Homework 1

Yuanteng Chen (3039725444)

1. Least Squares and the Min-norm problem from the Perspective of SVD

(a) How can we solve $\min_w \|Xw-y\|^2$

Solution: it's an ordinary Least Squares problem to solve $\min \|Xw-y\|^2$:

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

(b) plug in the SVD $X = U\Sigma V^T$ and simplify:

Solution: $\vec{w} = (X^T X)^{-1} X^T \vec{y}$

as U and V are orthonormal square matrics,

$$U^T = U^{-1},\ V^T = V^{-1},\ U^T U = U^{-1} U = E$$

$$\therefore \vec{w} = (V\Sigma^T U^T U \Sigma V^T)^{-1} V \Sigma^T U^T \vec{y}$$

$$= (V\Sigma^T E \Sigma V^T)^{-1} V \Sigma^T U^T \vec{y}$$

$$= (V\Sigma^T \Sigma V^T)^{-1} V \Sigma^T U^T \vec{y}$$

$$(V\Sigma^T \Sigma V^T)^{-1} = V \Sigma^{-1} (\Sigma^T)^{-1} V^{-1}$$

$$\therefore \vec{w} = V \Sigma^{-1} (\Sigma^T)^{-1} V^{-1} V \Sigma^T U^T \vec{y}$$

$$= V \Sigma^{-1} (\Sigma^T)^{-1} \Sigma^T U^T \vec{y}$$

$$= V \Sigma^{-1} U^T \vec{y}$$

$\Sigma^+ = \Sigma^{-1} \rightarrow$ an $n \times m$ matrix with the reciprocals of the single value ($\frac{1}{\sigma_i}$) along the diagonal.

(c) $w^* = Ay$. What happens if we left-multipy by our matrix A?

Solution: $Ax = V\Sigma^+ u^T u \Sigma V^T$

as $u^T u = u^{-1} u = E$

$Ax = V \Sigma^T \Sigma V^T$

assume: $\Sigma = \begin{pmatrix} \sigma_1 & \cdots & 0 \\ 0 & \sigma_2 & \vdots \\ \vdots & \ddots & \sigma_n \\ 0 & \cdots & 0 \end{pmatrix}$ $\Sigma^+ = \begin{pmatrix} \frac{1}{\sigma_1} & \cdots & \cdots & 0 \\ \vdots & \frac{1}{\sigma_2} & & \vdots \\ 0 & \cdots & \frac{1}{\sigma_n} & \cdots 0 \end{pmatrix}$

$n \times m \quad m \times n$

$\Sigma^+ \Sigma = E \ (n \times n)$

$\therefore Ax = V V^T = V V^{-1} = E$

(d) in the case $m<n$. we want to solve $\min \|w\|^2$, $Xw = y$. What is the minimum norm solution?

Solution: $Xw = y$

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

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

$$w = (X^T X)^{+} X^T y$$

But I don't know how to solve $\min \|w\|^2$

(e) Plug in the SVD $X = U \Sigma V^T$ and simplify.

Solution: same as (b).

$$w = V \Sigma^{+} U^T \vec{y}$$

(f) min-norm solution is in the form $w^* = By$

What happens if we right-multify X by matrix B?

Solution:

Same as (c): $XB = U \Sigma V^T V \Sigma^{+} U^T$

$$= U \Sigma \Sigma^{+} U^T$$

$$= U U^T$$

$$= E$$

2. The 5 Interpretations of Ridge Regression

(a) P1: Optimization problem

$$\underset{w}{\operatorname{argmin}} \|y - Xw\|_2^2 + \lambda \|w\|_2^2$$

$X \in R^{n\times d}$, $y \in R^n$ is the target vector of values.

Solution: $\|y - Xw\|_2^2 + \lambda\|w\|_2^2$

$$= y^T y + (Xw)^T (Xw) - 2y \cdot Xw + \lambda w^T w$$

$$= y^T y + w^T X^T X w - 2y \cdot Xw + \lambda w^T w$$

in order to find min

$$\frac{d}{dw}(y^T y + w^T x^T x w - 2y \cdot Xw + \lambda w^T w)$$

$$= 2X^T X w - 2X^T y + 2\lambda w = 0$$

$$X^T X w - X^T y + \lambda w = 0$$

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

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

(b) P2: "Hack of shifting the singular values,

$X = U \Sigma V^T$ be the full SVD of the X

Plug this into the Ridge Regression solution and simplify. What happens to the singular values of $(X^T X + \lambda E)^{-1} X^T$ when $\sigma_i \ll \lambda$ or $\sigma_i \gg \lambda$

Solution: as U and V are both square orthonormal matrices $U^TU = V^TV = E$

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

$$= (V\Sigma^T U^T U \Sigma V^T + \lambda E)^{-1} V\Sigma^T U^T \vec{y}$$

$$= (V\Sigma^T \Sigma V^T + \lambda E)^{-1} V\Sigma^T U^T \vec{y}$$

as $\lambda E = \lambda V E V^T$

$$W = (V\Sigma^T \Sigma V^T + \lambda V E V^T)^{-1} V \Sigma^T U^T \vec{y}$$

$$= (V(\Sigma^T\Sigma + \lambda E)V^T)^{-1} V\Sigma^T U^T \vec{y}$$

$$= (V^T)^{-1} (\Sigma^T\Sigma + \lambda E)^{-1} V^{-1} V \Sigma^T U^T \vec{y}$$

$$= V(\underbrace{\Sigma^T\Sigma}_{\substack{(d\times n)\cdot(n\times d) \\ d\times d}} + \lambda E)\, \Sigma^T U^T \vec{y}$$

$$\lambda E + \Sigma^T \Sigma = \operatorname{diag}(\sigma_i^2 + \lambda)$$

$$\therefore W = V \operatorname{diag}\left(\frac{1}{\sigma_i^2 + \lambda}\right) \cdot \Sigma^T U^T \vec{y}$$

$$\Sigma^T = \begin{bmatrix} \sigma_1 & \cdots & & 0 \\ \vdots & \ddots & & \vdots \\ 0 & \sigma_d & \cdots & 0 \end{bmatrix} \quad (d\times n)$$

$$\therefore W = V\left[\operatorname{diag}\left(\frac{\sigma_i}{\sigma_i^2 + \lambda}\right), O_{(d, n-d)}\right] U^T \vec{y}$$

in $W = V\left[\operatorname{diag}\left(\frac{\sigma_i}{\sigma_i^2 + \lambda}\right), O_{\in(d\times n-d)}\right] U^T \vec{y}$

$\frac{\sigma_i}{\sigma_i^2+\lambda}$ $(i=1,\cdots,d)$ are singular values and λ prevents denominator of any singular value to be 0

$$\begin{cases} \text{When } \sigma_i \ll \lambda, & \frac{\sigma_i}{\sigma_i^2+\lambda} \approx \frac{\sigma_i}{\lambda} \\ \text{when } \sigma_i \gg \lambda, & \frac{\sigma_i}{\sigma_i^2+\lambda} \approx \frac{1}{\sigma_i} \end{cases}$$

(c): P3: Maximum A posteriori (MAP) estimation. Ridge Regression can be viewed as finding the MAP estimate when we apply a prior on the W, we can think of the prior for W as being $N(0, I)$ and view the random Y as $Y = \vec{x}^T W + \sqrt{\lambda} N$,

(noise N is distributed iid as $N(0,1)$)

vector → $Y = XW + \sqrt{\lambda} N$ (rows of $X = n$)

Show that (1) is the MAP estimate for W given an observation $Y=y$.

Solution: MAP estimate for W is same as

$$\arg\max_W P(W \mid Y=y)$$

$$= \arg\max_w \frac{P(w,y)}{P(y)}$$

$$= \arg\max_w \frac{P(y \mid w)\cdot P(w)}{P(y)} = \arg\max_W \frac{\prod_{i=1}^{n} P(y_i \mid w) P(w)}{P(y)}$$

$$Y = XW + \sqrt{\lambda}\, N$$

$$y_i = x_i^T W + \sqrt{\lambda}\, N$$

$$N = \frac{y_i - x_i W}{\sqrt{\lambda}} \sim N(0,1)$$

for $X \sim N(0,1) \rightarrow p(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$

$$\therefore p(y_i|w) = \frac{e^{-\frac{1}{2}\left(\frac{y_i - x_i w}{\sqrt{\lambda}}\right)^2}}{\sqrt{2\pi}}$$

$$\therefore \underset{w}{\operatorname{argmax}} \frac{\prod_{i=1}^{n} p(y_i|w)\, p(w)}{p(y)} = \underset{w}{\operatorname{argmax}} \frac{\frac{e^{-\frac{\|w\|^2}{2}}}{\sqrt{2\pi}} \prod_{i=1}^{n} \frac{e^{-\frac{\left(\frac{y_i - x_i w}{\sqrt{\lambda}}\right)^2}{2}}}{\sqrt{2\pi}}}{p(y)}$$

$\because p(y), \sqrt{2\pi}$ will not change with W

$$\therefore MAP \Leftrightarrow \underset{w}{\operatorname{argmax}}\; e^{-\frac{\|w\|^2}{2}} \prod_{i=1}^{n} e^{\frac{-(y_i - x_i w)^2}{2\lambda}}$$

$$\ln(MAP) \Leftrightarrow \underset{w}{\operatorname{argmax}} -\frac{\|w\|^2}{2} + \sum_{i=1}^{n}\left(-\frac{(y_i - x_i w)^2}{2\lambda}\right)$$

$$\Leftrightarrow \underset{w}{\operatorname{argmax}} -\frac{\|w\|^2}{2} - \frac{1}{2\lambda}\sum_{i=1}^{n}(y_i - x_i w)^2$$

$$\Leftrightarrow \underset{w}{\operatorname{argmin}}\; \|w\|^2 + \frac{1}{\lambda}\sum_{i=1}^{n}(y_i - x_i w)^2$$

$$\Leftrightarrow \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - x_i w)^2 + \lambda \|w\|^2$$

(d) P4: Fake data

$$\vec{y} = \begin{bmatrix} \vec{y} \\ 0_d \end{bmatrix}, \quad \vec{X} = \begin{bmatrix} X \\ \sqrt{\lambda} I_d \end{bmatrix}$$

where 0_d is the zero vector in $\mathbb{R}^d$ and $I_d \in \mathbb{R}^{d\times d}$ is the identity matrix:

Solution:

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

$$= \left(\begin{bmatrix} X \\ \sqrt{\lambda} E \end{bmatrix}^T \begin{bmatrix} X \\ \sqrt{\lambda} E \end{bmatrix} \right)^{-1} \begin{bmatrix} X \\ \sqrt{\lambda} E \end{bmatrix}^T \begin{bmatrix} \vec{y} \\ \vec{0} \end{bmatrix}$$

$$= \left([X^T \ \sqrt{\lambda} E] \cdot \begin{bmatrix} X \\ \sqrt{\lambda} E \end{bmatrix} \right)^{-1} \begin{bmatrix} X \\ \sqrt{\lambda} E \end{bmatrix}^T \cdot \begin{bmatrix} \vec{y} \\ \vec{0} \end{bmatrix}$$

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

(e) P5: Fake Features

$$\vec{X} = [X, \sqrt{\lambda} E_n]$$

We are interested in the min-norm solution:

$$\underset{\eta}{\operatorname{argmin}} \|\eta\|_2^2 \quad \vec{X} \eta = \vec{y}$$

$$[X, \sqrt{\lambda} E_n] \begin{bmatrix} \vec{w} \\ \vec{\xi} \end{bmatrix} = \vec{y}$$

$X: n\times d$ $I: n\times n$ $\vec{w}: d\times 1$ $\vec{\xi}: n\times 1$

to solve the min-norm:

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

$$\begin{bmatrix} \vec{w} \\ \vec{f} \end{bmatrix} = \begin{bmatrix} X^T \\ \sqrt{\lambda} E \end{bmatrix} \left([X, \sqrt{\lambda} E_n] \begin{bmatrix} X^T \\ \sqrt{\lambda} E \end{bmatrix} \right)^{-1} \vec{y}$$

$$\begin{bmatrix} \vec{w} \\ \vec{f} \end{bmatrix} = \begin{bmatrix} X^T \\ \sqrt{\lambda} E \end{bmatrix} \cdot (XX^T + \lambda E)^{-1} \vec{y}$$

$$\therefore \begin{cases} \vec{w} = X^T (XX^T + \lambda E)^{-1} \vec{y} \\ \vec{f} = \sqrt{\lambda} E \cdot (XX^T + \lambda E)^{-1} \vec{y} \end{cases}$$

(g) $\vec{W_r} = (X^T X + \lambda E)^{-1} X^T \vec{y}$

what happens when $\lambda \to \infty$?

Solution, when $\lambda \to \infty$

$(X^T X + \lambda E) \simeq \mathrm{diag}_n(\lambda)$

$(X^T X + \lambda E)^{-1} \simeq \mathrm{diag}n(0)$

$\therefore \vec{W_r} \approx \vec{0}$

(h) what happens when $\lambda \to 0$

Solution:

when $\lambda \to 0$.

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

3. General Case Tikhonov Regularization

Consider the optimization problem:

$$\min_x \|W_1(A\vec{x}-\vec{b})\|_2^2 + \|W_2(\vec{x}-c)\|_2^2$$

W_1 can be viewed as a generic weighting of the residuals and W_2 along with c can be viewed as a general weighting of the parameters.

(a)

$$f(x) = \|W_1(A\vec{x}-\vec{b})\|_2^2 + \|W_2(\vec{x}-c)\|_2^2$$

$$= [W_1(A\vec{x}-\vec{b})]^T W_1(A\vec{x}-\vec{b}) + [W_2(\vec{x}-c)]^T W_2(\vec{x}-c)$$

$$= (A\vec{x}-\vec{b})^T W_1^T W_1 (A\vec{x}-\vec{b}) + (\vec{x}-c)^T W_2^T W_2 (\vec{x}-c)$$

$$= x^T A^T W_1^T W_1 A x - 2b^T W_1^T W_1 A x - b^T W_1^T W_1 \vec{b}$$
$$+ x^T W_2^T W_2 x - 2c^T W_2^T W_2 x + c^T W_2^T W_2 c$$

$$\frac{df}{dx} = 2A^T W_1^T W_1 A\vec{x} - 2b^T W_1^T W_1 A + 2W_2^T W_2 \vec{x}$$
$$- 2c^T W_2^T W_2$$

$$\frac{df}{dx} = 0 \Rightarrow (2A^T W_1^T W_1 A + 2W_2^T W_1)\vec{x} = 2b^T W_1^T W_1 A$$
$$+ 2c^T W_2^T W_2$$

$$\therefore (A^T W_1^T W_1 A + W_2^T W_1)\vec{x}$$
$$= (b^T W_1^T W_1 A + c^T W_2^T W_2)$$

$$\vec{x} = (A^T W_1^T W_1 A + W_2^T W_1)^{-1}(b^T W_1^T W_1 A + c^T W_2^T W_2)$$

(b) construct an appropriate matrix C and vector d that allows to rewrite this problem as

$$\min_x \|Cx-d\|^2$$

and use the OLS solution ($x^* = (C^TC)^{-1}C^Td$)

Solution,

$$\min_x \|W_1(Ax-b)\|_2^2 + \|W_2(x-c)\|_2^2$$

① the first part:

$$\|W_1(Ax-b)\|_2^2$$

$$\rightarrow C_1 = [W_1 A], \quad d_1 = [W_1 b]$$

$$\|C_1x-d_1\|^2 = \|W_1(Ax-b)\|_2^2$$

② the second part:

$$\|W_2(x-c)\|_2^2$$

$$\rightarrow C_2 = [W_2] \quad d_2 = [W_2 c]$$

$$\|C_2x-d_2\|_2^2 = \|W_2(x-c)\|_2^2$$

$$\therefore C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} W_1 A \\ W_2 \end{bmatrix}$$

$$d = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} W_1 b \\ W_2 c \end{bmatrix}$$

$$\therefore x^* = (C^TC)^{-1}C^Td$$

$$= \left([A^TW_1^T, W_2^T] \begin{bmatrix} W_1 A \\ W_2 \end{bmatrix} \right)^{-1} [A^TW_1^T, W_2^T] \begin{bmatrix} W_1 b \\ W_2 c \end{bmatrix}$$

$$= (A^TW_1^TW_1A + W_2^TW_2)^{-1} [A^TW_1^TW_1b + W_2^TW_2c]$$

(c) choose a W_1, W_2 and c such that this reduces to the simple case of ridge regression that you've seen in the previous problem,

$$x^* = (A^T A + \lambda E)^{-1} A^T b$$

Solution:

$$x = (A^T W_1^T W_1 A + W_2^T W_2)^{-1} (A^T W_1^T W_1 b + W_2^T W_2 c)$$

$$\Downarrow$$

$$x^* = (A^T A + \lambda E)^{-1} A^T b$$

$$\therefore \begin{cases} W_1^T W_1 = E \\ W_2^T W_2 = \lambda E \\ W_2^T W_2 c = 0 \end{cases} \Rightarrow \begin{cases} W_1 = E \\ W_2 = \sqrt{\lambda} E \\ c = \vec{0} \end{cases}$$

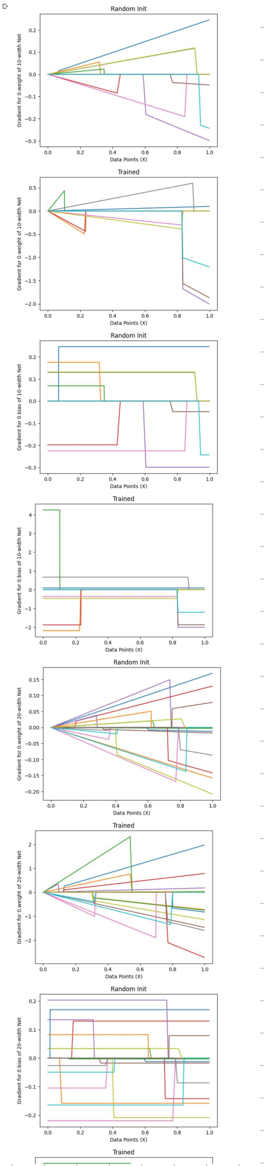
4. Coding Fully Connected Networks.

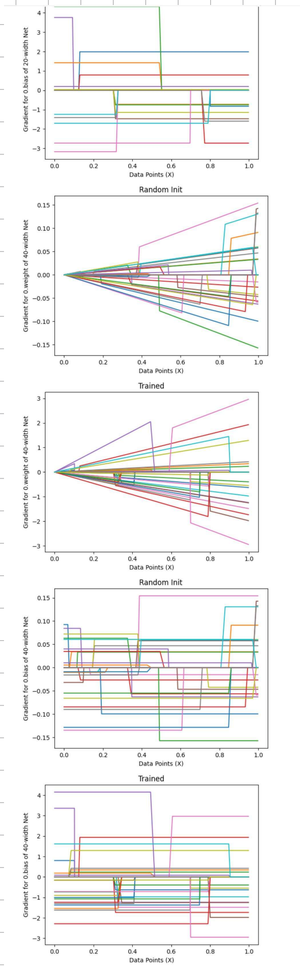
(a) ① higher learning rate is more suitable for three-layer network and we need to low down the learning-rate when training a five layers network.

(b) of course training five layers network costs more time.

5. Visualizing features from local linearization of neutral nets.

(a)



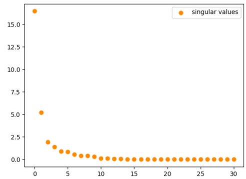


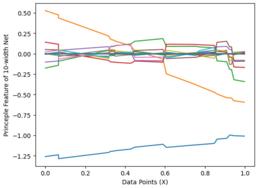
(b) SVD for feature matrix

During training, we can imagine that we have a generalized linear model with a feature matrix corre
corresponding to each learnable parameter. We know from our analysis of gradient descent, that th
corresponding to this feature matrix are important

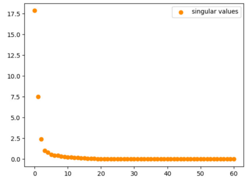
(b).

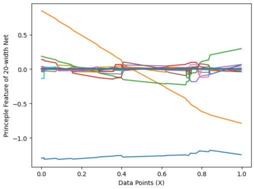
Width 10



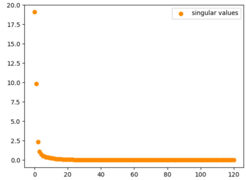


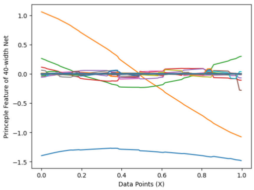
Width 20





Width 40





(c) Two-layer Network

Augment the jupyter notebook to add a second hidden layer of the same size as the first hidden layer, ful

6. Homework Process and Study Group.

(b) Yujie Zhao 3039725470

(c) 15 hours