Midterm Exam

- Total duration: 90 minutes.
- You **can** use one page as a cheat sheet.
- You **cannot** consult your notes, textbooks, Google, or any other form of external help.
- Maximum points: 60. Any score above 60 will be rounded to 60.
- Once you are finished, please scan, or take a picture of, your answers and upload on NYUClasses before 8pm ET. You will have to include your cheat sheet, if you used one. No late submissions will be accepted.
- Good luck and stay safe! Haotian Yi

1. **(3 points)** Please write down the time at the *start* and *end* of your exam. The difference should not exceed 90 minutes. Please also write down your *name* and *signature* below; by doing so, you are affirming the NYU Tandon School of Engineering student code of conduct.

Start 2:00 PM

end 3:05 PM

3:05 – 3:23 PM writing an additional SVMs for last question

Haotian Yi

N18800809

hy1651

2. **(10 points)** This is a slight variant of a homework problem. Let $\{x_1, x_2, \ldots, x_n\}$ be a set of points in d-dimensional space, and let $\{p_1, p_2, \ldots, p_n\}$ denote a probability distribution over the integers $[1, 2, \ldots, n]$. Suppose we wish to produce a single point estimate $\mu \in \mathbb{R}^d$ that minimizes the *weighted* squared-error:

$$L(\mu) = p_1\|x_1 - \mu\|_2^2 + p_2\|x_2 - \mu\|_2^2 + \ldots + p_n\|x_n - \mu\|_2^2$$

Find a closed form expression for the optimal μ and prove that your answer is correct.

$$L(\mu) = \sum_{i=1}^{n} p_i \|x_i - \mu\|_2^2$$

take derivative and set it to zero:

$$\frac{\partial L(\mu)}{\partial \mu} = \sum_{i=1}^{n} \left[-2 p_i (x_i - \mu)\right] = 0$$

$$\sum_{i=1}^{n} (p_i x_i) = \sum_{i=1}^{n} (p_i \mu)$$

$$\mu = \frac{\sum_{i=1}^{n} (p_i x_i)}{\sum_{i=1}^{n} p_i}$$

3. **(15 points)** This is a variant of a homework problem. Suppose x is a d-dimensional input, w is a d-dimensional variable, and λ is a regularization parameter.
   a. Show that the minimizer of the squared-error loss with ℓ_1 regularizer:

$$L(w) = \frac{1}{2}\|x - w\|_2^2 + \lambda \|w\|_1$$

   is given by:

$$w_i^* = \begin{cases} x_i - \lambda & \text{if} \quad x_i > \lambda, \\ x_i + \lambda & \text{if} \quad x_i < -\lambda, \\ 0 & \text{otherwise.} \end{cases}$$

   b. Show that the minimizer of the squared-error loss with ℓ_2 regularizer:

$$L(w) = \frac{1}{2}\|x - w\|_2^2 + \lambda \|w\|_2^2$$

   is given by:

$$w_i^* = \left(\frac{1}{1 + 2\lambda}\right) x_i.$$

   c. In class, we argued via contour plots that greater ℓ_2 regularization encourages "small" solutions, while greater ℓ_1 regularization encourages "sparse" solutions. Mathematically justify why that is the case by examining the structure of the optimal solutions derived above.

a. $L(w) = \frac{1}{2}\sum_{i=1}^{d}(x_i - w_i)^2 + \lambda \sum_{i=1}^{d}|w_i|$

$$\frac{\partial L(w)}{\partial w} = (-2)\cdot\frac{1}{2}\sum_{i=1}^{d}(x_i - w_i) + \lambda\sum_{i=1}^{d}\frac{\partial |w_i|}{\partial w_i} = 0$$

for each dimension i :

① when $w_i > 0$

$$\frac{\partial L_i(w_i)}{\partial w_i} = w_i - x_i + \lambda = 0$$

$\Rightarrow w_i^* = x_i - \lambda$ (corresponding to $x_i > \lambda$)

② when $w_i < 0$

$$\frac{\partial L_i(w_i)}{\partial w_i} = w_i - x_i - \lambda = 0$$

$\Rightarrow w_i^* = x_i + \lambda$ (corresponding to $x_i < -\lambda$)

③ when $w_i = 0$, $w_i^* = 0$ (corresponding to "otherwise")

thus question (a) solved.

b. $$\frac{\partial L(w)}{\partial w} = (-2)\cdot\frac{1}{2}(X-W) + 2\lambda W = 0$$

$$-X + W + 2\lambda W = 0$$

$$W = \frac{X}{1+2\lambda}$$

$$\Rightarrow W_i^* = \frac{X_i}{1+2\lambda} = \left(\frac{1}{1+2\lambda}\right) X_i$$

c. for l_2 regularization,

we can see from (b) that

$$W_i^* = \frac{1}{1+2\lambda} X_i$$

so assume $\lambda_1 > \lambda_2 > 0$

$$W_i^* \text{ for } \lambda_1 : \frac{1}{1+2\lambda_1} X_i$$

$$W_i^* \text{ for } \lambda_2 : \frac{1}{1+2\lambda_2} X_i$$

$$\left(\frac{1}{1+2\lambda_1} X_i\right) \Big/ \left(\frac{1}{1+2\lambda_2} X_i\right) = \frac{1+2\lambda_2}{1+2\lambda_1} < 1$$

thus we can see that if λ goes greater, W_i^* will be smaller, problem solved.

for L1 regularization,

we can see from ca) that

$w_i^* = 0$ when $-\lambda \leq x_i \leq \lambda$ $(\lambda > 0)$

set $\lambda_1 > \lambda_2 > 0$

for λ_1: $w_i^* = 0$ when $-\lambda_1 \leq x_i \leq \lambda_1$

for λ_2: $w_i^* = 0$ when $-\lambda_2 \leq x_i \leq \lambda_2$

we can see $-\lambda_1 < -\lambda_2$, $\lambda_1 > \lambda_2$,

so if λ goes greater, section for $w_i^* = 0$ will be wider, which encourages more 0 and leading to "sparse" solution.

4. **(10 points)** The following represents python code for an algorithm that attempts to perform linear regression. (a) Identify the algorithm. (b) Explain why this algorithm may not converge as implemented below, and identify the line in the algorithm that makes this happen. (c) Suggest a way to fix this algorithm.

```
def optim_alg(init, steps, grad):
  xs = [init]
  for step in steps:
    xs.append(xs[-1] - step * grad(xs[-1]))
  return xs

def linear_reg_grad(X,y,w):
  return X.T.dot(X.dot(w)-y)

input_to_optim_alg = lambda w: linear_reg_grad(X,y,w)
learning_rates = np.arange(start=1, stop=300, step=3)
ws = optim_alg(w0,learning_rates,input_to_optim_alg)
```

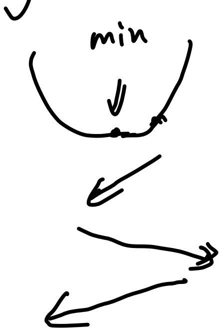
grad $\frac{1}{2}\|Xw-y\|_2^2$

$= X^T(Xw-y)$

(a). it's a gradient descent algorithm (linear regression)

(b) the wrong line is learning_rates = np.arange(start=1, stop=300, step=3)

it makes learning rate keep increasing with iteration, thus algorithm may keep result pass global minimum and fluctuate side to side and never converge because increasing learning rate.



(c) we can fix learning rate, such as:

learning_rates = [0.001] × 100

5. **(10 points)** To combat the COVID-19 pandemic, an enterprising NYU Tandon graduate student decides to build a logistic regression model to predict the conditional likelihood of a person being one of two states – *infected* or *clear* – based on daily forehead temperature measurements over the last 30 days. Fortunately, a dataset of such measurements for a population of 100,000 persons is available. $k=2$

$n=100000$ $d=30$

a. Identify the parameters of the problem (number of samples n, data dimension d, number of classes k.)

b. If X and y denote the arrays that encode the training data points and labels, what are the sizes of X and y?

c. Starting from the definition of conditional likelihood, derive the loss function used to train the model. You can assume the probabilities can be modeled as a sigmoid function.

a. $n = 100000$, $d = 30$, $k = 2$

b. size of X: $n \times (d+1) = 100000 \times 31$

size of $Y = 100000 \times 1$

c. 1 denote infected, 0 denote clear in y $\quad$ f denotes sigmoid

$$P(y_i = k \mid x_i) = f(x_i)^{y_i} (1-f(x_i))^{(1-y_i)}$$

$$P(y \mid X) = \prod_{i=1}^{100000} f(x_i)^{y_i} (1-f(x_i))^{1-y_i}$$

$\Downarrow -\log$

$$L(f) = -\sum_{i=1}^{100000} \{ y_i \log f(x_i) + (1-y_i) \log (1-f(x_i)) \}$$

$$= -\sum_{i=1}^{100000} \{ y_i \log f(z) + (1-y_i) \log (1-f(z)) \}$$

because sigmoid $= f(z) = \frac{1}{1+e^{-z}}$, $z = \langle w, x_i \rangle$

$$L(w) = -\sum_{i=1}^{100000} \{ y_i \log \frac{1}{1+e^{-\langle w, x_i \rangle}} + (1-y_i) \log \frac{e^{-\langle w, x \rangle}}{1+e^{-\langle w, x \rangle}} \}$$

6. **(15 points)** Suppose we are given real-valued scalar data (i.e., $d = 1$) belonging to one of two classes. We are given a set of three data samples with negative labels, $X_- = \{0, 1, -1\}$, and a set of three data samples with positive labels, $X_+ = \{-3, 3, -2\}$. Our goal is to build a classifier for this dataset. We will show that kernel methods are particularly useful in this case.

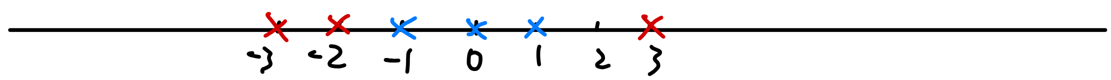
   a. Argue that no perfect linear separator in the original space can exist.

   b. Argue that if the data is mapped via the two-dimensional feature mapping $\phi(u) = (u, u^2)$, then a perfect linear separator exists.

   c. Explicitly draw the maximum-margin linear separator in the new feature space, and mark the closest points nearest to this linear separator.

   d. Calculate the equation of the maximum-margin separator.

a. because we can see from X_- & X_+, points are overlapping into each other's half area on the axis, there is no straight linear decision boundary that can divide them into two areas.

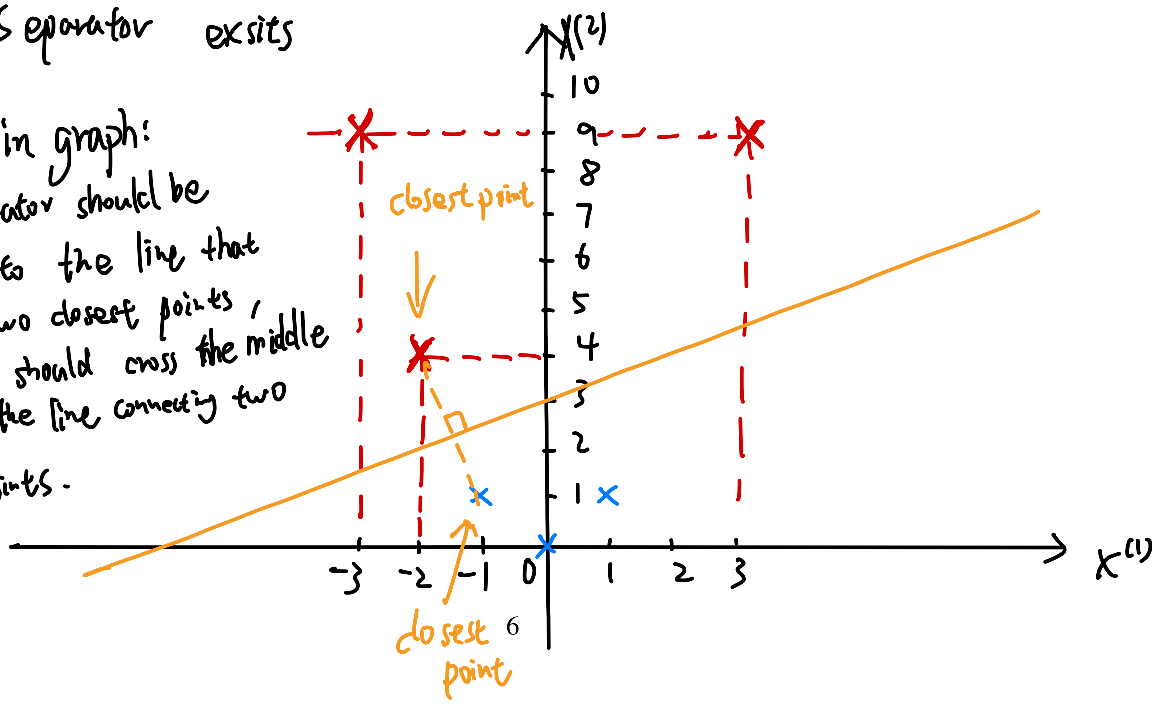


b. X_- after $\phi(u) = \{(0,0), (1,1), (-1,1)\}$

X_+ after $\phi(u) = \{(-3,9), (3,9), (-2,4)\}$

we can see from graph, points of different labels are no longer overlapping, so a perfect linear separator exsits

C. as in graph:
separator should be vertical to the line that connect two closest points, and it should cross the middle point of the line connecting two closest points.



d. call the line segment connecting two closest points C,

the slope of C is $k_C = \frac{(4)-(1)}{(-2)-(-1)} = -3$,

so the slope of separator S (decision boundary) is $k_S = \frac{-1}{k_C} = \frac{1}{3}$

$$x^{(2)} = \frac{1}{3}x^{(1)} + b.$$

The middle point of C is $(-\frac{3}{2}, \frac{5}{2})$,

$$\frac{5}{2} = \frac{1}{3}\times(-\frac{3}{2}) + b \Rightarrow b = \frac{5}{2} + \frac{1}{2} = 3$$

Thus decision boundary is: $x^{(2)} = \frac{1}{3}x^{(1)} + 3 \Leftrightarrow x^2 = \frac{1}{3}x + 3$

($x^{(1)}$ is original x) 3:05 PM

I am not sure above is right, I think maximum-margin separator is produced by SVMs, below is for additional reference: SVM:

$$Loss = \begin{cases} 0, & y\langle w, x_i\rangle \geq 1 \\ 1 - y\langle w, x_i\rangle, & \text{otherwise} \end{cases}$$

if $sign(\langle w, x\rangle)$ dosen't match label:

$$w_{t+1} = w_t + y_i x_i^T \quad (w_0 = 0)$$

equation is $\quad sign(-6 - x_1 + 3x_2)$

$\Rightarrow sign(-6 - u + 3u^2)$

iterations

$\Downarrow$

$$u^2 = \frac{1}{3}u + 2$$

label y_i	x_i	$y\langle w, x_i\rangle$	updated	W_t
				(0,0,0)
+	(1,-3,9)	0	✓	(1,-3,9)
−	(1,0,0)	-1	✓	(0,-3,9)
+	(1,3,9)	72	✗	(0,-3,9)
−	(1,1,1)	-6	✓	(-1,-4,8)
+	(1,-2,4)	39	✗	(-1,-4,8)
−	(1,-1,1)	-11	✓	(-2,-3,7)
+	(1,-3,9)	70	✗	(-2,-3,7)
−	(1,0,0)	2	✗	(-2,-3,7)
+	(1,3,9)	52	✗	(-2,-3,7)
−	(1,1,1)	-2	✓	(-3,-4,6)
+	(1,-2,4)	29	✗	(-3,-4,6)
−	(1,-1,1)	-7	✓	(-4,-3,5)
+	(1,-3,9)	-4+9+45	✗	(-4,-3,5)
−	(1,0,0)	4	✗	(-4,-3,5)
+	(1,3,9)	-4-9+45	✗	(-4,-3,5)
−	(1,1,1)	+2	✗	(-4,-3,5)
+	(1,-2,4)	-4+5+20	✗	(-4,-3,5)
−	(1,-1,1)	-(-4+3+5)	✓	(-5,-2,4)
+	(1,-3,9)	-5+6+36	✗	( )
−	(1,0,0)	5	✗	( )
+	(1,3,9)	-5 $\cancel{6}$+36	✗	( )
−	(1,1,1)	3	✗	( )
+	(1,-2,4)	25	✗	( )

$-$	$(1,-1,1)(-5+2+4)$		✓	$(-5,-2,4)$
$+$	$(1,-3,9)$	24	✗	$(-6,-1,3)$
$-$	$(1,0,0)$	6	✗	$(-6,-1,3)$
$+$	$(1,3,9)$	18	✗	$(-6,-1,3)$
$-$	$(1,1,1)$	4	✗	$(-6,-1,3)$
$+$	$(1,-2,4)$	8	✗	$(-6,-1,3)$
$-$	$(1,-1,1)$	2	✗	$(-6,-1,3)$

3:23 PM