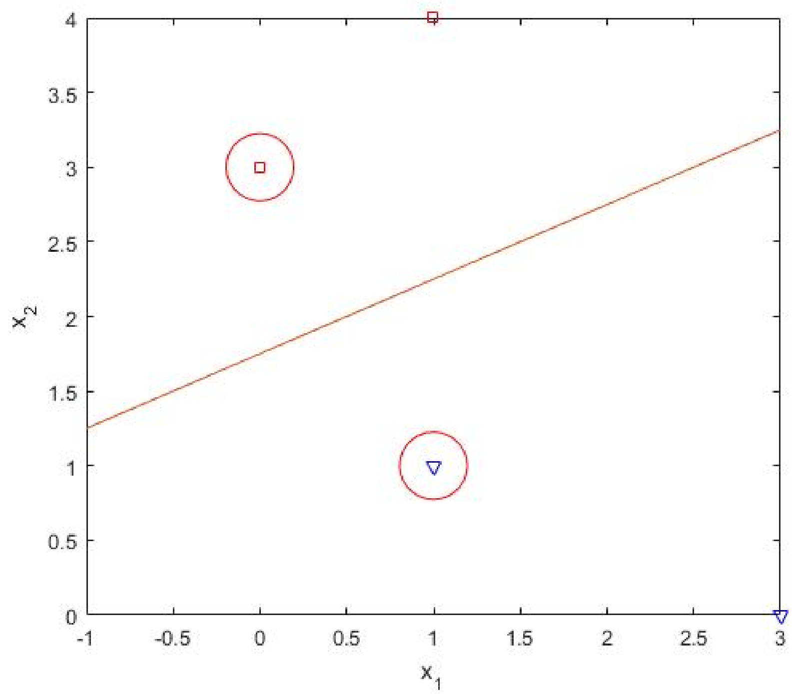
$$I(\text{ECE146;ECE102}) = H_b(\frac{3}{8}) - \frac{5}{8}H_b(\frac{1}{5}) - \frac{3}{8}H_b(\frac{1}{3}) \approx 0.1589;$$
$$I(\text{ECE146;ECE131}) = H_b(\frac{3}{8}) - \frac{1}{2}H_b(1) - \frac{1}{2}H_b(\frac{1}{4}) \approx 0.5488;$$
$$I(\text{ECE146;MUSC15}) = H_b(\frac{3}{8}) - \frac{1}{2}H_b(\frac{1}{2}) - \frac{1}{2}H_b(\frac{1}{4}) \approx 0.0488.$$

You want to ask how he/she did in ECE131 because the information gain is the highest.

6. (10 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?
Solution: Yes, data is linearly separable.



(b) 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:

$$\begin{aligned} \min_{w_1, w_2, b} \quad & w_1^2 + w_2^2 \\ s.t. \quad & w_1 + 4w_2 + b \geq 1, \\ & 3w_2 + b \geq 1, \\ & -w_1 - w_2 - b \geq 1, \\ & -3w_1 - b \geq 1. \end{aligned}$$

(c) 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 $[1,1]^T$ and $[0,3]^T$. The line that has normal vector $[-1,2]$ and also pass through the midpoint of support vectors $\left(\left[\frac{1}{2},2\right]^T\right)$ is $-x_1 + 2x_2 - 3.5 = 0$.

(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.
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

$$\begin{aligned}
W(\boldsymbol{\alpha}) &= \sum_{i=1}^{4} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{4} y_i y_j a_i a_j \boldsymbol{x}^{(i)T}\boldsymbol{x}^{(j)} \\
&= 2\alpha_0 - \frac{1}{2}\alpha_0^2 \boldsymbol{x}^{(2)T}\boldsymbol{x}^{(2)} + \alpha_0^2 \boldsymbol{x}^{(2)T}\boldsymbol{x}^{(3)} - \frac{1}{2}\alpha_0^2 \boldsymbol{x}^{(3)T}\boldsymbol{x}^{(3)} \\
&= 2\alpha_0 - \frac{5}{2}\alpha_0^2.
\end{aligned}$$

Maximizing $W(\boldsymbol{\alpha})$ over α_0, we get $\alpha_3 = \alpha_2 = \alpha_0 = \frac{2}{5}$. Using $\boldsymbol{w} = \sum_{m\in\mathcal{S}} \alpha_m y^{(m)} \boldsymbol{x}^{(m)}$, we get $\boldsymbol{w} = [-\frac{2}{5}, \frac{4}{5}]^T$. 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 $b = -\frac{7}{5}$. 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.