COL-870: Assignment 0

DEVANG MAHESH
2017EE10093

1) Linear Algebra

1.1 SVD

$A \in \mathbb{R}^{m \times n}$

$B \in \mathbb{R}^{n \times m}$

We know B is obtained by rotating matrix A clockwi

Mathematically, it can be written as:-

$$B_{n \times m} = A^T_{m \times n} \underbrace{\begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ & & \vdots & & \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix}_{m \times m}}_{\triangleq C}$$

$$\boxed{B_{n \times m} = A^T C} \quad \text{——— (1)}$$

Using the SVD of matrix A, we can write

$$A = UDV^T \quad \text{——— (2)}$$

where U and V are orthogonal matrices and D is a diagnol matrix with the diagnol element being the singular value

Putting (2) in (1), we get

$$B = (UDV^T)^T C$$
$$= V D^T U^T C$$
$$= V D^T [U^T C]$$

Here V is an orthogonal matrix, D^T is a diagnol matrix, If we prove that $[U^T C]$ is orthogonal as well, the above expression will be the SVD of matrix B.

$$(U^T C).(U^T C)^T = U^T C . C^T U$$
$$= U^T U \quad [\because \text{It can be shown } C^T C = I]$$
$$= I \quad [\because U \text{ is orthogonal}]$$

$\therefore$ We prove that $B = VD^T(U^T C)$ is the SVD of matrix B. The singular values of B are the elements of diagnol matrix $\underline{\underline{D^T}}$.

As D is a diagnol matrix, the diagnol elements of D are $\boxed{\text{equal}}$ to the diagnol elements of D^T.

Here $\boxed{\begin{aligned} \text{Singular Values of } A &= \text{Diagnol elements of } D \\ &= \text{Diagnol elements of } D^T \\ &= \text{Singular values of } B \end{aligned}}$

1.2 Norms

$A \in \mathbb{R}^{m \times n}$

$x \in \mathbb{R}^{m \times 1} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_m \end{bmatrix}$

(a) We know

$$\|x\|_\infty = \max_{i=1}^{m} |x_i|$$

Let j be the index for which $|x_j|$ maximises $(1 \le j \le m)$

$$\boxed{\therefore \ \|x\|_\infty = |x_j|}$$

$$\|x\|_2 \triangleq \sqrt[2]{\sum_{i=1}^{m} |x_i|^2}$$

$$= \sqrt[2]{|x_j|^2 + c}$$

[Where x_j is the $|x_j|$ maximising index and $c = \sum_{\substack{k=1 \\ k \ne j}}^{m} |x_k|^2$ and $\boxed{c \ge 0}$

$$\|x\|_2 = \sqrt[2]{|x_j|^2 + c} \ge \sqrt[2]{|x_j|^2}$$

$$\ge |x_j|$$

$$\ge \|x\|_\infty$$

$$\Rightarrow \quad \boxed{\|x\|_\infty \le \|x\|_2}$$

(b) $\|x\|_2 \overset{\Delta}{=} \sqrt[2]{\sum_{i=1}^{m} |x_i|^2}$

$\leq \sqrt[2]{\sum_{i=1}^{m} |x_j|^2}$ [where x_j is element of x s.t $|x_j|$ is maximum]

$\leq \sqrt{m}\,|x_j|$

$\leq \sqrt{m}\,\|x\|_\infty$ $\left[\because \|x\|_\infty = \max_{i=1}^{m} |x_i| = \underline{|x_j|}.\right]$

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

(c) $A \in \mathbb{R}^{m \times n}$

$x \in \mathbb{R}^{m \times 1}$

Let $y \in \mathbb{R}^{n \times 1}$, then $Ay \in \mathbb{R}^{m \times 1}$

We know that for a matrix A,

$$\|A\|_p = \max_{\|y\| \neq 0} \frac{\|Ay\|_p}{\|y\|_p}$$

As we take supremum,

$$\|A_p\| \geq \frac{\|Ay\|_p}{\|y\|_p} \quad \forall y \in \mathbb{R}^{n \times 1},\ y \neq 0^{n \times 1}$$

$$\Rightarrow \boxed{\|Ay\|_p \leq \|A_p\| \, \|y\|_p} \quad \forall y \in \mathbb{R}^{n \times 1}, \ y \neq 0^{n \times 1}$$

(1)

$$\|Ay\|_\infty \le \|Ay\|_2 \quad \left[\because Ay \in \mathbb{R}^{m\times 1} \text{ and } \|u\|_\infty < \|u\|_2 \ \forall \text{ vectors}\right]$$

$$\le \|A\|_2 \|y\|_2 \quad [\text{using } (1)]$$

$$\le \|A\|_2 \left(\sqrt{n}\, \|y\|_\infty\right) \left[\because l \in \mathbb{R}^{n\times 1} \ \|l\|_2 \le \sqrt{n}\, \|l\|_\infty\right]$$

$$\le \sqrt{n}\, \|A\|_2 \|y\|_\infty \quad \forall y \in \mathbb{R}^{n\times 1} \ y \neq 0$$

$$\Rightarrow \quad \frac{\|Ay\|_\infty}{\|y\|_\infty} \le \sqrt{n}\, \|A\|_2 \quad \forall y \in \mathbb{R}^{n\times 1}, y \neq 0$$

As ~~i~~the inequality is true for $\forall y \in \mathbb{R}^{n\times 1}, y \neq 0$, it will true for even ~~so~~ max value od <u>L.H.S</u>

$$\max\left(\frac{\|Ay\|_\infty}{\|y\|_\infty}\right) \le \sqrt{n}\, \|A\|_2$$

$$\Rightarrow \quad \boxed{\|A\|_\infty \le \sqrt{n}\, \|A\|_2}$$

(d) $A \in \mathbb{R}^{m \times n}$

$x \in \mathbb{R}^{m \times 1}$

$y \in \mathbb{R}^{n \times 1}$, $Ay \in \underline{\underline{\mathbb{R}^{m \times 1}}}$.

$$\|Ay\|_2 \leq \sqrt{m}\,\|Ay\|_\infty \quad \left[\because \text{ for } l \in \mathbb{R}^{m \times 1} \quad \|l\|_2 \leq \sqrt{m}\,\|l\|_\infty\right]$$

$$\leq \sqrt{m}\,\|A\|_\infty \|y\|_\infty \quad [\text{using (1)}]$$

$$\leq \sqrt{m}\,\|A\|_\infty \|y\|_2 \quad \left[\because \|x\|_\infty < \|x\|_2 \ \forall x \neq 0\right]$$

$$\Rightarrow \quad \frac{\|Ay\|_2}{\|y\|_2} \leq \sqrt{m}\,\|A\|_\infty \quad \forall y \in \mathbb{R}^{n \times 1} \quad y \neq 0^{n \times 1}$$

As it is true ~~for~~ $\forall y$, it will be true for the maximum as well

$$\max\left(\frac{\|Ay\|_2}{\|y\|_2}\right) \leq \sqrt{m}\,\|A\|_\infty$$

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

3) Probability

$$\Phi(x) = \int_{y=-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \quad \text{———} (1)$$

$$\left.\begin{array}{l} X \sim N(\mu, 1) \\ Z \sim N(0, 1) \end{array}\right] \text{Independent}$$

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

1) $E[I / X = x] = 1 . P(Z < X / X = x) + 0 . P(Z \geq X / X = x$

$= P(Z < x)$

$= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$

$= \Phi(x) \quad$ [from (1)]

2) ~~E~~ X is a R.V. $\phi(x)$ is a function of a R.V, hence also a R.V.

$$E[\phi(x)] = \int_{x=-\infty}^{\infty} \phi(x) . P(X=x)\, dx$$

$$= \int_{-\infty}^{\infty} \left(\int_{y=-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \right) . \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2} dx$$

(2)

This part is Continued in
Next Page

Let $P(X, Z)$ be the joint pdf of X and Z

$$P(X=x, Z=z) = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}\right)\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\right)$$

[$\because X, Z$ are independent
$P(X=x, Z=z) = P(X=x).P(Z$

$$P(Z<X) = \int_{x=-\infty}^{\infty}\int_{z=-\infty}^{x} P(X,Z)\, dz . dx$$

$$= \int_{x=-\infty}^{\infty}\int_{z=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} . \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2} dz\, dx$$

$$= \int_{x=-\infty}^{\infty}\left(\int_{z=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz\right) . \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2} dx \quad \text{--- (3)}$$

Comparing (2) and (3), we get

$$\boxed{E[\phi(X)] = P(Z<X)}$$

3) As X is normally distributed

and z is also normally distributed

$(z-x)$ will also be normally distributed

Let $(z-x) \sim N(a, b^2)$

$E(z-x) = E(z) - E(x)$

$\boxed{a = -\mu}$

$Var(z-x) = Var(z) + Var(-x) + 2\cancel{cov}^{0}(z, x)$

($\because$ Independent)

$b^2 = 1 + 1$

$= 2$

$\therefore (z-x) \sim N(-\mu, 2)$

Using transformation on $(z-x)$

$$k = \frac{(z-x)+\mu}{\sqrt{2}}$$

$k \sim N(0,1)$ (standard normal distribution)

$$\boxed{z - x < 0 \implies k < \frac{\mu}{\sqrt{2}}}$$

$E[\phi(x)] = P(z < x) = P\left(k < \frac{\mu}{\sqrt{2}}\right)$

$= \phi\left(\frac{\mu}{\sqrt{2}}\right)$

4) Machine Learning

4.1 PCA

$x_i \in \mathbb{R}^{p\times 1}$ $\quad u \in \mathbb{R}^p$

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

$$= \max_{u:\|u\|_2=1} \frac{1}{N}\sum_{i=1}^{N}(u^T x)^2$$

$$= \max_{u:\|u\|_2=1} \frac{1}{N}\sum_{i=1}^{N}(u^T x).(u^T x)^T \quad \left[\begin{array}{l} \because u^T x \text{ is scalar} \\ u^T x = (u^T x)^T \end{array}\right]$$

$$= \max_{u:\|u\|_2=1} \frac{1}{N}\sum_{i=1}^{N} u^T x x^T u$$

$$= \max_{u:\|u\|_2=1} u^T \underbrace{\left(\frac{1}{N}\sum_{i=1}^{N} x x^T\right)}_{\Sigma} u$$

$$\boxed{= \max_{u:\|u\|_2=1} u^T \Sigma u}$$

$$\boxed{\text{where } \Sigma = \frac{1}{N}\sum_{i=1}^{N} x x^T}$$

As $x \in \mathbb{R}^{p\times 1}$ $\quad x^T \in \mathbb{R}^{1\times p}$

$x x^T \in \mathbb{R}^{1\times p}$ ✓

2)
$$\min_{u:\|u\|_2=1} \frac{1}{N}\sum_{i=1}^{N} \|x_i - uu^T x_i\|_2^2$$

$u^T x_i$ will be a scalar. Let it be equal to $\underline{\underline{k}}$

$$= \min_{u:\|u\|_2=1} \frac{1}{N}\sum_{i=1}^{N} \|x_i - uk\|_2^2$$

$$= \min_{\substack{u:\|u\|_2=1 \\ k}} \frac{1}{N}\sum_{i=1}^{N} (x_i - u.k)^T (x_i - u.k)$$

$$= \min_{\substack{u:\|u\|_2=1 \\ k}} \frac{1}{N}\sum_{i=1}^{N} (x_i^T - k.u^T)(x_i - u.k)$$

$$= \min_{\substack{u:\|u\|_2=1 \\ k}} \frac{1}{N}\sum_{i=1}^{N} \Big(x_i^T x_i - (x_i^T u)k - k(u^T x_i) + k^2.\underbrace{1}_{[\because \|u\|_2^2=1]}\Big)$$

$$(k = u^T x_i = x_i^T u)$$

$$= \min_{\substack{u:\|u\|_2=1 \\ k}} \frac{1}{N}\sum_{i=1}^{N} (x_i^T x_i - k^2 - \cancel{k^2} + \cancel{k^2})$$

$$= \min_{u:\|u\|_2=1} \frac{1}{N}\sum_{i=1}^{N} \Big(\underbrace{x_i^T x_i}_{\text{Independent of } u\text{, can be ignored}} - (x_i^T u)^2\Big)$$

$$= \min_{u:\|u\|_2=1} \frac{1}{N}\sum_{i=1}^{N} \big(-(x_i^T u)^2\big)$$

$$= \max \frac{1}{N}\sum_{i=1}^{N} (x_i^T u)^2$$ [Absorbing negative sign min → max]

$$= \max_{u:\|u\|_2^2=1} \frac{1}{N}\sum_{i=1}^{N} (u^T x_i)(x_i^T u)$$

$$= \max_{u:\|u\|_2^2=1} \frac{1}{N}\sum_{i=1}^{N} u^T(x_i x_i^T)\, u$$

$$= \max_{u:\|u\|_2^2=1} \cancel{\sum_{i=1}^{N}}\; u^T \underbrace{\left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i^T\right)}_{=\Sigma \text{ (as before)}} u$$

$$\boxed{= \max_{u:\|u\|_2^2=1} u^T \Sigma u}$$

where $\Sigma \in \mathbb{R}^{p\times p}$

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

4.2 Bias - Variance Tradeoff

$$h_{\bar{\theta}}(x) = E_D[h_\theta(x)]$$

$$\text{bias} = h_{\bar{\theta}}(x) - y$$

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

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

$$= E_D\left(y^2 + \left(h_\theta(x)\right)^2 - 2y\,h_\theta(x)\right)$$

$$= y^2 + E_D\left(h_\theta(x)^2\right) - 2y\,\underbrace{E\left(h_\theta(x)\right)}_{h_{\bar{\theta}}(x)}$$

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

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

$$= \underbrace{\left(h_{\bar{\theta}}(x) - y\right)^2}_{(\text{bias})^2} + \left[\underbrace{E\left(\left(h_\theta(x)\right)^2\right) - \left(h_{\bar{\theta}}(x)\right)^2}_{\text{variance [from (1)]}}\right]$$

$$= (\text{bias})^2 + \text{variance}$$

$$\text{Variance} = E_D\left(\left(h_{\bar{\theta}}(x) - h_\theta(x)\right)^2\right)$$

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

$$= \left(h_{\bar{\theta}}(x)\right)^2 + E\left(\left(h_\theta(x)\right)^2\right) - 2\,h_{\bar{\theta}}(x) \cdot h_{\bar{\theta}}(x)$$

$$\boxed{\text{Variance} = E\left(\left(h_\theta(x)\right)^2\right) - \left(h_{\bar{\theta}}(x)\right)^2} \longrightarrow (1)$$

$[\because E(h_\theta(x)) = h_{\bar{\theta}}$

Now adding noise.

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

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

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

$$= f(x)^2 + E_D(\varepsilon^2) + E_D((h_\theta(x))^2) - 2\underbrace{\cancel{E_D(\varepsilon)}^{0} . E(h_\theta(x))}_{\because IID}$$

$$+ 2f(x)[\underbrace{\cancel{E_D(\varepsilon)}^{0}}_{\because \sim N(0,1)} - E_D(h_\theta(x))]$$

$$= \underbrace{f(x)^2 + \cancel{E_D(\varepsilon^2)} + E_D(h_\theta(x)^2) - 2f(x) E_D(h_\theta(x))}_{\text{previous expression} = (bias)^2 + variance} + E_D(\varepsilon^2)$$

$$= (bias)^2 + variance + E_D(\varepsilon^2) \left[\begin{array}{l} \varepsilon \sim N(0, \sigma^2) \\ Var = E(\varepsilon^2) - \cancel{E(\varepsilon)}^{0\,2} \end{array} \right.$$

$$= bias^2 + variance + \sigma^2 \leftarrow \left[\sigma^2 = E((\varepsilon)^2) - 0 \right]$$

4.3 Kernelising the Perceptron

1) $$\theta^{(i+1)} = \theta^{(i)} + \alpha\left[y^{(i+1)} - h_{\theta^{i}}(x^{(i+1)})\right] x^{(i+1)}$$

given we start with $\theta^0 = \vec{0}$,
even in higher dimension, θ^0 will be a zero vector

for $i \neq 0$, we can write

$$\theta^{i} = \sum_{k=1}^{i} \beta_k x^{(k)}$$ [where β_k come from solving the expression recursively and given prediction)

In the transformed space,

$$\boxed{\theta^{(i)} = \sum_{k=1}^{i} \beta_k \phi(x^{(k)})}$$

2) $h_{\theta^{(i)}}(x^{(i+1)}) = g\left(\theta^{(i)T} \cdot \phi(x^{(i+1)})\right)$

$$= g\left(\underbrace{\left(\sum_{k=1}^{i} \beta_k \phi(x^{(k)})\right)}_{\text{using the } \theta \text{ expressed in Part 1}} \cdot \phi(x^{(i+1)})\right)$$

$$= g\left(\sum_{k=1}^{i} \beta_k \underbrace{\phi(x^{(k)})\phi(x^{(i+1)})}_{k(x^{(k)},\, x^{(i+1)})}\right)$$

kernel

$$= \boxed{g\left(\sum_{k=1}^{i} \beta_k k(x^{(k)}, x^{(i+1)})\right)}$$

(3) The update is simple

$$\theta^{(i+1)} = \theta^{(i)} + \alpha \left[y^{(i+1)} - \underbrace{h_\theta(x^{(i+1)})}_{\text{Calculated in part 2}} \right] \phi(x^{(i+1)})$$

Comparing with $\theta^{(i+1)} = \sum_{k=1}^{(i+1)} \beta_k \phi(x^k)$

$$= \theta^i + (\beta_{i+1}) \phi(x^{(i+1)})$$

We set $\boxed{\beta_{i+1} = \alpha [y^{(i+1)} - h_\theta(x^{(i+1)})]}$.

$$\boxed{\theta^{i+1} = \theta^i + \beta_{i+1} \phi(x^{(i+1)})}$$

Q2) Subgradients :-

(a) $f(x) = \max_{i=1,2,\ldots m} (a_i^T x + b_i)$

For a given point x, let k be the index of b_k s.t.

$f(x) = a_k^T x + b_k$

Then $\boxed{a_k}$ will be the subgradient at that x.

If we multiple k's s.t. $a_k^T x + b_k = f(x)$, then we can choose any value b/w the highest and lowest a_k

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

For a given point x, let k be the index s.t.

$f(x) = |a_k^T x + b_k|$

If $a_k^T x + b_k \geq 0$ then subgradient $= \boxed{a_k}$

$a_k^T x + b_k < 0$ then subgradient $= \boxed{-a_k}$

(c) $p(t) = x_1 + x_2 t + \cdots + x_n t^{n-1}$

$$f(x) = \sup_{0 \le t \le 1} p(t)$$

For a given x, find $\vec{t^*}$ s.t. $f(x) = p(\vec{t})$

then $\vec{t^*} = \begin{bmatrix} 1 \\ t^* \\ t^{*2} \\ \vdots \\ t^{*n} \end{bmatrix}$ is the sub-gradient at $\underline{x}$.

(d) $f(x) = x_{(1)} + x_{(2)} + \cdots x_{(k)}$

For this case for a given x, we will ~~x~~ using only a ~~few~~ subset of few indices of x.

Let $A = \{ j \}$ s.t. j is among k largest indices.

Then subgradient at x will be g s.t.

$$g_i = \begin{cases} 1 & \text{if } i \in A \\ 0 & \text{o/w} \end{cases}$$

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

For a given x, Let y^* be the point for which $\|x-y\|^2$ is minimum. We move infinitesimally from $x \to x + \Delta x$

$$f(x + \Delta x) = \inf_{Ay^* \le b} \|x + \Delta x - y^*\|^2$$

$$= \inf_{Ay^* \le b} \left(\|x-y^*\|^2 + 2\Delta x \|x - y^*\|_1 + \cancel{\text{smaller terms}}^{0} \right)$$

$$\left[\because \sum (x_i + \Delta x_i - y_i)^2 = \sum (x_i - y_i)^2 + \underbrace{\sum 2\Delta x (x_i - y_i)}_{\text{1 norm}} \right]$$

$$= \inf_{Ay^* \le b} \left(\|x - y^*\|^2 + 2\Delta x \|x - y^*\|_1 \right)$$

We know $\inf(A+B) \ge \inf(A) + \inf(B)$

$$\Rightarrow f(x + \Delta x) \ge \inf(\|x-y^*\|^2) + \inf(2\Delta x \|x - y^*\|)$$

$$\ge \inf(\|x-y^*\|^2) + \underbrace{2(x-y^*)}_{\text{gradient}} \Delta x$$

$$\boxed{\text{gradient} = 2(x - y^*)}$$