Assignment 0.

COL 870.

By:-
Ankit Garg
[redacted]

Section I : Linear Algebra

1.1. $A \rightarrow m \times n$ matrix

$B \rightarrow n \times m$ matrix

B is obtained by rotating 90° clockwise on paper.

By taking examples I observed that A can be converted into B^T by multiplying a matrix H to A.

Here $H \rightarrow n \times n$ matrix

Such that $H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & & & & 1 & 0 \\ 0 & & & 1 & 0 & 0 \\ 0 & & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}_{n \times n}$

$h_{ij} = 1$ if $i = n-1-j$

0 else.

$HA = B^T$

Now, According to SVD any matrix Q can be

Page No.:
Date:

Written as

$$\underset{m\times n}{Q} = U \Sigma V^T \quad \text{— (1)}$$

U: $m \times m$, Σ: $m \times n$, V^T: $n \times n$

$\Sigma \rightarrow$ diagonal and U & V are orthogonal.

From (1), we can write

$$A = U_A \Sigma_A V_A^T \quad \text{— (2)}$$

Multiply H on both sides

$$HA = H U_A \Sigma_A V_A^T$$

$$B^T = H U_A \Sigma_A V_A^T$$

$$B = (H U_A \Sigma_A V_A^T)^T = V_A \Sigma_A^T U_A^T H^T$$

We get $$B = V_A \Sigma_A^T U_A^T H^T \quad \text{— (3)}$$

From (2) we already have that V_A is orthogonal, Σ_A is diagonal.

Σ_A is diagonal $\Rightarrow$ Σ_A^T is also diagonal.

Now, if we prove $U_A^T H^T$ is orthogonal, then we could say that (3) is SVD decomposition

Now, $H^T = H$ and $H^2 = I$

$$(U_A^T H^T)(U_A^T H^T)^T = U_A^T H^T H U_A$$

$$= U_A^T H^2 U_A = U_A^T U_A$$

$$= I$$

because U is orthogonal.

We can say, (3) is a SVD decomposition

$$U_B \Sigma_B V_B^T = B = U_B \Sigma_B V_B^T$$

where $U_B = V_A$

$$\Sigma_B = \Sigma_A^T \quad \text{---} \quad (4)$$

$$V_B^T = U_A^T H^T$$

From (4), we can say that A and B have same singular values.

1.2. $x \rightarrow$ ~~n×1~~ $m \times 1$ vector

$A \rightarrow m \times n$ matrix.

(a) Prove $\|x\|_\infty \leq \|x\|_2$

$$\|x\|_\infty = \lim_{p \to \infty} \left(\sum_{i=1}^{m} |x_i|^p \right)^{1/p}$$

Let $|x_{max}| = \max_{i=1 \to m} (|x_i|)$

$$\|x\|_\infty = \lim_{p \to \infty} |x_{max}| \left(\sum_{i=1}^{m} \left| \frac{x_i}{x_{max}} \right|^p \right)^{1/p}$$

$\left(\sum_{i=1}^{m} \left| \frac{x_i}{x_{max}} \right|^p \right)$ → This limit is 1 because it is of the form (1^∞)

$$= \lim_{p \to \infty} |x_{max}| \cdot$$

Hence $\|x\|_\infty = |x_{max}|$

Page No.:
Date:

Now $\|x\|_2 = \left(\sum_{i=1}^{m} x_i^2\right)^{1/2} = |x_{max}| * \left[1 + \frac{\sum_{\substack{i=1 \\ i \neq max}}^{m} x_i^2}{x_{max}^2}\right]^{1/2}$

because this sum is to be zero

$\geq |x_{max}|$

Hence. $\|x\|_2 \geq \|x\|_\infty$

or $\|x\|_\infty \leq \|x\|_2$

(b) Prove $\|x\|_2 \leq \sqrt{m}\, \|x\|_\infty$

$\|x\|_2 = \left(\sum_{i=1}^{m} x_i^2\right)^{1/2} \leq \left(\sum_{i=1}^{m} x_{max}^2\right)^{1/2}$.

$\leq (m\, x_{max}^2)^{1/2}$

$\leq \sqrt{m}\, |x_{max}|$

$\leq \sqrt{m}\, \|x\|_\infty$

Hence $\|x\|_2 \leq \sqrt{m}\, \|x\|_\infty$

(c) Prove $\|A\|_\infty \leq \sqrt{n}\, \|A\|_2$; $y \rightarrow n \times 1$ matrix vector
$A \rightarrow m \times n$ matrix

matrix norm $\rightarrow \|A\|_p = \max_{y \neq 0} \frac{\|Ay\|_p}{\|y\|_p}$ — (1)

From (1) we can say for any vector y.

$\|A\|_p \geq \frac{\|Ay\|_p}{\|y\|_p}$

$$\|A\|_p \|y\|_p \geq \|Ay\|_p \quad - (1).$$

$$\frac{\|Ay\|_p}{\|y\|_p} \leq$$

Now from part (a) of this question we can say

$$\|Ay\|_\infty \leq \|Ay\|_2$$

Using 1 we get

$$\leq \|A\|_2 \|y\|_2$$

from part (b) of this question we can say

$$\leq \|A\|_2 \sqrt{n} \|y\|_\infty$$

$$\|Ay\|_\infty \leq \|A\|_2 \sqrt{n} \|y_\infty\|$$

$$\frac{\|Ay\|_\infty}{\|y\|_\infty} \leq \|A\|_2 \sqrt{n}$$

↳ This is true for any arbitrary y. It must also hold for y, for which you get $\max_y \frac{\|Ay\|_\infty}{\|y\|_\infty}$.

Hence.

$$\|A\|_\infty \leq \sqrt{n} \|A\|_2$$

Hence proved.

(d) Prove $\|A\|_2 \leq \sqrt{m}\, \|A\|_\infty$

By def^n:-

$$\|A\|_2 = \max_{\|y\|_2 = 1} \|Ay\|_2.$$

For any arbitrary y, for which $\|y\|_2 = 1$

$\|Ay\|_2$. Now let $A = \begin{bmatrix} \vec{a}_1 \\ \vec{a}_2 \\ \vdots \\ \vec{a}_m \end{bmatrix}$

↳ There are m row vectors

$$\|Ay\|_2 = \left(\sum_{i=1}^{m} (\vec{a}_i \cdot \vec{y})^2 \right)^{1/2}$$

For i^{th} term.

$$(\vec{a}_i \vec{y})^2 = \left(\sum_{j=1}^{n} a_{ij} y_j \right)^2$$

Using Cauchy Swartz Inequality

$$\leq \left(\sum_{j=1}^{n} a_{ij}^2 \right)^1 \left(\sum_{j=1}^{n} y_j^2 \right)$$

$$\leq \sum_{j=1}^{n} a_{ij}^2 \quad [\text{Since } \|y\|_2 = 1].$$

$$\|Ay\|_2 \leq \left(\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \right)^{1/2} \quad — (1).$$

$\leq \left(\sum_i \sum_j |a_{ij}| \right)$ [Because $\|\ \|_2 \leq \|\ \|_1$]

Now, $\|Ay\|_\infty = \lim_{p \to \infty} \left(\sum_{i=1}^{m} (|\vec{a}_i \cdot \vec{y}|)^p \right)^{1/p}$

$\|A\|_\infty = \max_{\|x\|_\infty = 1} \|Ax\|_\infty$ = maximum absolute row sum = k (let say)

Page No.:
Date:

We will use the fact that $\|A\|_\infty$ = maximum absolute row sum

$$\sqrt{m}\,\|A\|_\infty = \sqrt{m\,\|A\|_\infty^2} = \sqrt{m\left(\sum_{j=1}^{n}|a_{kj}|\right)^2}$$

Let k be the row corresponding to maximum row sum

$$= \sqrt{m\left(\sum_{j=1}^{n}|a_{kj}|\right)^2} \geq \sqrt{\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|a_{ij}|\right)^2}$$

Now square of sum of positive values is greater than sum of square of those values.

$$\geq \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} \quad -(2)$$

Combining (1) & (2) we can say

$$\|Ay\|_2 \leq \sqrt{m}\,\|A\|_\infty.$$

↳ true for any arbitary y, must be true for the one $\sup_{\|y\|_2=1}\|Ay\|_2$

$$\|A\|_2 \leq \sqrt{m}\,\|A\|_\infty$$

Hence proved

Page No.:
Date:

Section II : Subgradients.

(a) $f(x) = \max_{i=1,\dots m} (a_i^T x + b_i)$

Let n^{th} line be the one, for which.

$$\max_{i=1\dots m} (a_i^T x + b_i) = a_n^T x + b_n \quad -(1)$$

If there is only one curve for which (1) is satisfied

the **Subgradient of $f(x) = a_n$**

If there are more than one curve, then the curve is non differentiable. We can choose any value between the left and right derivative. that lies between the derivatives around that point

$\nabla f_1 \rightarrow$ be the curve on left.

$\alpha_1 \nabla f_1 + \alpha_2 \nabla f_2 \dots \alpha_k \nabla f_k$; $\alpha_1 + \dots \alpha_k = 1$

These k curves pass through the point and contribute to the derivative in some direction.

(b) $f(x) = \max_{i=1\dots m} |(a_i^T x + b_i)|$

Let n^{th} line be the one for which.

$$\max_i |(a_i^T x + b_i)| = |a_n^T x + b_n|$$

if $a_n^T x + b_n > 0$ then gradient $= a_n$

if $a_n^T x + b_n < 0$ then gradient $= -a_n$

else $\alpha_1 (a_n) + \alpha_2 (-a_n)$; where $\alpha_1 + \alpha_2 = 1$.

If more than one curve passes through this point. Then look at the derivatives at this point, in various directions and take values in-between.

Let $f_1, f_2 \dots f_k$ be the curves.

then $\text{grad} = \alpha_1 \nabla f_1(x) + \dots \alpha_k \nabla f_k(x)$

$\alpha_1 + \alpha_2 + \dots \alpha_k = 1$

(c) $f(x) = \sup_{0 \le t \le 1} p(t)$ where $p(t) = x_1 + x_2 t + \dots x_n t^{n-1}$

Let t_0 be the value for which we get $\sup_{0 \le t \le 1} p(t)$ for our given x

$f(x) = x_1 + x_2 t_0 + \dots x_n t_0^{n-1}$

$\nabla f = [1 \quad t_0 \quad \dots \quad t_0^{n-1}]$.

↳ subgradient

(d) $f(x) = x_{[1]} + x_{[2]} + \dots x_{[k]}$.

$x_{[i]} \rightarrow i^{th}$ largest element.

$\nabla f(x)_{x[i], i \le k} = 1$; $\nabla f(x)_{x[i], i > k} = 0$

Here subgradient $= \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$ → Value is 1, if element i is one of the k largest values. else zero

Date:

(e) $f(x) = \inf_{Ay \leq b} \|x - y\|^2$

For given x, we have to minimize $\|x - y\|^2$, subject to the constraints $Ay \leq b$.

So this is constrained optimization, we can use KKT to solve this

$$\|x - y\|^2 + \lambda (Ay - b)$$

Create dual and then solve. Don't know further!

Page No.:
Date:
youva

Section II : Probability.

1. $E[I \mid X=x] = ?$

$Z \equiv N(0,1)$

$X \equiv N(\mu,1)$

X and Z are independent

and $\Phi(x)$ is the cdf of gaussian distribution.

$$I = \begin{cases} 1 ; & \text{if } Z < X \\ 0 ; & Z \geq X \end{cases}$$

$P(I=1) = P(Z<X)$

$P(I=0) = P(Z \geq X)$

$E(I \mid X=x) = 1 \cdot P(I=1 \mid X=x) + 0 \cdot P(I=0 \mid X=x)$

$= P(Z<x) = \Phi(x)$ [By the definition of cdf]

2. $E(\phi(X)) = ?$

Using the rule of conditional expectation

$= E(E(\phi(X) \mid X=x))$

$= E(E(P(Z<x)))$

→ By the definition of cdf.

2. In the previous part we derived

$$E(I|X=x) = \Phi(x) \quad \text{— (1)}$$

We can write

$$E(I) = 1 \cdot P(Z<X) + 0 \cdot P(Z \geq X)$$

Hence $E(I) = P(Z<X)$

$$E(E(I|X=x)) = P(Z<X)$$

↳ Using the law of conditional expectation

Using (1).

$$E(\Phi(X)) = P(Z<X)$$

Hence proved

3. In the previous part

$$E(\Phi(X)) = P(Z<X) = P(0<X-Z)$$

$$= P(X-Z>0)$$

$X \sim N(\mu,1)$; $Z \sim N(0,1)$ are independent

their diffrence is gaussian $\sim N(\mu, 2)$

Let $Y = X-Z$

$$P(Y>0) = P\left(\frac{Y-\mu}{\sqrt{2}} > \frac{-\mu}{\sqrt{2}}\right) = 1 - \Phi\left(\frac{-\mu}{\sqrt{2}}\right)$$

$$= \Phi\left[\frac{\mu}{\sqrt{2}}\right]$$

Property of cdf function

Page No.:
Date:

Section IV : Machine Learning

4.1 PCA.

1. $\max_{u:\|u\|_2=1} \tilde{V}[u^T x]$

$u \rightarrow p \times 1$ vector.

$X \rightarrow p \times N$ matrix.

Use $Var(x) = E(x^2) - E(x)^2$

$$\tilde{V}[u^T x] = E\left((u^T x_i)^2\right) - E(u^T x_i)^2$$

$$= \sum_{i=1}^{N} \frac{(u^T x_i)^2}{N} - \left[\frac{\sum_{i=1}^{N} u^T x_i}{N}\right]^2$$

$$= \sum_{i=1}^{N} \frac{u^T x_i x_i^T u}{N} - \left[u^T \sum_{i=1}^{N} \frac{x_i}{N}\right]^2$$

$\Rightarrow 0$; since data is zero mean.

$$= u^T \left(\sum_{i=1}^{N} \frac{x_i x_i^T}{N}\right) u$$

On comparing $\Sigma = \sum_{i=1}^{N} \frac{x_i x_i^T}{N}$

2. $$\min \frac{1}{N} \sum_{i=1}^{N} \|x_i - uu^T x_i\|_2^2$$

$$= \min \cdot \frac{1}{N} \sum_{i=1}^{N} (x_i - uu^T x_i)^T (x_i - uu^T x_i)$$

Page No.:
Date:

Now $\mu^T x_i = x_i^T u_i = k$ [scalar quantity]

$$= \min \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu k)^T (x_i - \mu k)$$

$$= \min \frac{1}{N} \sum_{i=1}^{N} \left(x_i^T x_i - \mu^T x_i k - x_i^T \mu k + k^2 \right)$$

$\mu^T x_i = k$

$$= \min \frac{1}{N} \sum_{i=1}^{N} \left(\underbrace{x_i^T x_i}_{\text{independent of } u} - \cancel{\mu^2} - k^2 + \cancel{\mu^2} \right)$$

independent of u. We can remove it

$$= \min \frac{1}{N} \sum_{i=1}^{N} (-k^2) = \max \frac{1}{N} (k^2)$$

$$= \max \frac{1}{N} \sum_{i=1}^{N} (\mu^T x_i x_i^T u)$$

$$= \mu^T \left(\frac{\sum_{i=1}^{N} x_i x_i^T}{N} \right) u$$

$$\sum = \frac{\sum_{i=1}^{N} x_i x_i^T}{N}$$

Hence proved.

4.2.

Bias - Variance Trade-off.

(a) $(x, y) \rightarrow$ new data point

$$J(\theta) = (y - h_\theta(x))^2$$

$h_\theta(x) \rightarrow$ classifier trained over a dataset D.

D is which is dataset is random variable in this question

$$E_D[J(\theta)] = E_D\left[(y - h_\theta(x))^2\right]$$

$$= E_D\left[(y - h_{\bar{\theta}}(x) + h_{\bar{\theta}}(x) - h_\theta(x))^2\right]$$

$$= E_D\left[(y - h_{\bar{\theta}}(x))^2 + (h_{\bar{\theta}}(x) - h_\theta(x))^2 + 2(y - h_{\bar{\theta}}(x))(h_{\bar{\theta}}(x) - h_\theta(x))\right]$$

$$= E_D[(y - h_{\bar{\theta}}(x))^2] + E_D[(h_{\bar{\theta}}(x) - h_\theta(x))^2] + E_D[2(y - h_{\bar{\theta}}(x))(h_{\bar{\theta}}(x) - h_\theta(x))]$$

$$= (y - h_{\bar{\theta}}(x))^2 + E_D[(h_{\bar{\theta}}(x) - h_\theta(x))^2] + 2(y - h_{\bar{\theta}}(x))\underbrace{(h_{\bar{\theta}}(x) - h_{\bar{\theta}}(x))}_{=0}$$

$$= \text{Bias}^2 + \text{Variance}$$

(b) Now for noise

$$J(\theta) = (y - h_\theta(x))^2 = (f(x) + \epsilon - h_\theta(x))^2$$

$$= (f(x) - h_\theta(x))^2 + \epsilon^2 + 2\epsilon(f(x) - h_\theta(x))$$

$$E_D[J(\theta)] = E_D[(f(x) - h_\theta(x))^2] + E_D(\epsilon^2) + E_D(2\epsilon(f(x) - h_\theta(x)))$$

Using results from the previous part

$$= \text{Bias}^2 + \text{Variance} + \text{Var}(\epsilon) - E_D^2(\epsilon) + 2E_D(\epsilon)E(f(x) - h_\theta(x))$$

Since $\epsilon \sim N(0, \sigma^2)$.

$$= \text{Bias}^2 + \text{Variance} + \sigma^2 - 0 + 0.$$

$$\doteq \text{Bias}^2 + \text{Variance} + \sigma^2$$

4.3.

Let's make some observation, then I will describe the 3 parts.

We are doing only one epoch and we take one example at a time.

Therefore we have

$\theta^{(0)}, \theta^{(1)}, \dots\dots \theta^{(M)}$ parameters

[assuming dataset has M examples]

According to perceptron

$$\theta^{(i+1)} = \theta^{(i)} + \alpha \underbrace{\left[y^{(i+1)} - h_\theta(x^{(i+1)}) \right]}_{\text{Note that this term is either } 0, \alpha \text{ or } -\alpha.} x^{(i+1)}$$

Depending upon whether $(i+1)^{th}$ example was misclassified, and what was it actual label.

Let this value be $\gamma^{(i+1)}$ corresponding to $\theta^{(i+1)}$

Then $\theta^{(i+1)} = \theta^{(i)} + \gamma^{(i+1)} x^{(i+1)}$ — (1).

where $\gamma^{i+1} \in \{ \alpha, -\alpha, 0 \}$

From (1)

for any $\forall\ i \in \{0, N\}$ we can say one of the form

$$\theta^{(i)} = \sum_{j=1}^{i} \gamma^{(j)} x^{(j)} + \theta^{(0)}$$

$\theta^{(0)} \to \vec{0}$

So $\theta^{(i)}$ can represented in the basis formed by examples i.e. $\{ x^{(1)}, x^{(2)}, \ldots x^{(m)} \}$

For example

$$\theta^{(i)} = \begin{bmatrix} \alpha \\ -\alpha \\ 0 \\ \vdots \\ \alpha \end{bmatrix}_{m \times 1}$$ → is the representation.

Notice that this representation is independent of the dimension of x or θ.

1. Now, in transformed space, from the above arguments.

$$\theta^{(i)} = \sum_{j=1}^{i} \gamma_j^{(j)} \phi(x^{(j)})$$

$\theta^{(i)}$ can be represented on the basis

$$\text{Basis} = \left[\phi(x^{(1)}), \phi(x^{(2)}) \ldots \phi(x^{(m)}) \right]$$

— (1).

$\theta^{(i)}$ will be represented as $(M \times 1)$ vector.

Note that, the representation is independent of the dimension of $\theta^{(i)}$ or even $\phi(x^{(i)})$.

For example

$$\theta^{(0)} = [0 \; 0 \; 0 \; 0 \; \ldots \; 0]_{1 \times m}$$

$$\theta^{(1)} = [\gamma_1^{(1)} \; 0 \; 0 \; \ldots \; 0]_{1 \times m}$$

$$\theta^{(2)} = [\gamma_1^{(1)} \; \gamma_2^{(2)} \; 0 \; 0 \; \ldots \; 0]_{1 \times m}$$

$$\vdots$$

$$\theta^{(m)} = [\gamma_1^{(1)} \; \gamma_2^{(2)} \; \ldots \; \gamma_m^{(m)}]_{1 \times m}$$

The above representation holds when basis is in Eq (1).

2. Now, we have already represented $\theta^{(i)}$ in our own defined basis.

For prediction on $\theta^{(i+1)}$ example.

$$h_{\theta^{(i)}}(x^{(i+1)}) = g(\theta^{(i)T} \phi(x^{i+1})) \quad \text{— (2)}$$

Now $\theta^{(i)} = \sum_{j=1}^{i} \gamma_{ij} \phi(x^{(j)})$ in our basis.

(2) becomes.

$$= g\left(\sum_{j=1}^{i} \gamma_{ij} \phi^T(x^{(j)}) \phi(x^{i+1})\right)$$

all the γ^1 to γ^i have been calculated till now

So. $\phi^T(x^j)\, \phi(x^{i+1}) = K(x^j, x^{i+1})$

↳ Mercer's Kernel for this basis function

$$= g\left(\sum_{j=1}^{i} \gamma_{ij} K(x^j, x^{i+1})\right)$$

We can compute this.

And hence can make our prediction.

3.

When at $(i+1)$ example, we have already calculated $\gamma^{(1)}$ to $\gamma^{(i)}$, and we need to calculate $\gamma^{(i+1)}$ at this step.

$$\theta^{(i+1)} = \theta^{(i)} + \alpha \left(y^{(i+1)} - g\left(\theta^{(i)T} \phi(x^{(i+1)})\right) \right) \phi(x^{(i+1)})$$

$$\sum_{j=1}^{i+1} \gamma^j \phi(x^{(j)}) = \sum_{j=1}^{i} \gamma^j \phi(x^{(j)}) + \alpha \left(y^{(i+1)} - g(\theta^{(i)T} \phi(x^{(i+1)})) \right) \phi(x^{(i+1)})$$

We say how to compute this in the last part

$$\sum_{j=1}^{i+1} \gamma^j \phi(x^{(j)}) - \sum_{j=1}^{i} \gamma^j \phi(x^{(j)}) = \alpha \left(y^{(i+1)} - g(\theta^{(i)T} \phi(x^{(i+1)})) \right) \phi(x^{(i+1)})$$

$$\gamma^{(i+1)} \phi(x^{(i+1)}) = \alpha \left(y^{(i+1)} - g(\theta^{(i)T} \phi(x^{(i+1)})) \right) \phi(x^{(i+1)})$$

$$\boxed{\gamma^{(i+1)} = \alpha \left(y^{(i+1)} - g(\theta^{(i)T} \phi(x^{(i+1)})) \right)}$$

Update Rule.

Calculation of prediction has been discussed in the previous part.