Hanze Yao
3033093286

CS189 HW2

1. Linear Regression, Projection and Pseudoinverse

(a) Proof: $\|y-u\|^2 = \|y - P_x(y) + P_x(y) - u\|^2$

$$= \|y - P_x(y)\|^2 + \|P_x y - u\|^2$$

$$\geq \|y - P_x(y)\|^2$$

$$\therefore P_x(y) = \underset{u \in \text{range}(X)}{\arg\min} \|y - u\|^2$$

(b) First, proof: P is a rank-d orthogonal projection matrix $\Rightarrow$ there exists a $U \in R^{n \times d}$, $P = UU^T$, $U^TU = I$

P is a $\ldots$

$\text{rank}(P) = d$, $P = P^T$, $P^2 = P$

$\therefore P = V\Sigma V^T$

$\therefore P^TP = (V\Sigma V^T)^T(V\Sigma V^T) = V\Sigma V^T V\Sigma V^T = V\Sigma\Sigma V^T$

$= P = V\Sigma V^T$

$\therefore \Sigma^2 = \Sigma$ $\therefore \Sigma$ is a identity matrix

$P^TP = VV^T$

$\therefore U = V$

$V^TV = I$ $\therefore U^TU = I$

Second proof: there exists a $U \cdots$ $\Rightarrow$ P is a rank-d $\cdots$

$\therefore U \in R^{n \times d}$ $\therefore U^TU \in R^{d \times d} = I$ $\therefore \text{rank}(U) = d$

$\therefore P = UU^T$ $\therefore P^T = (UU^T)^T = (U^T)^TU^T = UU^T = P$

$P^2 = UU^TUU^T = UU^T = P$

$\therefore \text{rank}(U) = d$ $\therefore \text{rank}(U^T) = d$ $\therefore \text{rank}(UU^T) = \text{rank}(P)$

$= d$

$\therefore P$ is a rank-d $\cdots$

(c) Proof: $\text{tr}(P) = \text{tr}(UU^T) = \text{tr}(U^TU) = \text{tr}(I)$

$U \in R^{n \times d}$ $\therefore U^TU \in R^{d \times d}$ $\therefore \text{tr}(I) = d = \text{tr}(P)$

(d) Proof: $\therefore X \in R^{n \times d}$ $\text{rank}(X) = d$ $n > d$ $\therefore \text{rank}(X^T) = d$ $\text{rank}(X^TX)$

$\text{rank}((X^TX)^+) = d$ $\therefore \text{rank}(X(X^TX)^+X^T) = d$ $= d$

$(X(X^TX)^+X^T)^T = X((X^TX)^+)^TX^T$

$\therefore (X^TX)^T = X^TX$ $\therefore X^TX$ is symmetric $\therefore (X^TX)^+$ is also

symmetric since X^TX has full rank (d).

$\therefore (X(X^TX)^+X^T)^T = X(X^TX)^+X^T$

$\therefore (X(X^TX)^+X^T)^2 = X(X^TX)^+X^TX(X^TX)^+X^T = X(X^TX)^+X^T$

$\therefore$ $X(X^TX)^{-1}X^T$ is a rank-d orthogonal projection matrix

$\therefore$ $UU^T = X(X^TX)^{-1}X^T$, $U^TU = I$

$\therefore$ $UU^TU = X(X^TX)^{-1}X^TU$

 $U = X(X^TX)^{-1}X^TU$

 $U = X(X^TX)^{-1}X^T$

(e)

(a) $\because$ $X = \sum\limits_{i=1}^{\min(n,d)} \sigma_i u_i v_i^T$

 $= \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_d u_d v_d^T$ (given $n > d$)

 $\therefore$ $v_1^T, v_2^T, \cdots, v_d^T$ are all orthonormal $R^{d\times 1}$ row vectors

$\because \sigma_i > 0$ $\therefore$ Each row of X is a linear combination of

$\therefore \sigma_i u_i \neq 0$ $v_1^T, v_2^T, \cdots, v_d^T$

 $\therefore$ $\{v_i : \sigma_i > 0\}$ are an orthonormal basis for the row space

 of X

(b) $X^T = \sum\limits_{i=1}^{\min(n,d)} \sigma_i v_i u_i^T$

For the same reason as above

 $\{u_i : \sigma_i > 0\}$ are an orthonormal basis for the row space

 of X^T

 $\{u_i : \sigma_i > 0\}$ are an orthonormal basis for the column space

 of X.

(f) $X^\dagger = \sum\limits_{i, \sigma_i > 0} \sigma_i^{-1} v_i u_i^T = V\Sigma^{-1}U^T$

$\tilde{P}_X(y) = XX^\dagger y = U\Sigma V^T V\Sigma^{-1} U^T y$

 If $\text{rank}(X) = d$ then $V \in R^{d\times d}$ which is full rank.

 but $U \in R^{n\times n}$ which is not full rank

 $\therefore$ $\tilde{P}_X(y) = UU^T y$

 If $\text{rank}(X) = d$ then there are multiple solutions for P_X

 $\tilde{P}_X(y)$ will be one of them.

 If $\text{rank}(X) = d = n$ then there is only one solution for P_X

 and $\tilde{P}_X(y) = P_X$

2. The Least Norm Solution

(a) Using Lagrange Multiplier Method

Lagrangian function should be

$$L(w, \lambda) = w^T w + \lambda \times (X^T X w - X^T y)$$

$\hookrightarrow$ Because the constraint

$$\therefore \frac{\partial}{\partial w} L(w, \lambda) = 2w^T + \lambda X^T X = 0$$

is w must minimize

$$\|Xw - y\|_2^2 \quad \therefore \quad \nabla_w \|Xw - y\|_2^2$$

$$w^T = -\frac{\lambda}{2} X^T X$$

$$= 0$$

$$w = -X^T \left[X \left(-\frac{\lambda}{2} \right)^T \right]$$

which gives $X^T X w - X^T y = 0$

$$\therefore \hat{w}_{LN} \text{ is on the row space of } X.$$

Next, proof of uniqueness.

Assume w_1 and w_2 are two solutions for $\hat{w}_{LN}$

Then $\| w_1 + \lambda (w_2 - w_1) \|_2^2$ has two minimum values

as $\lambda = 1$ and $\lambda = 0$.

$$\| w_1 + \lambda (w_2 - w_1) \|_2^2$$

$$= \|w_1\|^2 + 2\lambda \|w_1\| \|w_2 - w_1\| + \lambda^2 \| w_2 - w_1 \|_2^2$$

$$\therefore w_1 \neq w_2 \quad \therefore \|w_2 - w_1\|_2^2 \neq 0 \text{ and } > 0$$

$$\therefore \| w_1 + \lambda (w_2 - w_1) \|_2^2 \text{ has single minimum value}$$

which contradicts my assumption

$$\therefore \hat{w}_{LN} \text{ is unique.}$$

(b) $X = \sum\limits_{i: \sigma_i > 0} \sigma_i u_i v_i^T \qquad X^T = \sum\limits_{i: \sigma_i > 0} \sigma_i v_i u_i^T$

$$\therefore \tilde{w} = \sum\limits_{i: \sigma_i > 0} \frac{1}{\sigma_i} v_i (u_i^T y) = \sum\limits_{i: \sigma_i > 0} (\sigma_i v_i u_i^T) \times \left(\frac{1}{\sigma_i^2} y \right)$$

$$= X^T (\Sigma^{-2} y)$$

$$\therefore \tilde{w} \text{ is in the row space of } X$$

Then check the optimality.

$$X = U \Sigma V^T \qquad \tilde{w} = V \Sigma^{-1} U^T y$$

check $\quad X^T X \tilde{w} = X^T y$

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

$$= V \Sigma U^T y = X^T y$$

$$\therefore \tilde{w} = \hat{w}_{LN}$$

(C)

(1) $X^+ X = V\Sigma^+ U^T U \Sigma V^T = VV^T$

∴ $\{V_i : \sigma_i > 0\}$ is an orthonormal basis for the row space of X

∴ $X^+ X$ is the orthogonal projection matrix onto the row space of X.

(2) $X = \sum_i \sigma_i U_i V_i^T$ $X^T = \sum_i \sigma_i V_i U_i^T$

$X^T X = \sum_i \sigma_i^2 V_i V_i^T$

$(X^T X)^+ = \sum_i \frac{1}{\sigma_i^2} V_i V_i^T$

$(X^T X)^+ X^T = \sum_i \frac{1}{\sigma_i^2} V_i V_i^T (\sigma_i V_i U_i^T)$

$= \sum_i \frac{1}{\sigma_i} V_i U_i^T = X^+$

(3) Need to show that $X^+ X w = X^+ y$ when w satisfies the least square optimality condition.

w satisfies the least square optimality

∴ $(X^T X) w = X^T y$

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

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

↳ according to C(2)

∴ $P_X(w) = X^+ X w = X^+ y$

(4) ∴ According to Z(1), $X^+ y$ is a orthogonal projection onto the row space of X.

According to Z(3), $X^+ y$ satisfies the least square optimality

∴ $X^+ y = \hat{W}_{LN}$ is consistent with previous part of the problem

$$V^T V = I$$
$$V^{-1} = V^T$$

3. The Ridge Regression Estimator

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

∵ λI is full rank ∴ $X^T X + \lambda I$ is also full rank

∴ $(X^T X + \lambda I)$ is invertible ∴ W_λ is unique.

$X = U S V^T$, $X^T = V S U^T$

$X^T X = V S U^T U S V^T = V S^2 V^T$

∴ $W_\lambda = (V S^2 V^T + \lambda I)^{-1} V S U^T y$

$= (V^{-1})(V^{-1} S^2 V^T V + V^{-1} \lambda I V)^{-1} (V^T)^{-1} V S U^T y$

$= (V^{-1})(S^2 + \lambda I)^{-1} V V S U^T y$

$= \dfrac{S}{S^2 + \lambda I} V U^T y$

$= \displaystyle\sum_{i=1}^{d} \dfrac{\sigma_i}{\sigma_i^2 + \lambda} V_i \langle u_i, y \rangle$

(b) $\|\widehat{W_\lambda}\|^2 = \widehat{W_\lambda}^T \widehat{W_\lambda} = \displaystyle\sum_{i=1}^{d} \dfrac{\sigma_i}{\sigma_i^2 + \lambda} V_i^T \langle u_i, y \rangle \times \underbrace{\dfrac{\sigma_i}{\sigma_i^2 + \lambda} V_i \langle u_i, y \rangle}_{\text{is a scalar}}$

$= \displaystyle\sum_{i=1}^{d} \left(\dfrac{\sigma_i}{\sigma_i^2 + \lambda}\right)^2 V_i^T V_i \langle u_i, y \rangle^2$

$= \displaystyle\sum_{i=1}^{d} \left(\dfrac{\sigma_i}{\sigma_i^2 + \lambda}\right)^2 \langle u_i, y \rangle^2$

(c) ∵ $\widehat{W_{LN}} = 0$ ∴ $\displaystyle\sum_{i: \sigma_i > 0} \sigma_i^{-1} \langle u_i, y \rangle V_i = 0$

∴ $\sigma_i > 0$, $\langle u_i, y \rangle$ is a scalar

∴ $\langle u_i, y \rangle = 0$

∴ $\widehat{W_\lambda} = \displaystyle\sum_{i=1}^{d} \dfrac{\sigma_i}{\sigma_i^2 + \lambda} V_i \underbrace{\langle u_i, y \rangle}_{0} = 0$ for all λ

(d) If $\widehat{W_{LN}} \neq 0$, then $\langle u_i, y \rangle \neq 0$, then $\widehat{W_\lambda} \neq 0$

∴ as λ increases and $\sigma_i > 0$, $\dfrac{\sigma_i}{\sigma_i^2 + \lambda}$ strictly decreases

∵ $\lambda \in (0, \infty)$ and $\sigma_i > 0$ ∴ $\dfrac{\sigma_i}{\sigma_i^2 + \lambda} > 0$

∴ as λ increases, $\left(\dfrac{\sigma_i}{\sigma_i^2 + \lambda}\right)^2$ decreases $\Rightarrow \|W_\lambda\|^2$ decreases

$\therefore \hat{w}_\lambda \neq 0 \quad \therefore \|\hat{w}_\lambda\|^2 > 0$

$\therefore$ map $\lambda \mapsto \|\hat{w}_\lambda\|^2$ is strictly decreasing and positive on $(0, \infty)$

(e) As $\lambda \to 0$

$$\hat{w}_\lambda = \sum_{i=1}^{d} \frac{\sigma_i}{\sigma_i^2} V_i \langle u_i, y \rangle$$

$$= \sum_{i=1}^{d} \frac{1}{\sigma_i} V_i \langle u_i, y \rangle$$

$$= \sum_{i=1}^{d} \frac{1}{\sigma_i} V_i u_i^T y = X^\dagger y = \hat{w}_{LN}$$

$$\therefore \lim_{\lambda \to 0} \hat{w}_\lambda = \hat{w}_{LN}$$

(f) Because of part (d), as λ increases, $\|\hat{w}_\lambda\|^2$ will decrease. Therefore, giving λ a reasonable value can control the value of $\|w\|^2$, which controls the complexity of the model.

4) Patrick vs Alvin

(a) For Patrick's problem.

$X_1, X_2 \ldots X_n$ are inputs, which is X

$Y_1, Y_2 \ldots Y_n$ are observations with noise, which is

$$Y = f(X) + N$$

Now we want to fit a degree d polynomial to this data. So this is a linear regression problem with feature matrix as

$$\begin{pmatrix} 1 & X_1 & X_1^2 & \cdots & X_1^d \\ 1 & X_2 & X_2^2 & \cdots & X_2^d \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_n & X_n^2 & \cdots & X_n^d \end{pmatrix}$$

(b) Seen in the code

(c) The average training error decrease suddenly as the increase of d at first, Then the average training error starts to decrease slowly as the d increases. (There is a slight increase from 18 to 19) This is because as the increase of d, the model becomes more complex and thus fits more closely to the data (with noise) and thus the average training error will decrease. If I try to fit a polynomial of degree n with a standard matrix inversion method, in the plot, the training error will suddenly increase from degree 18 to degree 19.

(d) For fresh error, after degree 4, the error starts to increases as degree d increases.

I think this is because, as d increases, Patrick starts to fit the noise in the data. As a result, overfit problem will apply, the fresh error (which is like test error) will increase.

(e) With these 2 plots, I think using degree 4 polynomial may be a good choice

(f) Not enough space. To next page

According to the test put the best degree is 4 (has the smallest validation error at $\lambda = 0.1$)

(f) According to the output, the best degree is 4, best lambda
is 0.15
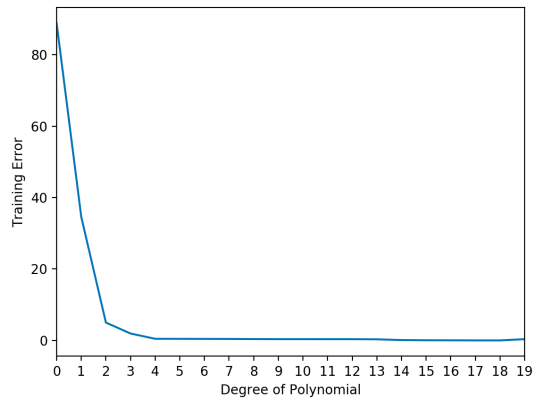
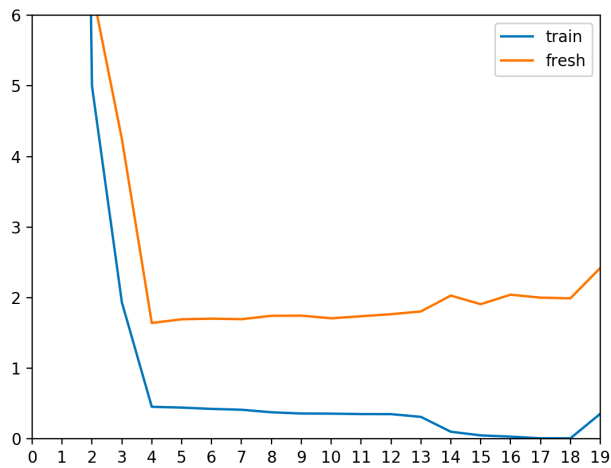With $\lambda = 0.1$, select the degree with the lowest Average validation
error which is degree

(g)

Q4 output

b.

d.

f & g.

```
YaohanzedeMacBook-Pro:hw2-data yaohanze$ python fg.py
0
1
2
3
4
5
Average train error:
[[0.05856762013 0.05856762066 0.05856762225 0.05856762491 0.05856762862]
 [0.05848920002 0.05848948229 0.0584902034  0.05849122    0.05849243261]
 [0.05846063295 0.05848702853 0.05848893458 0.05849036454 0.05849178744]
 [0.05839260718 0.05848700898 0.05848892445 0.05849035741 0.05849178171]
 [0.05831118703 0.05848700861 0.05848892426 0.05849035728 0.05849178161]
 [0.05815215946 0.0584870086  0.05848892426 0.05849035728 0.05849178161]]
Average valid error:
[[0.05857468619 0.05857468536 0.0585746856  0.05857468691 0.05857468927]
 [0.05852250667 0.05852112813 0.05852038752 0.05852010745 0.05852016181]
 [0.05854617135 0.0585210268  0.05852032221 0.05852006042 0.0585201252 ]
 [0.05860046897 0.05852102354 0.05852032035 0.05852005896 0.05852012386]
 [0.05869725681 0.05852102357 0.05852032036 0.05852005896 0.05852012386]
 [0.05887534948 0.05852102357 0.05852032036 0.05852005896 0.05852012386]]
Best degree: 4
Best lambda: 0.15
```