CS189 HW2

Hanze Yao
3033093286

1. Linear Regression, Projection and Pseudoinverse

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

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

$$\therefore P_X(y) = \underset{u \in \text{range}(X)}{\text{argmin}} \|y-u\|^2$$

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

$\because P$ is a ...

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

$\therefore P^TP = VV^T$

$\therefore U = V$

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

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

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

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

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

$\because$ rank$(U) = d$ $\therefore$ rank$(U^T) = d$ $\therefore$ rank$(UU^T) =$ rank(P) $= d$

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

(c) Proof: $tr(P) = tr(UU^T) = tr(U^TU) = tr(I)$

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

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

rank$((X^TX)^{-1}) = d$ $\therefore$ rank$(X(X^TX)^{-1}X^T) = d$

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

$\because (X^TX)^T = X^TX$ $\therefore X^TX$ is symmetric $\therefore (X^TX)^{-1}$ is also symmetric since X^TX has full rank (d).

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

$\therefore (X(X^TX)^{-1}X^T)^2 = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T$

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

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

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

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

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

(e)

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

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

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

$\because \sigma_i > 0$ $\therefore \sigma_i u_i \neq 0$

$\therefore$ Each row of X is a linear combination of $v_1^T, v_2^T, \ldots v_d^T$

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

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

For the same reason as above

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

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

(f) $X^+ = \sum_{i:\sigma_i > 0} \sigma_i^{-1} v_i u_i^T = V\Sigma^{-1}U^T$

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

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

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

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

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

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

If $\mathrm{rank}(X) = d = n$ 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

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

↳ Because the constraint is w must minimize $\|Xw - y\|_2^2$ ∴ $\nabla_w \|Xw - y\|_2^2 = 0$ which gives $X^T X w - X^T y = 0$

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

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

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

∴ $\hat{W}_{LN}$ is on the rowspace 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 $\lambda = 0$.

$$\|W_1 + \lambda (W_2 - W_1)\|_2^2$$

$$= \|W_1\|^2 + 2\lambda \|W_1\| \|W_2 - W_1\| + \lambda^2 \|W_2 - W_1\|_2^2$$

∵ $W_1 \neq W_2$ ∴ $\|W_2 - W_1\|_2^2 \neq 0$ and > 0

∴ $\|W_1 + \lambda (W_2 - W_1)\|_2^2$ has single minimum value which contradicts my assumption

∴ $\hat{W}_{LN}$ is unique.

(b) $X = \sum_{i: \sigma_i > 0} \sigma_i u_i v_i^T \quad X^T = \sum_{i: \sigma_i > 0} \sigma_i v_i u_i^T$

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

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

∴ $\tilde{w}$ is in the rowspace of X

Then check the optimality.

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

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

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

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

∴ $\tilde{w} = \hat{w}_{LN}$

(C)

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

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

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

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

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

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

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

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

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

$\because$ w satisfies the least square optimality

$\therefore (X^TX)w = X^Ty$

$w = (X^TX)^+X^Ty$

$\therefore X^+Xw = X^+X\underbrace{(X^TX)^+X^T}_{\text{according to C(2)}}y = X^+XX^+y = X^+y$

$\therefore P_{X^T}(w) = X^+Xw = X^+y$

(4) $\because$ 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

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

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

3. The Ridge Regression Estimator

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

$\because \lambda I$ is full rank $\therefore X^T X + \lambda I$ is also full rank

$\therefore (X^T X + \lambda I)$ is invertible $\therefore W_\lambda$ is unique.

$X = USV^T$, $X^T = VSU^T$

$X^T X = VSU^T USV^T = VS^2V^T$

$\therefore W_\lambda = (VS^2V^T + \lambda I)^{-1} VSU^T y$

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

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

$= \frac{S}{S_i^2 + \lambda I} VU^T y$

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

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

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

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

(c) $\because \hat{W}_{LN} = 0 \therefore \sum_{i:\sigma_i > 0} \sigma_i^{-1} \langle u_i, y \rangle V_i = 0$

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

$\therefore \langle u_i, y \rangle = 0$

$\therefore \hat{W}_\lambda = \sum_{i=1}^{d} \frac{\sigma_i}{\sigma_i^2 + \lambda} V_i \underbrace{\langle u_i, y \rangle}_{0} = 0$ for all λ

(d) If $\hat{W}_{LN} \neq 0$, then $\langle u_i, y \rangle \neq 0$, then $\hat{W}_\lambda \neq 0$

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

$\because \lambda \in (0, \infty)$ and $\sigma_i > 0$ $\therefore \frac{\sigma_i}{\sigma_i^2 + \lambda} > 0$

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

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

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

(e) As $\lambda \to 0$

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

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

$$= \sum_{i=1}^{d} \frac{1}{\sigma_i} V_i U_i^T y = X^+ y = \hat{W}_{LN}$$

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

(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,

$x_1, x_2 \dots x_n$ are inputs, which is X

$y_1, y_2 \dots y_n$ 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 see next page

(e) According to the output 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.