CSE-4819/5820 Introduction to Machine Learning
Written Assignment 1
Linear Algebra; Convex Optimization; Linear Regression

Total: 50pt

Due: End of Day (23:59), September 20th

Requirements:

1. The written assignment will be graded based on correctness, accuracy and clarity.
2. Please prepare your answers, and submit the final assignment in a **pdf** file through HuskyCT assignment portal. It is recommended that you prepare the assignment using MS Word and then convert it into pdf. Please put your **full name** at the beginning of the document for ease of grade recording.
3. Please explicitly indicate the ids of the questions you are answering for ease of grading.
4. Even if your answer is not correct, you may still get certain partial marks based on your calculation/analysis process. Please present the necessary calculation process if any.
5. Each student will have a total of 5 free late (calendar) days to use for homework. Each 24 hours or part thereof that an assignment is late uses up one full late day. Please note: once these late days are exhausted, no late assignments will be accepted for any reason. Students are highly encouraged to reserve your late days for unavoidable emergencies, planned travel, etc.

Q1. Linear Algebra and Probability (22pt)

(1) (8pt) Given the following four vectors,

x_1 = [0, 0.2, 1.0, 2.2]

x_2 = [0.7, 0.2, 0.5, 2.0]

x_3 = [0, 1.0, 1.5, 2.2]

x_4 = [0.8, 0.1, 1.2, 2.0]

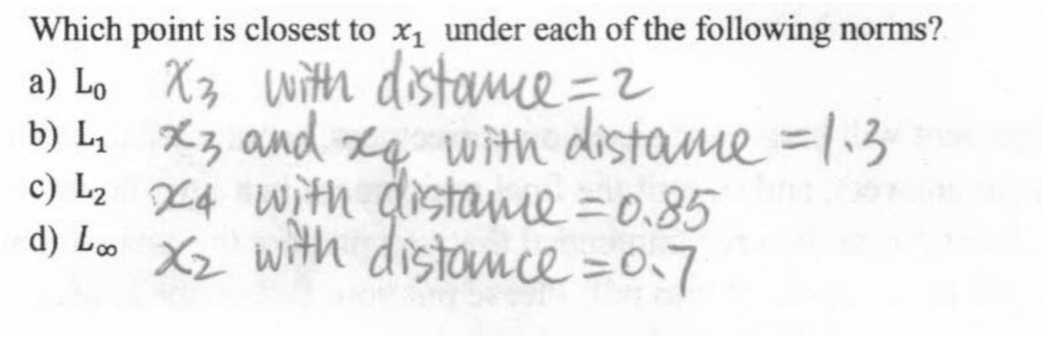
Which point is closest to x_1 under each of the following norms?

a) L_0

b) L_1

c) L_2

d) L_∞



Which point is closest to x_1 under each of the following norms?

a) L_0 X_3 with distance = 2

b) L_1 x_3 and x_4 with distance = 1.3

c) L_2 x_4 with distance = 0.85

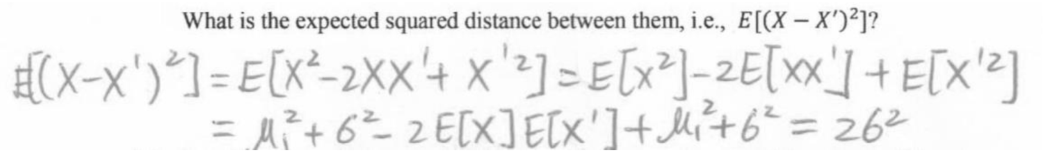
d) L_∞ X_2 with distance = 0.7

(2) (4pt) If $X \sim N(\mu, \sigma^2)$, E[X] = μ, Var[X] =σ^2, and $E[X^2] = \mu^2 + \sigma^2$. Also, recall that expectation is linear, so it obeys the following three properties:

$E[X + c] = E[X] + c$ for any constant c,
$E[X + Y] = E[X] + E[Y]$,
$E[aX] = aE[X]$ for any constant a.

We note that if X and X' are independent, then $E[XX'] = E[X]E[X']$.
Consider two points (sampled independently) from the same class follow: $X \sim N(\mu_1, \sigma^2)$ and $X' \sim N(\mu_1, \sigma^2)$.

What is the expected squared distance between them, i.e., $E[(X - X')^2]$?

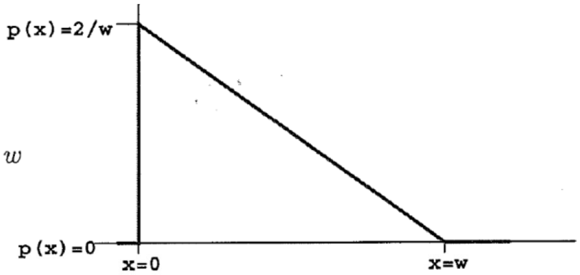


What is the expected squared distance between them, i.e., $E[(X - X')^2]$?

$$E[(X-X')^2] = E[X^2 - 2XX' + X'^2] = E[X^2] - 2E[XX'] + E[X'^2]$$
$$= \mu_1^2 + \sigma^2 - 2E[X]E[X'] + \mu_1^2 + \sigma^2 = 2\sigma^2$$

(3) (4pt) Consider the probability density function shown in the following figure and equations.

$$
\begin{aligned}
p(x) &= 0 && \text{if } x < 0 \\
p(x) &= \frac{2}{w} - \frac{2x}{w^2} && \text{if } 0 \le x \le w \\
p(x) &= 0 && \text{if } w < x
\end{aligned}
$$



(2pt) Which *one* of the following expressions is true?

(a) $E[X] = \int_{x=-\infty}^{\infty} w(\frac{2}{w} - \frac{2x}{w^2})dx$

(b) $E[X] = \int_{x=0}^{w} x(\frac{2}{w} - \frac{2x}{w^2})dx$

(c) $E[X] = \int_{x=0}^{w} w(\frac{2}{w} - \frac{2x}{w^2})dx$

(d) $E[X] = \int_{x=-\infty}^{\infty} (\frac{2}{w} - \frac{2x}{w^2})dx$

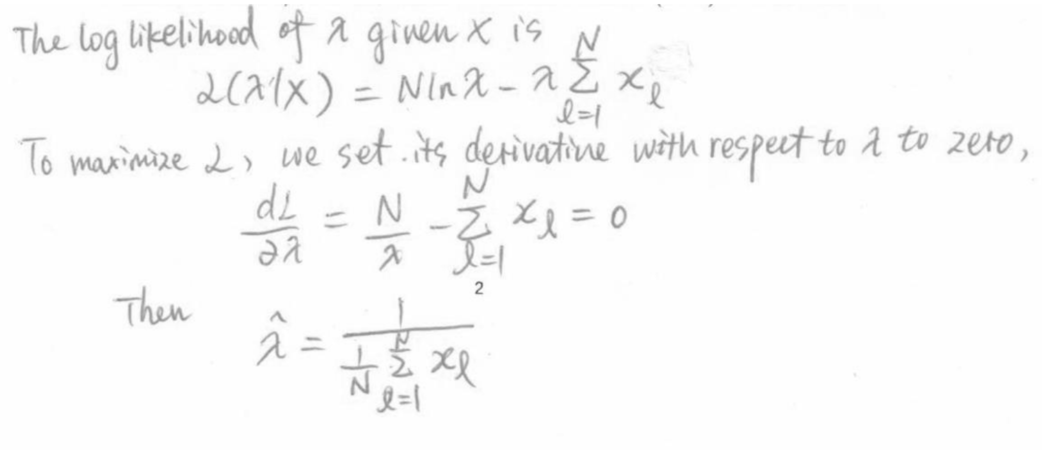
(2pt) What is $p(x = 1|w = 2)$? 0.5

(4) (6pt) Consider a feature x which is a continuous random variable with possible outcomes being all the nonnegative real numbers. The random variable follows a distribution with the following probability density function (PDF):

$$p(x \mid \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0, \end{cases}$$

where the parameter λ of the distribution is a positive real number.

Given a data set X = $\{x_1, x_2, \dots, x_N\}$ drawn *i.i.d.* (independent and identically distributed) from the distribution, derive the maximum likelihood estimate (MLE) of λ based on X.



The log likelihood of λ given X is

$$\mathcal{L}(\lambda|X) = N\ln\lambda - \lambda\sum_{l=1}^{N} x_l$$

To maximize $\mathcal{L}$, we set its derivative with respect to λ to zero,

$$\frac{d\mathcal{L}}{\partial\lambda} = \frac{N}{\lambda} - \sum_{l=1}^{N} x_l = 0$$

2

Then

$$\hat{\lambda} = \frac{1}{\frac{1}{N}\sum_{l=1}^{N} x_l}$$

Q2. Introduction to Optimization (16pt)

(1) (10pt) Please justify if the following statement is correct or not (2 pt for correct T/F; 3 pt each for correct explanation).

(a) In machine learning, the optimization problem we are solving to train a model is always a maximization problem.

Q2 (a) True or False. True: ∵ Any optimization with minimization can be converted int a maximization problem. False: Sometimes we are maximizing a likelihood, and (by flipping the sign) sometimes we are minimize a cost function.

(b) For any constant $c \in R$, $f(x) = \frac{x^2}{c-x}$ is convex on $-\infty < x < c$.

Q2(b)

$$\frac{x^2}{c-x} = -\frac{x^2}{x-c} = -\left[x+c+\frac{c^2}{x-c}\right]$$

$$= -(x+c) - \frac{c^2}{x-c}$$

Each of $-(x+c)$ and $-\frac{c^2}{x-c}$ is a convex function on $-\infty < x < c$.

(2) (6pt) Use the method of Lagrange multipliers to find the maximum values of the objective function. Please provide the maximum values and the corresponding variables.
objective function: maximize $f(x, y) = 6xy$
subject to: $x^2/9 + y^2/16 = 1$.

(2) $L = f(x, y) + \lambda\left(\frac{x^2}{9} + \frac{y^2}{16} - 1\right)$.

$$\frac{\partial L}{\partial x} = 6y + \frac{2\lambda x}{9} = 0 \quad \frac{\partial L}{\partial y} = 6x + \frac{2\lambda y}{16} = 0, \quad \frac{\partial L}{\partial \lambda} = \frac{x^2}{9} + \frac{y^2}{16} - 1 = 0$$

$$\Rightarrow \lambda = \frac{-27y}{x} = \frac{-48x}{y} \quad \Rightarrow \frac{x^2}{9} = \frac{y^2}{16}.$$

Since $\frac{x^2}{9} + \frac{y^2}{16} = 1 \quad \Rightarrow \quad x = \pm\frac{3\sqrt{2}}{2} \quad y = \pm 2\sqrt{2}$

For maximizing $6xy$,

(x, y) is either $\left(\overset{3}{\frac{3\sqrt{2}}{2}}, 2\sqrt{2}\right)$ or $\left(-\frac{3\sqrt{2}}{2}, -2\sqrt{2}\right)$
maximum value of $6xy$ is 36.

Q3. Linear Regression (12pt)

Given known $\mathbf{X} \in \mathrm{R}^{n\times d}$, $\mathrm{y} \in \mathrm{R}^{n\times 1}$, and unknown $\mathrm{w} \in \mathrm{R}^{d\times 1}$, $\mathrm{y} = \mathbf{X}w + \epsilon$, where $\epsilon \sim \mathrm{N}(0, \sigma^2\mathbf{I})$. The task is to estimate w.

(a) (6pt) Please write down the loss function for the linear regression. Then derive the closed form estimation for w based on the least square method. Note that the derive process is required and we assume that $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, *i.e.*, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ exists.

(b) (6pt) Given $\mathrm{X} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 1 & 3 \end{bmatrix}$, and $\mathrm{y} = \begin{bmatrix} 5 \\ 3 \\ 2 \end{bmatrix}$. Using the closed form estimation for w based on the least square method, please compute $\boldsymbol{X}^T\boldsymbol{X}$, $\boldsymbol{X}^Ty$ and the estimated w.

$$L(w) = \|y - \mathbb{X}w\|_2^2$$

or $$\sum_{i=1}^{n} (y_i - w^T x_i)^2$$

or $$L(w) = (y - \mathbb{X}w)^T (y - \mathbb{X}w)$$

$$\begin{aligned} L(w) &= (y - \mathbb{X}w)^T (y - \mathbb{X}w) \\ &= (y^T - w^T \mathbb{X}^T)(y - \mathbb{X}w) \\ &= y^T y - y^T \mathbb{X} w - w^T \mathbb{X}^T y + w^T \mathbb{X}^T \mathbb{X} w \end{aligned}$$

$$\frac{\partial L(w)}{\partial w} = 0 - 2\mathbb{X}^T y + 2\mathbb{X}^T \mathbb{X} w = 0 .$$

$$\Rightarrow \mathbb{X}^T \mathbb{X} w = \mathbb{X}^T y \quad \Rightarrow w = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T y$$

$$x^T x = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 6 & 8 \\ 8 & 14 \end{bmatrix}$$

$$x^T y = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 5 \\ 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 13 \\ 17 \end{bmatrix}$$

$$\hat{w} = (x^T x)^{-1} x^T y = \begin{bmatrix} 2.3 \\ -0.1 \end{bmatrix}$$