Q1 $E(\text{Score}) = 4\cdot P(\text{score}=4) + 3\cdot P(\text{score}=3) + 2\cdot P(\text{score}=2) + 0.$

$4\cdot P(\text{score}=4) = 4\int_0^{\frac{\sqrt{3}}{3}} f(x)\,dx = 4\cdot\frac{1}{3} = \frac{4}{3}$

$3\cdot P(\text{score}=3) = 3\int_{\frac{\sqrt{3}}{3}}^{1} f(x)\,dx = 3\cdot\frac{1}{6} = \frac{1}{2}.$

$2\cdot P(\text{score}=2) = 2\int_1^{\sqrt{3}} f(x)\,dx = 2\cdot\frac{1}{6} = \frac{1}{3}.$

therefore

Expected score of Single shot

$= \frac{4}{3} + \frac{1}{2} + \frac{1}{3} = \frac{8+3+2}{6} = \frac{13}{6}$

Q2 n observations $x_1, x_2, x_3 \cdots x_n$ independent

$P(X=x_1, X=x_2 \cdots X=x_n) = P(X=x_1)P(X=x_2)\cdots P(X=x_n)$

$= \theta e^{-\theta x_1}\,\theta e^{-\theta x_2}\cdots\theta e^{-\theta x_n}$

Want to Maximize this value. $\theta^n e^{-\theta(\sum_{i=1}^n x_i)}$

$$\left[\theta^n e^{-\theta(\sum_{i=1}^n x_i)}\right]' = n\theta^{n-1} e^{-\theta(\sum_{i=1}^n x_i)} + \theta^n e^{-\theta(\sum_{i=1}^n x_i)}\cdot\left(-\sum_{i=1}^n x_i\right) = 0$$

$$n = \theta\sum_{i=1}^n x_i$$

$$\theta = \frac{n}{\sum_{i=1}^n x_i}$$

Q3 a). $X^T A X = [x_1 \cdots x_n] \begin{bmatrix} \sum a_{1i} x_i \\ \sum a_{2i} x_i \\ \vdots \\ \sum a_{ni} x_i \end{bmatrix}$

$$= x_1 \sum_{i=1}^{n} a_{1i} x_i + x_2 \sum_{i=1}^{n} a_{2i} x_i \cdots + x_n \sum_{i=1}^{n} a_{ni} x_i$$

$$= \sum_{j=1}^{n} \sum_{i=1}^{n} a_{ji} x_i x_j$$

b) Symmetric then $a_{ij} = a_{ji}$

plug in $x = e_1, e_2, e_3 \cdots$

$x_i = e_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \rightarrow$ ith

$x_i^T A x_i = a_{ii} > 0$ according to the definition of positive matrix

Q4: a). A positive $X^TAX>0$

$X^T(A+\gamma I)X = X^TAX + X^T\gamma IX = X^TAX + \gamma X^TX$

from definition. $X^TAX>0$ and γX^TX also >0.

then $A+\gamma I$ also positive.

b). V eigenvector and λ the corresponding eigenvalue.

then $AV=\lambda V$.

$V^TAV = V^T\lambda V = \lambda V^TV > 0$. $\Leftarrow V^TAV>0$

and $V^TV>0$ then $\lambda>0$

$\therefore$ all eigenvalue of A greater than zero

c). A positive $\Rightarrow \forall x \in \mathbb{R}^n \quad X^TAX>0$.

then Null space of A is ϕ

Because if $\exists X \; AX=0$ then $X^TAX=0$ ⊗

Null space $=0$ A has n independent column span $\mathbb{R}^n$

thus invertible.

d). A Symmetric then A can be written as $U^T\Lambda U$.

U^T constructed by normalized eigenvectors and Λ is diagonal, eigenvalues.

assume $v_1, v_2, v_3, \dots v_n$ are n normalized eigenvector

Λ diagonal and all entries >0

$\lambda_1, \lambda_2, \lambda_3 \dots \lambda_n$ are n eigenvalues all of them are positive

then $\begin{vmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \lambda_3 & \\ 0 & & & \ddots \; \lambda_n \end{vmatrix} = \begin{vmatrix} \sqrt{\lambda_1} & & & 0 \\ & \sqrt{\lambda_2} & & \\ & & \sqrt{\lambda_3} & \\ 0 & & & \ddots \; \sqrt{\lambda_n} \end{vmatrix}^2 = (\sqrt{\Lambda})^2$

therefore $A = U^T\sqrt{\Lambda}^T\sqrt{\Lambda}U$

$= (\sqrt{\Lambda}U)^T(\sqrt{\Lambda}U)$.

Q5.

a). $x, a \in \mathbb{R}^n$ get $\frac{\partial (x^T a)}{\partial x}$

$f(x) = x^T a = x_1 a_1 + x_2 a_2 + x_3 a_3 + x_4 a_4 \cdots$

$$\frac{\partial (x^T a)}{\partial x} = \begin{vmatrix} \frac{\partial f(x)}{x_1} \\ \frac{\partial (Ax)}{x_2} \\ \vdots \end{vmatrix}^T = a^T$$

$$(a^T)^T = a$$

b). $A \in \mathbb{R}^{n \times n}$ $x \in \mathbb{R}^n$ $\frac{\partial (x^T (Ax))}{\partial x}$

$$Ax = \begin{vmatrix} \sum A_{1i} x_i \\ \sum A_{2i} x_i \\ \sum A_{3i} x_i \\ \vdots \end{vmatrix} \quad f(x) = x^T Ax = x_1 \sum A_{1i} x_i + x_2 \sum A_{2i} x_i \cdots + x_n \sum A_{ni} x_i$$

$$\frac{\partial f(x)}{\partial x_k} = x_1 A_{1k} + x_2 A_{2k} + \cdots \cancel{x_k} + \cdots x_n A_{nk}$$

$$\cancel{\sum A_{ki}} \sum_{i \neq k} A_{ki} x_i + 2 A_{kk} x_k$$

$$= \sum_{i=1}^{n} A_{ki} x_i + \sum_{i=1}^{n} A_{ik} x_i$$

$$\frac{\partial f(x)}{\partial x} = \left| \left(\sum_{i=1}^{n} A_{1i} x_i + \sum_{i=1}^{n} A_{ik} x_i \right), \left(\sum_{i=1}^{n} A_{2i} x_i + \sum_{i=1}^{n} A_{i2} x_2 \right) \cdots \right|$$

$$= x^T A^T + x^T A$$

$$(x^T A^T + x^T A)^T = A^T x + A x$$

c) $(XA)_{ij} = \sum_{k=1}^{n} X_{ik}A_{kj}$ $\quad \text{Trace}(XA) = \sum_{i=1}^{n}(XA)_{ii}$

$$= \sum_{i=1}^{n}\sum_{k=1}^{n} X_{ik}A_{ki}$$

$$\frac{\partial(XA)}{\partial X_{ab}} = A_{ba} \qquad \frac{\partial(XA)}{\partial X} = \begin{vmatrix} A_{11} & A_{21} & A_{31} & \cdots \\ A_{12} & & & \\ \vdots & & & \end{vmatrix} = A^T$$

d) want to proof

$$\sqrt{\sum X_i^2} \le \sum |X_i| \le \sqrt{n}\sqrt{\sum X_i^2}$$

$$\left(\sqrt{\sum X_i^2}\right)^2 \le \left(\sum |X_i|\right)^2$$

$$\sum X_i^2 \le \sum |X_i|^2 + 2\sum_{i=1}^{n}\sum_{j=1} |X_i|\cdot|X_j| \quad i \neq j$$

(the double sum: greater or equal to zero)

therefore

$$\|X\|_2 \le \|X\|_1$$

want to show $\sum |X_i| \le \sqrt{n}\sqrt{\sum |X_i|^2}$,

$$\left(\sum |X_i|\cdot 1\right)^2 \le \sum |X_i|^2 \sum (1^2) \qquad \text{apply Cauchy inequality}$$

$$\sum a_i^2 \sum b_i^2 \ge \left(\sum a_i b_i\right)^2$$

$$\sum |X_i|^2 \sum 1^2 \ge \left(\sum |X_i|\cdot 1\right)^2. \qquad \text{Done}$$

e) Maximize $g(x) = X^T Z \mid \|x\|_1 \le 1$

$$X^T Z = X_1 Z_1 + X_2 Z_2 + \cdots X_n Z_n \le \text{Max}(|Z_i|)(Z_1 + Z_2 + \cdots Z_n)$$

$$\le \text{Max}(|Z_i|)\cdot 1$$

Q6 a). $P(X|W_i) = N(\mu_i, \sigma^2)(X)$. $P(W_i|X) = \frac{P(X|W_i)P(W_i)}{P(X)}$

$P(X) = P(X|W_1)P(W_1) + P(X|W_2)P(W_2)$
$= \frac{1}{2}N(\mu_1, \sigma^2)(X) + N(\mu_2, \sigma^2)(X)\frac{1}{2}$

$= \frac{\frac{1}{2} \cdot N(\mu_1, \sigma^2)(X)}{\frac{1}{2}\left(N(\mu_1, \sigma^2)(X) + N(\mu_2, \sigma^2)(X)\right)}$

$P(W_1|X) = \frac{\frac{1}{2}N(\mu_1, \sigma^2)(X)}{\frac{1}{2}[N(\mu_1, \sigma^2) + N(\mu_2, \sigma^2)](X)}$ $= P(W_2|X) = \frac{\frac{1}{2}N(\mu_2, \sigma^2)(X)}{\frac{1}{2}[N(\mu_1, \sigma^2) + N(\mu_2, \sigma^2)](X)}$

$\mu_1 \le \mu_2$

$N(\mu_1, \sigma^2)(X) = N(\mu_2, \sigma^2)(X)$

$\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(X-\mu_1)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(X-\mu_2)^2}{2\sigma^2}}$

$(X-\mu_1)^2 = (X-\mu_2)^2$ $X_0 = \frac{\mu_1+\mu_2}{2}$

Boundary $X_0 = \frac{\mu_1+\mu_2}{2}$. when $X < X_0$ W_2

else W_1

b) $P(\text{error}) = P(\text{misclassified as } w_2 | w_1) P(w_1) + P(\text{misclassified as } w_1 | w_2) P(w_2)$

$x_c = \frac{u_1 + u_2}{2}$

A B

$$A = \frac{1}{2} \times \int_{x_c}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-u_1)^2}{2\sigma^2}} dx$$

$$B = \frac{1}{2} \cdot \int_{-\infty}^{x_c} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-u_2)^2}{2\sigma^2}} dx$$

$z_1 = \left(\frac{x-u_1}{\sigma}\right)$ replace x in A as z_1

$$A \Rightarrow \frac{1}{2} \int_{\frac{x_c-u_1}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

$z_2 = \left(\frac{x-u_2}{\sigma}\right)$ replace x in B as z_2

$$B \Rightarrow \frac{1}{2} \int_{-\infty}^{\frac{x-u_2}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

~~$A = \frac{1}{2} \int \frac{(x_c - u_1)}{\sigma}$~~

$A = B = P(\text{error})$

$$B = \frac{1}{2} \int_{\frac{(x-u_2)}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

$= A$

since $e^{-z^2/2}$ is even function.

therefore

$$P(\text{error}) = \int_{\frac{u_2-u_1}{2\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

Q7. a). $f(w) = \frac{1}{2}\|Xw-y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$.

$= \frac{1}{2}(Xw-y)^T(Xw-y) + \frac{\lambda}{2}w^Tw$

$\frac{\partial f(w)}{w} \cdot (Xw-y) = k_{(w)}$ $\quad \frac{\partial k^Tk}{\partial k}\frac{\partial k}{\partial w} = 2k^T \cdot X$

therefore $\frac{\partial f(w)}{w} = \frac{1}{2}\cdot 2(Xw-y)^T\cdot X + \frac{\lambda}{2}\cdot 2w^T$

$= (Xw-y)^T X + \lambda w^T$

b). want derivative $=0$ $= \lambda w^T + (Xw-y)^T X$

$= \lambda w^T + w^TX^TX - y^TX$

$y^TX = w^T(\lambda I + X^TX)$.

$X^Ty = (\lambda I + X^TX)^Tw$.

$X^Ty = (\lambda I + X^TX)\cdot w$.

Since $\lambda > 0$ therefore $(\lambda I + X^TX)$ to ~~some space~~ invertable

$\Rightarrow (\lambda I + X^TX)^{-1}X^Ty = w$.

c).

the first term

$\frac{1}{2}\|Xw-y\|_2^2 = \frac{1}{2}(Xw-y)^T(Xw-y)$ $\quad w = w_n + X^Ta$.

$Xw = Xw_n + XX^Ta$ Since $w_n \in Null(X)$

$\Downarrow$

$Xw = XX^Ta$.

therefore $\frac{1}{2}\|Xw-y\|_2^2 = \frac{1}{2}(XX^Ta-y)^T(XX^Ta-y)$

only depends on a.

d). $\frac{\lambda}{2}w^Tw$ $\quad w = w_n + X^Ta$

$\frac{\lambda}{2}(w_n^T + a^TX)(w_n + X^Ta)$

$= \frac{\lambda}{2}(w_n^Tw_n + w_n^TX^Ta + a^TXw_n + a^TXX^Ta)$.

$w_n^TX^Ta = w_n \cdot X^Ta$

$a^TXw_n = (Xa)^T\cdot w_n$

$= X^Ta \cdot w_n = 0$

therefore $\frac{\lambda}{2}\|W\|_2^2 = \frac{\lambda}{2}[\underbrace{W_n^T W_n}_{\geq 0} + \underbrace{(\vec{\pi}^T a)^T(\vec{\pi}^T a)}_{\geq 0}]$

if a fixed $\frac{\lambda}{2}\|W\|_2^2 \geq \frac{\lambda}{2}(\vec{\pi}^T a)^T(\vec{\pi}^T a)$.

equal when $W_n = 0$

e. minimize $\frac{1}{2}\|X\vec{\pi}^T a - y\|_2^2 + \frac{\lambda}{2}\|\vec{\pi}^T a\|_2^2$ ⇒ depend only on a since from c and d. We find that

f. the first term of (1) not depend on W_n and the second theron of (1) is minimized (over W_n) when $W_n = 0$.

g. then take derivative of $\frac{1}{2}\|XX^T a - y\|_2^2 + \frac{\lambda}{2}\|X^T a\|_2^2$ (over a)

$$\frac{d}{da}\left[\frac{1}{2}(a^T XX^T - y^T)(XX^T a - y) + \frac{\lambda}{2}(a^T X)(X^T a)\right]$$

$$= \frac{d}{da}\left[\frac{1}{2}(a^T XX^T XX^T a - a^T XX^T y - y^T XX^T a + y^T y) + \frac{\lambda}{2}a^T XX^T a\right]$$

$$= \frac{1}{2}(a^T XX^T XX^T + a^T XX^T XX^T - y^T X^T X - y^T XX^T) + \frac{\lambda}{2}a^T XX^T$$

$XX^T = A$ (A symmetric)

then $a^T A^2 - y^T A + \lambda a^T A = 0$

$a^T(A^2 + \lambda A) = y^T A$ (A invertible)

$(A^T A + \lambda A)a = Ay$

$A^{-1}(AA^T - A\lambda)a = A^{-1}Ay$

$\Rightarrow (A^T + \lambda I)a = y$ $A^T = X\vec{\pi}^T X^T$ semipositive.

~~⇒ proved~~

$V^T X\vec{\pi}^T XV = \|XV\|_2^2 \geq 0$.

By Q4. gave a.

$(A^T + \lambda I)$ $\lambda > 0$ invertible

Since $A^T = X\vec{\pi}^T X^T$ semipositive therefore $a = (A^T + \lambda I)^{-1}y$

$= (X\vec{\pi}^T X^T + \lambda I)^{-1}y$

#7

g7. find a^*: $(XX^T+\lambda I)^{-1}y$.

$X = n \times d$.

get XX^T takes $n \times n \times d$.

inverse takes n^3.

$y = n \times 1$

then takes n^2.

$O(n^3 + n^2 + n^2 d)$.

$= O(n^3 + n^2 d)$ if n, d huge.

find w^* $(\lambda I - X^T X)^{-1} X^T y$

X^T $d \times n$

$X^T X$ takes $d^2 n$

inverse takes d^3.

$X^T y$ takes $d \times n$

multiply $(\lambda I - X^T X)^{-1}$ ($d \times d$) and $X^T y$ ($d \times 1$)

takes d^2

therefore $O(d^2 + nd + d^3 + d^2 n)$ if d, n large

$= O(d^3 + d^2 n)$

Compare

a^*: $O(n^3 + n^2 d)$ and w^*: $O(d^3 + d^2 n)$

if $n \geq d$ should choose w^*

if $n \ll d$ a^*

Q8 a) W^* 使

$$\sum_{0}^{n} (x_i^T W - y_i^T)(W^T x_i - y_i) + \lambda \underbrace{\sum_{j=1}^{d}\sum_{i=1}^{2} W_{ij}}_{= tr(W^T W).}$$

$$= \sum_{0}^{n} [x_i^T W W^T x_i + x_i^T W y_i - y_i^T W^T x_i + y_i^T y_i] + tr(W^T W)$$

take derivative (over W). $\Rightarrow$

$$= \left\{ \sum_{i=0}^{n} [\cancel{2WXX^T} + 2W^T x_i x_i^T - 2(y_i x_i^T)] + 2\lambda W^T \right\}^T$$

$$= \left\{ 2W^T \sum_{i=1}^{n} x_i x_i^T - 2\sum \text{(}y_i\text{)} x_i^T + 2\lambda W^T \right\}^T$$

$\sum x_i x_i^T$: $d\times 1$ $1\times d$ $= d\times d$ 矩阵相加 $\underbrace{\sum x_i y_i}$

$$= XX^T \quad X = [x_1, x_2, x_3, \cdots]$$

$$\sum y_i x_i^T = YX \qquad Y = [y_1\ y_2\ y_3 \cdots y_n]_{k\times n}$$

therefore gradient $= \left\{ 2W^T XX^T - 2YX^T + 2\lambda W^T \right\}^T$

Want gradient $= 0$ then $2W^T XX^T + 2\lambda W^T = 2YX^T$

$$W^T(XX^T + \lambda I) = YX^T$$

if $\lambda > 0$ $(XX^T + \lambda I)$ positive

thus invertable.

$(XX^T + \lambda I)$ ~~W~~ $W = XY^T$

$$W = (XX^T + \lambda I)^{-1} \cdot XY^T$$

$$= (XX^T + \lambda I)^{-1} \cdot XY^T$$

b) > python code

c) > python code

d) to compute $(XX^T + \lambda I)^{-1} \cdot X^T$ (dxd, dxn, nxk)

X: dxn

XX^T takes d^2n

inverse takes d^3.

X^T takes dnk

~~therefore $O(d^2n + d^3 + dnk)$.~~

and $(XX^T + \lambda I)^{-1} \cdot X^T$

takes d^2k

therefore $O(d^2k + dnk + d^3 + d^2n)$

e) $W = [w_1, w_2, \dots w_k]$. X: dxn

original classifier $\sum_{i=0}^{n} \|W^T x_i - y_i\|_2^2 + \lambda \|W\|_F^2$.

$$W^T x_i = \begin{vmatrix} w_1 \cdot x_i \\ w_2 \cdot x_i \\ \vdots \\ w_k \cdot x_i \end{vmatrix}$$

$\|W\|_F^2 = \|w_1\|_2^2 + \|w_2\|_2^2 + \|w_3\|_2^2 + \dots \|w_k\|^2$

$X = [x_1, x_2, \dots x_n]$

So we can choose w_i orthogonal to all x_i

then $X^T w_i = 0$.

and we can take W apart $W_\perp$ and $W_{||}$

$W_\perp$ is consisted by w_i $X^T w_i = 0$

$W_{||}$ is consisted by w_i Column Space of X.

then $W = W_\perp + XA$ A: nxk matrix

We then plug this w in the equation.

$w = w_\perp + XA$

the first term

$$\sum_{i=0}^{n} \|(w_\perp^T + A^T X^T) x_i - y_i\|_2^2 = \sum_{i=0}^{n} \|A^T X^T x_i - y_i\|_2^2 \quad (w_\perp^T x_i = 0)$$

$$= \sum_{i=1}^{n} \left(x_i^T X A A^T X^T x_i - 2 x_i^T X A y_i + y_i^T y_i \right).$$

$$= \text{trace}(X^T X A A^T X^T X) - 2\,\text{trace}[X^T X A^T] + \text{trace}(Y^T Y)$$

not about A

tail: $\|w\|_F^2 = \|w_\perp + XA\|_F^2 = \text{trace}(w_\perp + XA)^T (w_\perp + XA)$

$= \text{trace}(A^T X^T X A)$

$(w_\perp^T + A^T X^T)(w_\perp + XA) = w_\perp^T w_\perp + A^T X^T w_\perp + w_\perp^T w_\perp + A^T X^T X A$ (the $A^T X^T w_\perp$ term is 0)

the first term does not depend on $w_\perp$ ....

and for the second term. when $w_\perp = 0$ has Minimum value

We take derivative

$$\frac{\partial}{\partial A} \left| \text{trace}(X^T X A A^T X^T X) - 2\,\text{trace}[X^T X A^T] + \lambda\,\text{trace}(A^T X^T X A) \right| = 0$$

then $\lambda \cdot 2 A^T X^T X + 2 A^T X^T X X^T X = 2 Y^T X^T X$ Assume $X^T X$ invertiable

then $\lambda \cdot I A^T + I A^T X^T X = Y^T$.

$A^T (\lambda I + X^T X) = Y^T$. Since $(\lambda I + X^T X)$ invertiable

$A = (X^T X + \lambda I)^{-1} Y^T$.

Complexity: $X^T X$ takes nd^2 inverse n^3 times Y^T $n^2 k$.

$O(n^2 d + n^3 + n^2 k)$. if $n \ll d$ compute A instead of w.

#2.

a) $X = AZ + b$ $\quad Z = \begin{vmatrix} N(0,1) \\ N(0,1) \\ N(0,1) \\ \vdots \end{vmatrix}$

$\Sigma_{ij} = E[(X_i - b_i)(X_j - b_j)]$ since $b = \bar{X}$

$= E\left[\left(\sum_{k=1}^{n} A_{ik} Z_k\right)\left(\sum_{p=1}^{n} A_{jp} Z_p\right)\right]$

$= \sum A_{ik} \sum A_{jp} E(Z_k Z_p)$

$\hookrightarrow = \delta_{kj} = \begin{cases} 0 & k \neq j \\ 1 & k = j \end{cases}$

$\Sigma_X = AA^T$ invertable $\quad = (AA^T)_{ij}$

iff AA^T invertiable

$\det(AA^T) = \det(A)\det(A^T) \Leftrightarrow A$ invertiable

Σ_X^{-1} not exist $\Leftrightarrow$ A not invertiable

$X = AZ + b \Rightarrow X_1\ X_2\ X_3 \cdots X_n$ Linearly dependant.

We can get A from X, $\sqrt{\Sigma_X} = A$ Σ_X semipositive therefore $\sqrt{\Sigma_X}$ exist

then we do Row Reduce of A get $A' = \begin{vmatrix} \cdots \\ \cdots \\ \text{n Row all zero} \\ 0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0 \end{vmatrix}$

then we cut the Last n terms in X

convert X to X'. Since the Last n are Linearly dependent term

we do not Lose Information.

b). Σ symmetric and positive defined.

Σ^{-1} also symmetric and positive defined.

$\Sigma^{-1} = U \Lambda U^T$ $\quad \Lambda$ positive $\Lambda = \sqrt{\Lambda}\cdot\sqrt{\Lambda}$

$X^T U \sqrt{\Lambda}\sqrt{\Lambda} U^T X = \|\sqrt{\Lambda} U^T X\|_2^2$

$A = \sqrt{\Lambda} U^T$ $\quad \Lambda$ = diagonal Matrix

after diagonalizing Σ^{-1}

c). Sample from

convert X Back to ~~Encoding~~ ~~independent~~ $Z = \begin{vmatrix} N(0,1) \\ N(0,1) \\ N(0,1) \\ \vdots \end{vmatrix}$

$\Sigma = VV^T$ $X = VZ + b \rightarrow b = 0$

$\Sigma^{-1} = V^{T-1}V^{-1}$ $= VZ$

$X^T\Sigma^{-1}X = X^TV^{T-1}V^{-1}X$ $b = VZ + b$. X, Z are Samples from X, Z

$= Z^TV^TV^{T-1}V^{-1}VZ = Z^TZ = \|Ax\|_2^2$

change Sample space.

d). $\|Ax\|_2^2 = X^T\Sigma^{-1}X = X^TU\Lambda U^TX$. U^T change ~~base~~ base

therefore $\|U^TX\|_2 = 1$

Max = Maximum entry of Λ

Min = ~~Min entry~~ Minimum entry of Λ

entries of Λ are eigenvalues of Σ^{-1}

if $X_i \perp X_j$ then Σ diagonal Matrix entries correspond to variance of X_i

Maximum of $\|Ax\|_2^2 = \frac{1}{a}$ a = Minimum variance among X_i

Minimum of $\|Ax\|_2^2 = \frac{1}{b}$ b = Maximum variance among X_i

choose X_i makes $\|Ax\|_2^2 = \frac{1}{b}$ b = Max variance

#4. a) code

and use form solution from HW1.

$W = (\lambda I + X^T X)^{-1} X^T y$

b) $W_{n+1} = W_n - \alpha(2XX^T W_n - 2XY^T + 2\lambda W_n)$.

c) $W_{t+1} = W_t - \alpha(2x_i x_i^T W_t - 2x_i y_i^T + 2\lambda W_t)$.

In order to help the iterative algorithm converge. α should decrease generally. can use exp or $(1 - \frac{i}{\#\text{iteration}})$ ---

d) gradient descent : α = default value reg = 0.1 (default reg).

Stochastic gradient descent : α = default reg = 0.1

and $\alpha_i = \alpha * (1 - \frac{i}{\#\text{iteration}})$

From the Graph we can see that the error rate of gradient descent is always decreasing and the curve is very smooth. Because we compute the real derivative and with a proper α the update direction is always correct.

The graph of stochastic gradient descent is a little different. Although it trends to decrease, we can see a lot of noise during the process Because we randomly choose vector to compute the "expected" value of derivative, some times the update direction is not correct.

e) My score is 0.94460

#5.

a). $A(A^TA+\mu I) = (AA^T+\mu I)A \Leftarrow AA^TA + \mu IA = AA^TA + A\mu I$.

and ↓ invertiable ← ⇐ Both positive

$(AA^T+\mu I)^{-1}A(A^TA+\mu I)(A^TA+\mu I)^{-1} = (AA^T+\mu I)^{-1}(AA^T+\mu I)A(A^TA+\mu I)^{-1}$

$\Downarrow$

$(AA^T+\mu I)^{-1}A = A(A^TA+\mu I)^{-1}$ □

b). From Hw1 we have $w_* = w = (\lambda I + X^TX)^{-1}X^Ty$

from part a we write as $= X^T(XX^T+\mu I)^{-1}y$

$= X^Tk \qquad k = (XX^T+\lambda I)^{-1}y$

$= \sum_{i=1}^{n} a_i x_i$

$a_i = [(XX^T+\mu I)^{-1}y]_i$

~~and which can also indicate that a_i is unique~~

~~Since $(XX^T+\mu I)^{-1}$ is unique~~

w_* is unique since $w_* = (\lambda I + X^TX)^{-1}X^Ty$

and $(\lambda I + X^TX)$ invertiable then w_* uniquely defined.

c). $w = w_* + w_\perp$ $w_* = \sum a_i x_i$ and $w_\perp$ from Null(X)

$X_i w_\perp = 0$. $X_i w_\perp = 0$.

$\frac{1}{n}\sum_{i=1}^{n} Loss(w^Tx_i, y_i) + \mu\|w\|_2^2$ ← B

A → takes parameter w^Tx_i and y_i

$w^Tx_i = (w_*^T + w_\perp^T)x_i = w_*^Tx_i$

therefore w^Tx_i's value not depends on $w_\perp$

the value of part A not depends on $w_\perp$

c) $\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \text{loss}(w^T x_i, y_i) + \mu \|w\|_2^2$

Loss function convex for $\forall k_1 < k_2$

$\lambda \text{Loss}(k_1, y_i) + (1-\lambda) \text{Loss}(k_2, y_i) \geq \text{loss}(\lambda k_1 + (1-\lambda) k_2, y_i)$.

For Part B $= \mu \|w\|_2^2 = \mu (w_*^T + w_\perp^T)(w_* + w_\perp)$

$= \mu (w_*^T w_* + w_\perp^T w_\perp)$

if fix w_*, has minimum value when $w_\perp = 0$

therefore to ensure $\frac{1}{n} \sum_{i=1}^{n} \text{Loss}(w^T x_i, y_i) + \mu \|w\|_2^2$ has minimum value, $w_\perp$ should be zero

$w_* = \sum_{i=1}^{n} x_i a_i$

We never use the condition "Loss function is convex" therefore the proof above is a general proof.

Loss not convex, optimal solution still has the form

$w_* = \sum_{i=1}^{n} a_i x_i$

#6. P(see those data)

$\prod P(y_i|x_i)$

$= \prod N(w_0+w_1 x_i, \sigma^2)(y_i)$

$= \prod \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i-(w_0+w_1 x_i))^2}{2\sigma^2}}$

$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\frac{[(y_1-(w_0+w_1 x_1))^2+(y_2-(w_0+w_1 x_2))^2-\cdots]}{2\sigma^2}}$

Want to Maximize $\prod P(y_i|x_i)$

then Minimize $\sum_{i=1}^{n}(y_i-(w_0+w_1 x_i))^2 = \phi(w_0, w_1)$

$\phi_{w_0} = \sum -2(y_i-(w_0+w_1 x_i)) = 0$

$\phi_{w_1} = \sum -2(y_i-(w_0+w_1 x_i))(-x_i) = 0$

$\frac{\phi_{w_0}}{2} = \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} w_0 + w_1 \sum_{i=1}^{n} x_i = 0.$

$+\sum_{i=1}^{n} w_0 = +nw_0$

$w_0 = \bar{y} - w\bar{x}$

plug w_0.

~~$\sum y_i - nw_0 - w_1 \sum x_i = 0$~~

~~$\bar{y} = \bar{y}$~~

$\phi_{w_1} = 0$

$= -\sum y_i x_i + w_0 \sum x_i + w_1 \sum x_i^2 = 0.$

$-\sum x_i y_i + \bar{y}\sum x_i - w_1 \bar{x}\sum x_i + w_1 \sum x_i^2 = 0$

$w_1 = \frac{\sum_i x_i y_i - \bar{y}\sum x_i}{\sum x_i^2 - \bar{x}\sum x_i} = \frac{\sum_i x_i y_i - \bar{y}\bar{x}\cdot n}{\sum_i x_i^2 - n\bar{x}^2}$

#7. a). (0,1) (0,-1) (-1,0) (1,0) all $\frac{1}{4}$

$Cov(X,Y) = E(X \cdot Y) = 0$ uncorrelated

But $P(X=0, Y=0) = 0 \neq P(X=0) \cdot P(Y=0)$ not independent.

b). Consider X, Y.

$P(X=1) = \frac{1}{2}$ $P(Y=0) = \frac{1}{2}$ $P(X=1, Y=0) = \frac{1}{4} = P(X=1) \cdot P(Y=0)$

for all other cases $P(X,Y) = P(x) P(y)$

for XZ YZ pair due to ~~symmetry~~. symmetric properity they are all ~~mutually~~ pairwise independent

But they are not mutually independent

$P(X=1, Y=1, Z=1) = 0 \neq \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}$