CSCI 5521 – Introduction to Machine Learning

Homework 0

goelx029 | 5138568

Index

CSCI 5521
Homework 0

1

Name - Saksham Goel
ID - goelx029 | 5138568

Q1

1) Standard Linear Regression

$$\underset{\omega}{\text{minimize}} \; \phi(\omega) = \|X\omega - y\|_2^2 = \langle X\omega - y, X\omega - y \rangle$$

to find the $\min_{\omega} \phi(\omega)$ we can find the partial derivation of $\phi(\omega)$ with respect to each ω_j for all $j \leq n$ s.t. ω_j is the j^{th} element of ω vector.

$$\Rightarrow \frac{\delta}{\delta \omega_j} \phi(\omega) = \frac{\delta}{\delta \omega_j} \|X\omega - y\|_2^2 = \frac{\delta}{\delta \omega_j} \langle X\omega - y, X\omega - y \rangle$$

$$\Rightarrow \frac{\delta}{\delta \omega_j} \phi(\omega) = \frac{\delta}{\delta \omega_j} (X\omega - y)^T (X\omega - y)$$

$$\Rightarrow \frac{\delta}{\delta \omega_j} \phi(\omega) = \frac{\delta}{\delta \omega_j} (\omega^T X^T - y^T)(X\omega - y) = \frac{\delta}{\delta \omega_j} \left(\omega^T X^T X \omega - y^T X \omega - \omega^T X^T y + y^T y \right)$$

$$\Rightarrow \frac{\delta}{\delta \omega_j} \phi(\omega) = \frac{\delta (\omega^T X^T X \omega)}{\delta \omega_j} - \underbrace{2 \frac{\delta}{\delta \omega_j} \omega^T X^T y}_{} + \frac{\delta}{\delta \omega_j} (y^T y)$$

$\because \omega^T X^T y = (y^T X \omega)^T$

as $y^T X \omega$ = one number

$\Rightarrow (y^T X \omega)^T$ = same number

hence $\boxed{\omega^T X^T y = y^T X \omega}$

$$y^T X \omega = [y_1 \; \cdots \; y_n] \begin{bmatrix} \leftarrow x_1 \rightarrow \\ \leftarrow x_2 \rightarrow \\ \vdots \\ \leftarrow x_n \rightarrow \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_m \end{bmatrix}$$

$$= [y_1 \; \cdots \; y_n] \begin{bmatrix} x_1 \omega \\ x_2 \omega \\ \vdots \\ x_n \omega \end{bmatrix} = \sum_{i=1}^{n} y_i (x_i \omega)$$

2

now $y^T y = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} y_i^2$

so $\frac{\partial}{\partial \omega_j} y^T y = \frac{\partial}{\partial \omega_j} \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \frac{\partial y_i^2}{\partial \omega_j} = \sum_{i=1}^{n} 0 = \boxed{0} \Rightarrow \boxed{\frac{\partial y^T y}{\partial \omega} = \vec{0}}$

Now $y^T X \omega = \begin{bmatrix} y_1 & \dots & y_n \end{bmatrix} \begin{bmatrix} - x_1^T - \\ \vdots \\ - x_n^T - \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_m \end{bmatrix} = \sum_{i=1}^{n} y_i (x_i \cdot \omega)$

$\Rightarrow y^T X \omega = \sum_{i=1}^{n} y_i \sum_{k=1}^{m} x_{ik} \omega_k$

$\Rightarrow \frac{\partial}{\partial \omega_j} y^T X \omega = \frac{\partial}{\partial \omega_j} \sum_{i=1}^{n} y_i \sum_{k=1}^{m} x_{ik} \omega_k = \sum_{i=1}^{n} y_i \frac{\partial}{\partial \omega_j} \sum_{k=1}^{m} x_{ik} \omega_k$

$\frac{\partial}{\partial \omega_j} y^T X \omega = \sum_{i=1}^{n} y_i \, x_{ij}$

$\downarrow$

$\boxed{\frac{\partial}{\partial \omega} y^T X \omega = X^T y}$

now $\omega^T X^T X \omega = \begin{bmatrix} \omega_1 & \dots & \omega_m \end{bmatrix} \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \dots & x_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} - x_1^T - \\ \vdots \\ - x_n^T - \end{bmatrix} \begin{bmatrix} \omega_1 \\ \vdots \\ \omega_m \end{bmatrix}$

$= \begin{bmatrix} x_1 \cdot \omega & x_2 \cdot \omega & \dots & x_n \cdot \omega \end{bmatrix} \begin{bmatrix} x_1 \cdot \omega \\ \vdots \\ x_n \cdot \omega \end{bmatrix} = \sum_{i=1}^{n} (x_i \cdot \omega)^2 = \sum_{i=1}^{n} \left(\sum_{k=1}^{m} x_{ik} \omega_k \right)^2$

$$\frac{\delta}{\delta w_j} w^T X^T X w = \frac{\delta}{\delta w_j} \sum_{i=1}^{n} \left(\sum_{k=1}^{m} (x_{ik} w_k) \right)^2 = \frac{\delta}{\delta w_j}$$

$$= \sum_{i=1}^{n} \frac{\delta}{\delta w_j} \left(\sum_{k=1}^{m} x_{ik} w_k \right)^2$$

$$\cancel{= 2 \sum_{i=1}^{n} \sum_{k=1}^{m} \frac{\delta x_{ik} w_k}{\delta w_j} = 2 \sum_{i=1}^{n}}$$

$$= \sum_{i=1}^{n} \left(2 \sum_{k=1}^{m} x_{ik} w_k \right) \cdot \sum_{k=1}^{m} \frac{\delta}{\delta w_j} x_{ik} w_k$$

$$\frac{\delta}{\delta w_j} w^T X^T X w = 2 \sum_{i=1}^{n} \left(\sum_{k=1}^{m} x_{ik} w_k \right) \cdot x_{ij}$$

$$\boxed{\frac{\delta}{\delta w} (w^T X^T X w) = 2 X^T X w}$$

$$\frac{\delta}{\delta w} \phi(w) = \frac{\delta}{\delta w} (w^T X^T X w) - 2 \frac{\delta}{\delta w} (y^T X w) + \frac{\delta}{\delta w} (y^T y)$$

$$\boxed{\frac{\delta}{\delta w} \phi(w) = 2(X^T X w) - 2 X^T y}$$

for optimal w, $\frac{\delta}{\delta w} \phi(w) = 0 \rightarrow \cancel{2}(X^T X w) - \cancel{2} X^T y = 0$

$$X^T X w = X^T y$$

$$\boxed{w = (X^T X)^{-1} X^T y}$$

(4)

5 Ridge Regression

$$\min_{\phi} \phi(\omega) = \|X\omega - y\|^2 + \lambda\|\omega\|^2$$

want to find ω such that $\phi(\omega)$ is minimum
hence find ω such that $\frac{\delta}{\delta\omega}\phi(\omega) = 0$

from part A we already know → $\boxed{\frac{\delta}{\delta\omega}\|X\omega - y\|_2^2 = 2(X^TX)\omega - 2X^Ty}$

so finding $\frac{\delta}{\delta\omega}\lambda\|\omega\|^2 \to \frac{\delta}{\delta\omega}\lambda\, \omega^T\omega$

$$\to \frac{\delta}{\delta\omega_j}\lambda\sum_{i=1}^{m}\omega_i^2 = \lambda\sum_{i=1}^{m}\frac{\delta}{\delta\omega_j}\omega_i^2$$

$$\frac{\delta}{\delta\omega_j}\lambda\|\omega\|^2 = 2\lambda\omega_j$$

$$\Rightarrow \frac{\delta}{\delta\omega}\lambda\|\omega\|^2 = \boxed{2\lambda\omega}$$

$$\Rightarrow \frac{\delta}{\delta\omega}\phi(\omega) = 2(X^TX)\omega - 2X^Ty + 2\lambda\omega$$

set $\frac{\delta}{\delta\omega}\phi(\omega) = 0 \to \cancel{2}(X^TX)\omega - \cancel{2}X^Ty + \cancel{2}\lambda\omega = 0$

$$(X^TX\omega) + \lambda(\omega) = X^Ty$$

$$\boxed{\omega = (X^TX + \lambda I)^{-1}X^Ty}$$

5

Q2 Given $P(H) = p$
$P(T) = 1-p$

1) the probability → $p \times p \times (1-p) \times (1-p) \times p$

$= p^3 (1-p)^2$

$P(HHTTH) = p^3(1-p)^2$

$\log(P(HHTTH)) = \log(p^3(1-p)^2)$

$= \log(p^3) + \log((1-p)^2)$

$= 3\log p + 2\log(1-p)$

2) a) P(chosen fair coin) × P(HHTTH | Fair coin)

$= \frac{1}{2} \times \left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 = \left(\frac{1}{2}\right)^6 = \boxed{1/64}$

b) P(Biased) × P(HHTTH | Biased)

$= \frac{1}{2} \times \left(\frac{2}{3}\right)^3 \times \left(\frac{1}{3}\right)^2 = \frac{8}{2 \times 3^5} = 4/3^5 = \boxed{\frac{4}{243}}$

3) want $\max_p \log(P(HHTTH))$

so setting $\boxed{\frac{d}{dp} \log(P(HHTTH)) = 0}$

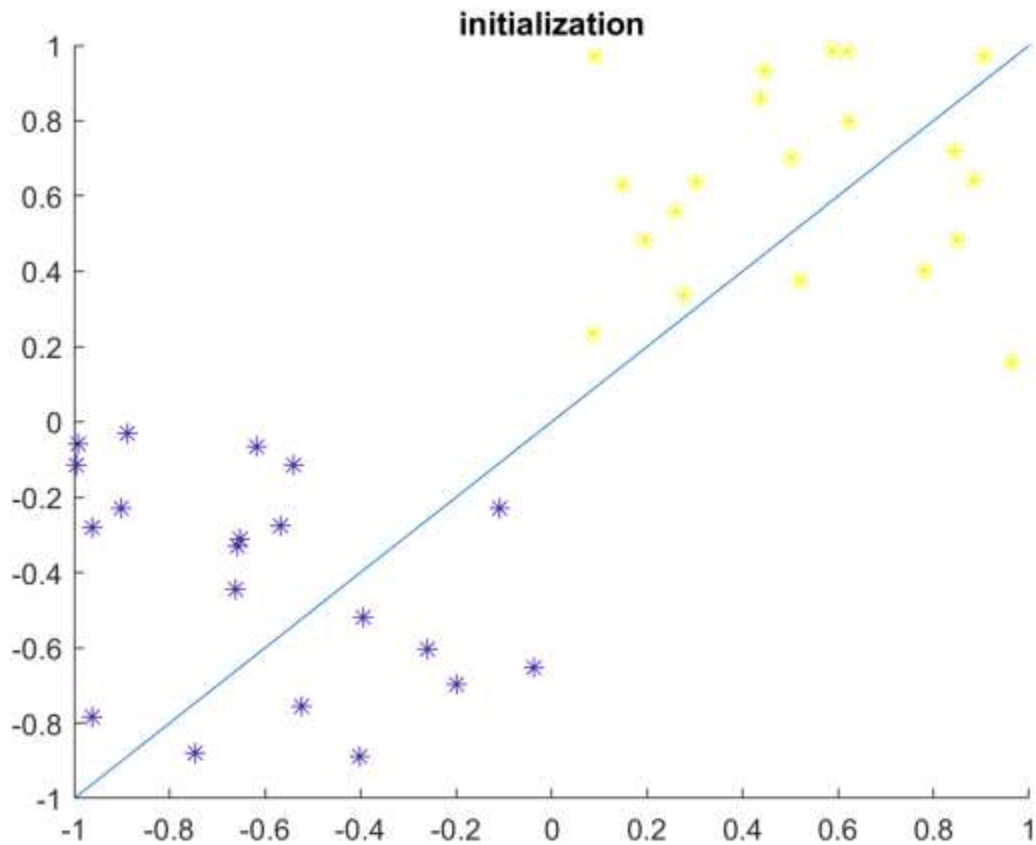
6

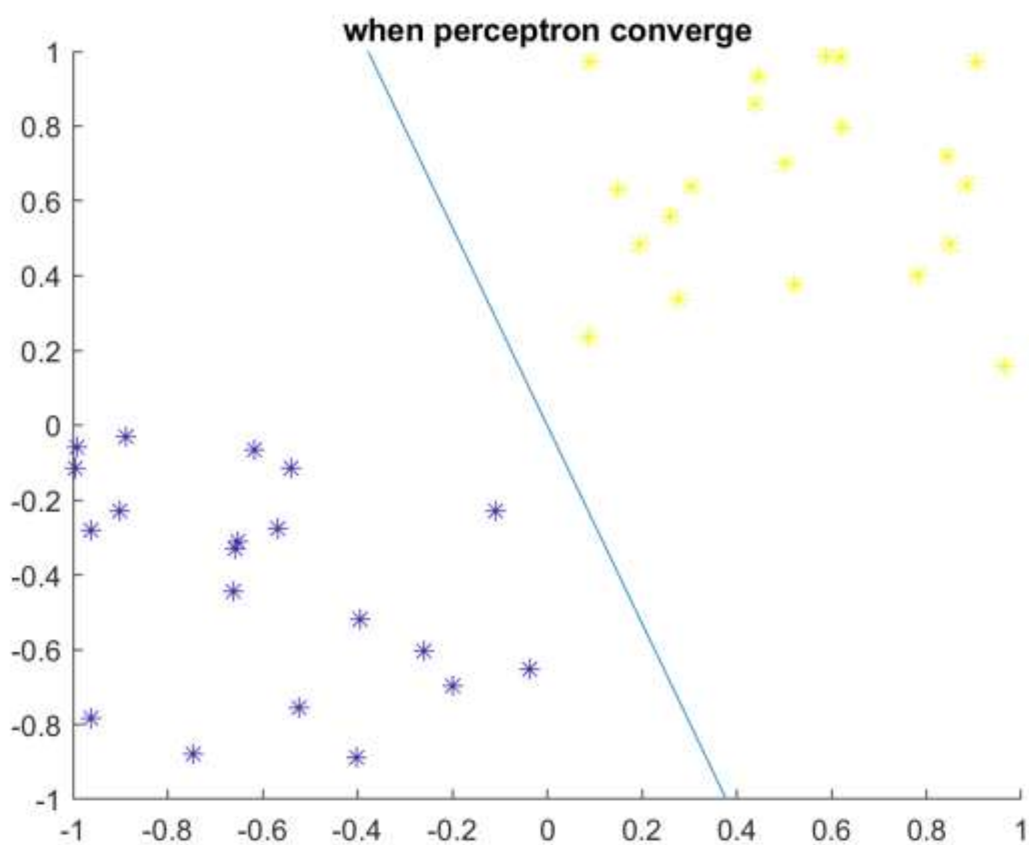
$$= \frac{d}{dp}\left(3\log(p) + 2\log(1-p)\right) = 0$$

$$= \frac{3}{p} + 2 \cdot \times \frac{1}{1-p} \times -1 = 0$$

$$= \frac{3}{p} = \frac{2}{1-p} \Rightarrow 3 - 3p = 2p$$

$$\Rightarrow \boxed{p = 3/5}$$


initialization
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1


when perceptron converge
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1