Machine Learning Fundamentals **Name:** ____________________

CISC 435-01-2023/Fall

Exam 01

10/19/2021

Time Limit: 90 Minutes Mina Gabriel

Answer four out of six questions, you can choose any four. Good luck!

1. (25 points) For function $f\left(\mathrm{x}_1, \mathrm{x}_2\right) = x_1^2 + 2x_2^2 + 3x_1x_2 + 2x_1 + 6$ show that you can write it as the quadratic form $\mathbf{x}^T \mathrm{Q} \mathbf{x} + \mathbf{b}\mathbf{x} + c$, where $\mathbf{x}$ is a 2×1 vector, Q is symmetric 2×2 matrix $\mathrm{Q} = \begin{bmatrix} 1 & \frac{3}{2} \\ \frac{3}{2} & 2 \end{bmatrix}$, $\mathbf{b}$ is 2×1 vector $\mathbf{b} = \begin{bmatrix} 2 & 0 \end{bmatrix}$, and c is a scalar $c = 6$.

   Hint:

$$\text{let } X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$X^T Q X + bx + c = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & 3/2 \\ 3/2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 2 \\ 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + 6$$

2. (25 points) The following problem is related to Bayes Theorem that is stated as $P(A \mid B) = P(B \mid A)P(A)/P(B)$, where A and B are random variables. Let us assume A is a binary random variable stating whether a woman has breast cancer and B is a binary random variable saying whether a woman tested positive on a mammogram. Let us assume the following background knowledge: 0.1% of women who have breast cancer; 90% of women who have breast cancer test positive on mammograms; 8% of healthy women who test positive on mammograms. What is the probability that a woman has cancer if she has a positive mammogram result? What is the probability that a woman has cancer if she has a negative mammogram result?

   Hint:

   Probability that a woman has cancer given she tested positive:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)}$$

3. (25 points) Gradient Descent for Linear Regression

Consider a simple linear regression model defined as $\hat{y} = w_0 + w_1 x_1 + w_2 x_2$

Given a dataset of n samples, the mean squared error (MSE) for this model is represented as:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Your tasks are:

a. Derive the gradient of the mean squared error (MSE) with respect to each of the weights w_0, w_1, and w_2. Show all your steps.

b. Express the gradient of all the weights in a vector form, commonly denoted as $\nabla_w \text{MSE}$ in other words:

$$\nabla_w MSE = \begin{bmatrix} \frac{\partial}{\partial w_0} MSE \\ \frac{\partial}{\partial w_1} MSE \\ \frac{\partial}{\partial w_2} MSE \end{bmatrix} = \frac{-2}{n} \begin{bmatrix} \cdots \\ \cdots \\ \cdots \end{bmatrix}$$

Extra Points: $\text{ReLU}(x) = \max(0, x)$ applying ReLU into the linear model, we get $\hat{y} = \text{ReLU}\left(w_0 + w_1 x_1 + w_2 x_2\right)$ Discuss the effect of incorporating this activation function to your protection on the gradient update.

4. (25 points) Given the following data points:

Point	**x**	**y**	**Class**
P1	1	3	A
P2	4	2	B
P3	2	6	A
P4	5	4	B

Compute:

(a) [10 points] Using KNN with $k = 3$ and the Euclidean distance, predict the class of the point P(3, 3).

(b) [10 points] Repeat the prediction for the same point P using $k = 2$.

(c) [5 points] How do we know the best value of k?

5. (25 points) Given the following 2D data points:

Point	**x**	**y**
A	2	3
B	5	4
C	3	7
D	8	5
E	2	6
F	6	8
G	7	2
H	1	4

Perform the following tasks:

(a) [5 pts] Plot the data points on a 2D plane. Label each point according to the table.

(b) [10 pts] Using the k-Means clustering algorithm, cluster the data points into 2 clusters. If doing by hand:

- Initialize the centroids by choosing two random points.
- Show the iterations and the updated centroids two times only.
- Highlight each cluster using different colors or symbols.

(c) [5 pts] Plot the final clusters with their centroids. Label each point and centroid clearly.

(d) [5 pts] Discuss the choice of initial centroids and its potential impact on the final clusters obtained. Would a different choice of initial centroids change the result? How come the Elbow Method will help us decide about the number of centroids? What does it mean to use WCSS $= \sum_{x\in c} d(x, c_1)^2 + \sum_{x\in c} d(x, c_2)^2 + \sum_{x\in c} d(x, c_3)^2 + \ldots$ as an objective function

6. (25 points) Consider the linear regression model given by $\hat{y} = w_0 + w_1 x$ Where $\hat{y}$ is the predicted value, x is the input feature, and w_0 and w_1 are the bias and weight respectively.

   Given the mean absolute error (MAE) loss function:

$$L = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

   You are provided with the following data point: $x = 2, y = 3$.

   Initially, $w_0 = 0.5$ and $w_1 = 1$.

   Using a learning rate (η) of 0.05, perform a single gradient descent update to adjust w_0 and w_1.

   Hint:

$$w^{new}_{\in\{0,1\}} \leftarrow w^{old}_{\in\{0,1\}} - \eta \frac{\partial L}{\partial w_{\in\{0,1\}}}$$