Lecture 9 :

SVD in Machine Learning

Ridge Regression

$(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}, \quad i = 1, \ldots, n$

until now, we considered $n > p$, $X = \begin{bmatrix} - x_1^T - \\ - x_2^T - \\ \vdots \\ - x_n^T - \end{bmatrix}$ w/ p LI columns. Then $\hat{w}_{LS} = (X^T X)^{-1} X^T y$

These assumptions were necessary to ensure there was a unique set of weights minimizing squared error — and that $X^T X$ had an inverse.

Let $X = U \Sigma V^T \implies X^T X = (U\Sigma V^T)^T (U \Sigma V^T) = V \Sigma^T U^T U \Sigma V^T = V \Sigma^T \Sigma V^T$

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

Now $\Sigma^T \Sigma = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ & & \ddots \\ 0 & & & \sigma_p^2 \end{bmatrix}$, so we can only invert $\Sigma^T \Sigma$ if all the σ_i's > 0

That is,

$$\boxed{X \text{ has } p \text{ LI columns} \iff X^T X \text{ invertible} \iff X \text{ is positive definite} \iff \text{all singular values} > 0}$$

If X has $r < p$ LI columns, then $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 = \sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_p$

> Can we still learn a predictor when X has $r < p$ LI columns?
> e.g., if $n < p$ (more features than training samples), then there are at most n LI cols

Recall ridge regression:

$$\hat{\underline{w}}_\lambda = \underset{\underline{w}}{\operatorname{argmin}} \ \|\underline{y} - X\underline{w}\|_2^2 + \lambda \|\underline{w}\|_2^2 = (X^TX + \lambda I)^{-1} X^T \underline{y}$$

Let $X = U\Sigma V^T$, as before.

Then
$$X^TX + \lambda I = V\Sigma^T\Sigma V^T + \lambda I = V\Sigma^T\Sigma V^T + \lambda \underbrace{VV^T}_{I} = V\Sigma^T\Sigma V^T + V(\lambda I)V^T$$

$$= V(\Sigma^T\Sigma + \lambda I)V^T$$

$$\Sigma^T\Sigma = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_p^2 \end{bmatrix} \Rightarrow \Sigma^T\Sigma + \lambda I = \begin{bmatrix} \sigma_1^2+\lambda & & & 0 \\ & \sigma_2^2+\lambda & & \\ & & \ddots & \\ 0 & & & \sigma_p^2+\lambda \end{bmatrix}$$

Even if $\sigma_p = 0$,
$\sigma_p^2 + \lambda > 0$

$\Rightarrow \Sigma^T\Sigma$ may not be invertible but $\Sigma^T\Sigma + \lambda I$ is <u>always</u> invertible

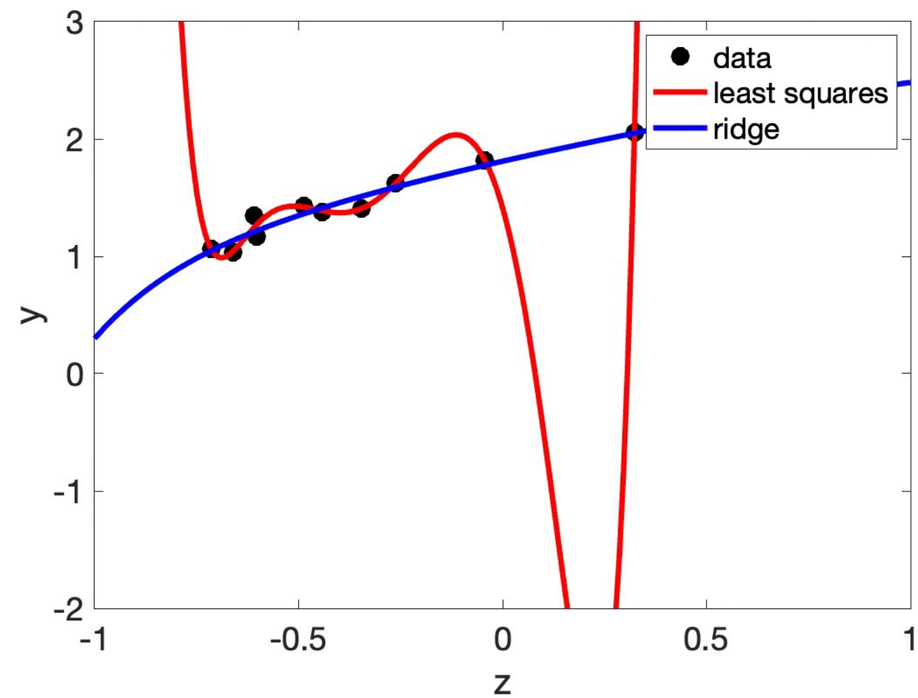
$\Rightarrow$ We can <u>always</u> compute the ridge regression estimate, even when a unique least squares estimate does not exist

$$\hat{\underline{w}}_\lambda = (X^TX + \lambda I)^{-1} X^T y$$

$$= V(\Sigma^T\Sigma + \lambda I) V^T V \Sigma^T U^T y$$

$$= V \underbrace{(\Sigma^T\Sigma + \lambda I)^{-1}\Sigma^T}_{\begin{bmatrix} \frac{\sigma_1}{\sigma_1^2+\lambda} & & 0 \\ & \frac{\sigma_2}{\sigma_2^2+\lambda} & \\ & & \ddots \\ 0 & & \frac{\sigma_p}{\sigma_p^2+\lambda} \end{bmatrix}} U^T y$$

$\boxed{\text{Ridge regression can help prevent overfitting}}$



WHY?

Imagine X has some very small singular values, e.g. 10^{-6}.

Also, imagine $\underline{y} = X\underline{w}^* + \underline{\varepsilon}$ where $\underline{\varepsilon}$ is random error or noise added to each sample

Then $\hat{\underline{w}}_{LS} = (X^TX)^{-1}X^T\underline{y} = (X^TX)^{-1}X^T(X\underline{w}^* + \underline{\varepsilon})$

$= \cancel{(X^TX)^{-1}X^TX}\,\underline{w}^* + (X^TX)^{-1}X^T\underline{\varepsilon}$

$$= \underbrace{\underline{w}^*}_{\substack{\text{good!} \\ \text{what we want}}} + V^T\underbrace{(\Sigma^T\Sigma)^{-1}\Sigma^T}_{\begin{bmatrix} 1/\sigma_1 & & & \\ & 1/\sigma_2 & & \\ & & \ddots & \\ & & & 1/\sigma_p \end{bmatrix}}U^T\underline{\varepsilon}$$

$\sim 10^6 \Rightarrow$ a little noise in $\underline{\varepsilon}$ gets multiplied by a huge number $\Rightarrow$ least squares might fit observations very closely but give strange predictions on test data

In contrast, $\hat{\underline{w}}_R = (X^TX + \lambda I)^{-1}X^T(X\underline{w}^* + \underline{\varepsilon})$

$= (X^TX + \lambda I)^{-1}X^TX\underline{w}^* + (X^TX + \lambda I)^{-1}X^T\underline{\varepsilon}$

$$= V\underbrace{(\Sigma^T\Sigma + \lambda I)^{-1}\Sigma^T\Sigma}_{\begin{bmatrix} \frac{\sigma_1^2}{\sigma_1^2+\lambda} & & \\ & \ddots & \\ & & \frac{\sigma_p^2}{\sigma_p^2+\lambda} \end{bmatrix}}V^T\underline{w}^* + V(\Sigma^T\Sigma + \lambda I)^{-1}\Sigma^TU^T\underline{\varepsilon}$$

$\approx \underline{w}^*$ for small λ

$$\frac{\sigma_i}{\sigma_i^2 + \lambda} \approx \begin{cases} 1/\sigma_i & \sigma_i \gg \lambda \\ 1/\lambda & \sigma_i \approx 0 \end{cases}$$

none of these values "blow up" by being close to dividing by 0, so we do not magnify noise $\Rightarrow$ better predictions on test data

The Elephant In the Room – how do we choose λ?

1. Split data into 2 sets: $(\underline{x}_i, y_i),\ i = 1, \dots, m$ = training set

$(\underline{x}_i, y_i),\ i = m+1, \dots, n$ = validation set.

2. For each $\lambda \in \{\lambda_1, \lambda_2, \dots, \lambda_g\}$, find $\underline{\hat{w}}_\lambda$ using training data

Measure the loss of each λ using validation set: $L_\lambda = \sum_{i=m+1}^{n} (y_i - \underline{x}_i^T \underline{\hat{w}}_\lambda)^2$

Choose λ with smallest L_λ.

Alternative to ridge regression (when some singular values = 0 so many $\underline{w}$ fit y perfectly)

with least squares, we needed to compute $(\Sigma^T\Sigma)^{-1}\Sigma^T = \begin{bmatrix} 1/\sigma_1 & & & \\ & 1/\sigma & & \\ & & \ddots & \\ & & & 1/\sigma_p \end{bmatrix}$,

which was problematic for any $\sigma_i = 1$

Alternative: define pseudoinverse of Σ as

$$(\Sigma^+)_{ii} = \begin{cases} 1/\sigma_i & \text{if } \sigma_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

that is, Σ^+ corresponds to taking the transpose of Σ and only inverting the nonzero diagonal entries

Special case: $p > n$, X has n linearly independent rows.

$\Rightarrow \Sigma \sim n \times p$

$$\underset{n\times p}{X} = \underset{n\times n}{U}\;\underset{n\times p}{\Sigma}\;\underset{p\times p}{V^T}$$

$$\Sigma = \Big[\underbrace{\begin{matrix}\sigma_1 & & \\ & \sigma_2 & \\ & & \ddots \\ & & & \sigma_n\end{matrix}}_{\text{n nonzero singular vals.}} \;\Big|\; \underbrace{\begin{matrix} 0 \end{matrix}}_{\text{p-n columns of zeros.}}\Big] \Rightarrow \Sigma^+ = \begin{bmatrix} 1/\sigma_1 & & & \\ & 1/\sigma_2 & & \\ & & \ddots & \\ & & & 1/\sigma_n \\ \hline & & 0 & \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \left.\right\} \text{p-n rows of zeros} \end{matrix} = \Sigma^T(\Sigma\Sigma^T)^{-1}$$

pseudoinverse solution $\hat{\underline{w}} = V\Sigma^+ U^T y$

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

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

Claim: when X has n LI rows, the choice $\hat{w} = V\Sigma^+ U^T y = X^T(XX^T)^{-1} y$

has the smallest norm $\|w\|_2^2$ of any w satisfying $y = Xw$.

Proof: for any w,

$$\|w\|_2^2 = \|w - \hat{w} + \hat{w}\|_2^2 = \|w - \hat{w}\|_2^2 + 2(w-\hat{w})^T\hat{w} + \|\hat{w}\|_2^2$$

Assume $Xw = y = X\hat{w} \Rightarrow X(w-\hat{w}) = 0$.

Then $(w-\hat{w})^T\hat{w} = (w-\hat{w})^T X^T (XX^T)^{-1} y$

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

$= 0$

$\Rightarrow \|w\|_2^2 = \|w - \hat{w}\|_2^2 + \|\hat{w}\|_2^2 \geq \|\hat{w}\|_2^2 \Longrightarrow \hat{w}$ has smallest norm!

Why does minimum norm make sense?

e.g. $X = \begin{bmatrix} 1 & 0 & 0.1 \\ 0 & 1 & 0 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

possible $\underline{w}$'s include $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 10 \end{bmatrix}, \begin{bmatrix} -10 \\ 0 \\ 100 \end{bmatrix}, \begin{bmatrix} -1000 \\ 0 \\ 10000 \end{bmatrix}$

minimum norm!