ECE M146 | Spring 2021 Exam 1
Introduction to Machine Learning | April 19, 2021
Instructor: Lara Dolecek | Maximum score: 100 points

You have 2 hours to submit your work **directly on Gradescope under the Exam 1 submission link**.

Please read and carefully follow all the instructions.

Instructions

- The exam is accessible from 10 am PST on April 19th to 10 am PST on April 20th. Once you open the exam, you will have 2 hours to upload your work (therefore open the exam at least 2 hours before the closing time).
- This exam is open book, open notes. You are allowed to consult your own class notes (homework, discussion, lecture notes, textbook). You are not allowed to consult with each other or solicit external sources for help (e.g., an online forum).
- For each question, start a new sheet of paper. Therefore, the number of pages of your scan should be at least the number of questions. It is ok to write multiple parts of a question on one sheet. Properly erase or cross out any scratch work that is not part of the answer.
- Please submit your exam through the corresponding submission link on Gradescope.
- Make sure to include your **full name** and **UID** in your submitted file.
- Make sure to **show all your work**. Unjustified answers will be at a risk of losing points.
- Calculators are allowed for matrix inversion, entropy calculation and etc.
- **Policy on the Academic Integrity**
  "During this exam, 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 10 points** if we can not find this statement.

I $\underline{\quad YourName \quad}$ with UID ____ have read and understood the policy on academic integrity.

1. (25 pts) **Perceptron**

(a) Write down the perceptron learning rule by filling in the blank below with a proper sign (+ or -).

i. Input x is falsely classified as negative:

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

ii. Input x is falsely classified as positive:

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

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

Sample #	1	2	3	4
x	[10,10]	[1,0]	[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 $w = [w_0, w_1, w_2] = [0, 1, 1]$.
Solution:
Starting weights: $w = [0, 1, 1]$.
Update weights based on $[10, 10]^T$: no update.
Update weights based on $[0, 0]^T$: $w \leftarrow w - [1, 1, 0] = [-1, 0, 1]$.
Update weights based on $[3, 3]^T$: $w \leftarrow w - [1, 3, 3] = [-2, -3, -2]$.
Update weights based on $[4, 8]^T$: $w \leftarrow w + [1, 4, 8] = [-1, 1, 6]$.

(c) Did the perceptron algorithm converge after the single iteration in (b) and why?
Solution:
No, the algorithm did not converge because with the current weight $[-1, 1, 6]$, data 3 is still misclassified.

(d) Suppose we run the algorithm for more iterations, will the algorithm converge and why?
Solution:
Yes, the data is linearly separable and the perceptron algorithm will converge.

(e) Suppose we get $w = [w_0, w_1, w_2] = [-10, 2, 1]$ when the algorithm terminates. What is the distance from Sample #3 to this learned hyperplane?
Solution:
Using the distance formula, we get:

$$d = \frac{|w_1 x_1 + w_2 x_2 + w_0|}{\|[w_1, w_2]^T\|} = \frac{1}{\sqrt{5}}.$$

2. (20 pts) **Linear Regression**
You are given the following three data points:

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 4 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \end{bmatrix}.$$

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

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

In matrix-vector form, the objective function is

$$J(w) = \|Xw - y\|^2,$$

for some X, y and $w = [w_0, w_1]^T$.

(a) What are X and y?

(b) What is the optimal w that minimizes the objective function?

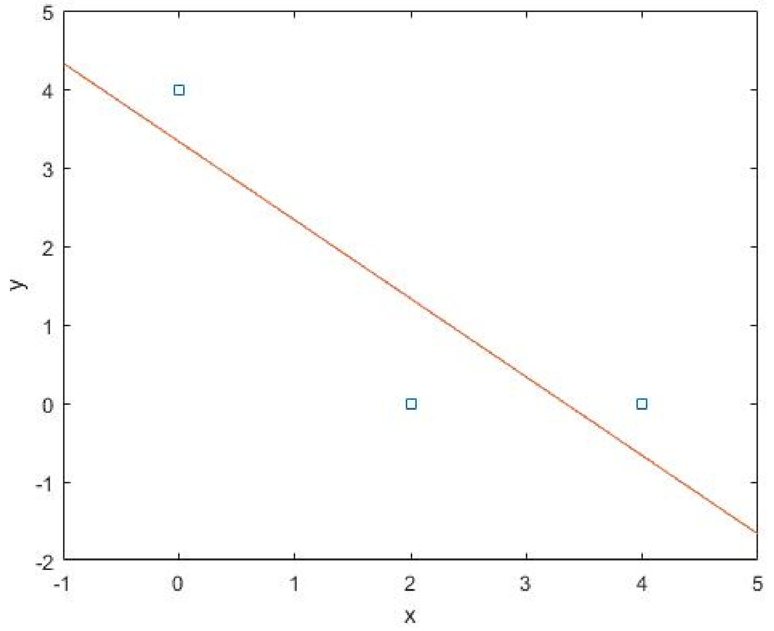
(c) Draw the three data points and the fitted line.

Solution:

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 2 \\ 1 & 4 \end{bmatrix}, y = \begin{bmatrix} 4 \\ 0 \\ 0 \end{bmatrix}.$$

$$w = (X^T X)^{-1} X^T y = \begin{bmatrix} 3 & 6 \\ 6 & 20 \end{bmatrix}^{-1} \times \begin{bmatrix} 4 \\ 0 \end{bmatrix} = \frac{1}{24} \begin{bmatrix} 20 & -6 \\ -6 & 3 \end{bmatrix} \times \begin{bmatrix} 4 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{10}{3} \\ -1 \end{bmatrix}.$$

The plot:



3. (20 pts) **Regularized Least Square and Gradient Descent**
Suppose you have n data points $\{x_i, y_i\}$ where $x_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$, you want to minimize the following loss function:

$$J(w) = \frac{1}{2}\sum_{i=1}^{n}(w^T x_i - y_i)^2 + \frac{\lambda}{2}\sum_{j=1}^{m} w_j^2,$$

where λ is a constant, $w \in \mathbb{R}^m$ and w_j are the j-th element of w.

(a) Let

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Write the loss function in matrix-vector form using X, y and w.
Solution: In matrix vector form:

$$J(w) = \frac{1}{2}\|Xw - y\|^2 + \frac{\lambda}{2}\|w\|^2.$$

(b) Derive the analytical solution of w that minimizes the loss function.
Note: you may assume the matrix $X^T X + \lambda I$ is invertible. You also don't have to show the w you found minimizes the lost function instead of maximizing it.
Solution: We find the gradient of $J(w)$ and set it equals to 0.

$$\begin{aligned} \nabla_w J(w) = X^T X w - X^T y + \lambda w &= 0 \\ (X^T X + \lambda I) w &= X^T y \\ w &= (X^T X + \lambda I)^{-1} X^T y. \end{aligned}$$

(c) Denote the j-th element of x_i as $x_{i,j}$. Find the gradient of $J(w)$ with respect to w_j.
Solution: Using the chain rule or the gradient in part (b), we get:

$$\frac{\partial J(w)}{\partial w_j} = \sum_{i=1}^{n}\left[(w^T x_i - y_i)x_{i,j}\right] + \lambda w_j. \tag{1}$$

(d) With a learning rate η, derive the stochastic gradient descent (SGD) rule to minimize the loss function for parameter w_j.
Solution: When performing SGD, only a single data point is considered at a times so we update w_j based the following rule:

$$w_j = w_j - \eta\left[(w^T x_i - y_i)x_{i,j} + \frac{\lambda}{n} w_j\right].$$

Note that we have a normalization factor n for the latter part because we can rewrite the latter part of (1) as $\frac{\lambda}{n}\sum_{i=1}^{n} w_j$.

4. (10 pts) **Logistic Regression**
Show that the derivative of the sigmoid function $\sigma(x) = \frac{1}{1+\exp(-x)}$ satisfies: $\sigma'(x) = \sigma(x)(1 - \sigma(x))$.
Solution:
We rewrite the sigmoid function as $\sigma(x) = \frac{\exp(x)}{1+\exp(x)}$. Use this expression and the quotient rule we get

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

5. (25 pts) **Decision Tree**
Decision tree has been used for medical diagnostic because it can be easily interpreted by human. There are 8 patients who have been asked to answer yes (1) or no (0) for several symptoms. Their answer and whether they are diagnosed with $XDisease$ are summarized in the following table:

Patient #	Fatigue	Fever	Cough	Headache	XDisease
1	0	0	0	1	0
2	0	1	0	1	1
3	1	0	0	1	0
4	1	1	1	1	1
5	0	1	1	0	1
6	1	1	0	0	1
7	0	0	0	0	0
8	1	0	1	1	1

(a) What is the binary entropy of this data set, i.e., $H(XDisease)$?
Solution: Define the binary entropy function as follows:

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

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

(b) Calculate the conditional entropy of

$$H(XDisease|X), \text{for}\, X \in \{Fatigue, Fever, Cough, Headache\},$$

i.e., the conditional entropy of $XDisease$ conditioning on the features.
Solution:

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

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

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

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

(c) Calculate the information gain:

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

for

$$X \in \{Fatigue, Fever, Cough, Headache\}.$$

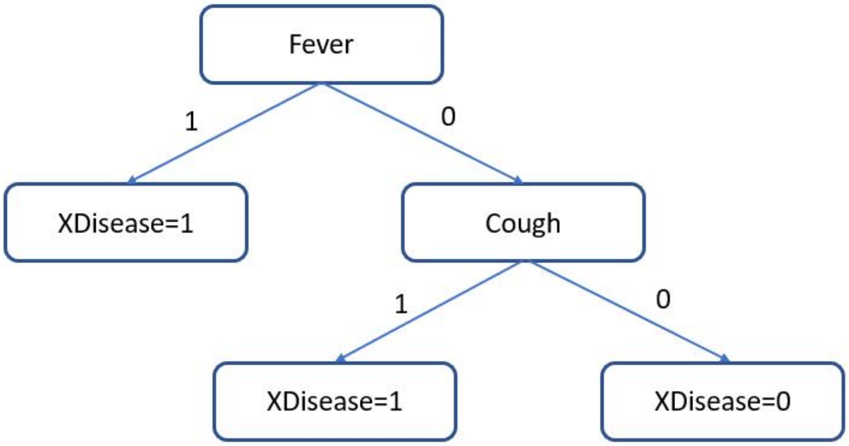
Solution:

$$\begin{aligned} I(XDisease|Fatigue) &= 0.9544 - 0.9056 = 0.0488; \\ I(XDisease|Fever) &= 0.9544 - 0.4056 = 0.5488; \\ I(XDisease|Cough) &= 0.9544 - 0.6068 = 0.3476; \\ I(XDisease|Headache) &= 0.9544 - 0.9512 = 0.0032. \end{aligned}$$

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

Solution: We choose *Fever* which has the largest information gain.

(e) Suppose we get the following tree as our final decision tree. Determine if patients 9 and 10 are positive for *XDisease* or not based on the decision tree you made.



Patient #	Fatigue	Fever	Cough	Headache	XDisease
9	1	0	1	0	?
10	1	0	0	0	?

Solution:

Patient 9: Positive

Patient 10: Negative