Regularization and Ridge Regression for Supervised Learning

Objectives

- Extend supervised learning to rank deficient $\underline{A}$

- Introduce ridge regression and regularization to control bias/variance tradeoff

- Solve ridge regression problem

Supervised Learning Solves Linear Equations

Classification/Modeling: $\underset{(N\times P)}{\underline{A}}\ \underset{(P\times 1)}{\underline{w}} \approx \underset{(N\times 1)}{\underline{d}}$ $\quad \underline{x}_i^T \underline{w} \simeq d_i$ (feature, model, label)

Least Squares Solution $\min_{\underline{w}} \|\underline{A}\underline{w} - \underline{d}\|_2^2$

Overdetermined system: $N \geq P$, $\text{rank}(\underline{A}) = P$

$\underline{w} = (\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{d}$ $\quad \underline{A}^T\underline{A}$ invertible

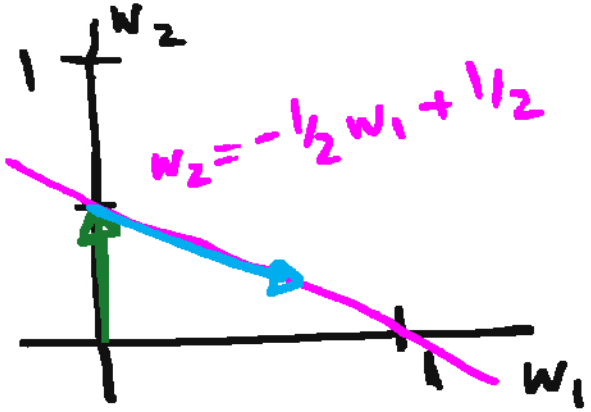
Now ☆ Underdetermined system: $N < P$ or $\text{rank}(\underline{A}) < P$

No unique solution $\underline{w}$!

Example: $N=1, P=2$ $\quad \underline{A} = [1 \ \ 2], d=1$

$\underline{A}\underline{w} = d \Rightarrow w_1 + 2w_2 = 1 \Rightarrow w_2 = -\frac{1}{2}w_1 + \frac{1}{2}$

infinite # solns zero error $\quad \underline{w} = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} + \alpha \begin{bmatrix} 1 \\ -0.5 \end{bmatrix}$ (Infinite w)



How do we choose a unique solution?

if $\text{rank}(A) < P$ and $Aw_0 = d$, then $Aw = d$ for $w = w_0 + Vb$

where cols(V) span space $\perp$ span$\{A\}$ $(AV = 0)$

Squared error is not sufficient to give unique solution

Consider robustness to errors/noise ε

$$x_i^T w = d_i \Rightarrow (x_i + \varepsilon)^T w = d_i + \varepsilon^T w$$

$\varepsilon^T w$: "variance" term

$$|\varepsilon^T w|^2 = \|w\|_2^2 \, \|\varepsilon\|_2^2 \cos^2\theta_{\varepsilon w}$$

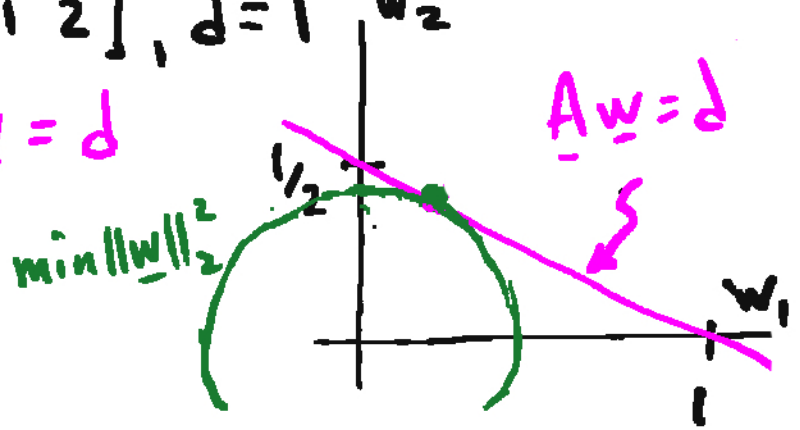
no info on $\varepsilon \Rightarrow \min \|w\|_2^2$

Example: $A = [1 \; 2]$, $d = 1$

$\min_w \|w\|_2^2$ s.t. $Aw = d$

$\min_{w_2} 5w_2^2 - 4w_2 + 1$

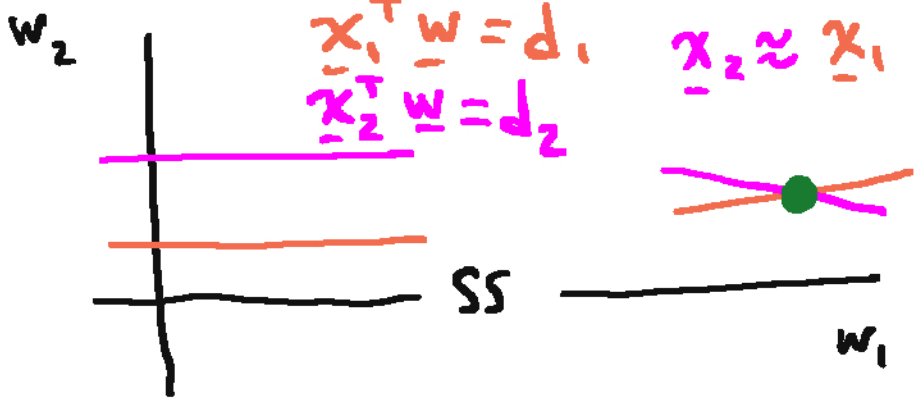
$\Rightarrow w = \begin{bmatrix} .2 \\ .4 \end{bmatrix}$



Similar features $\Rightarrow$ large $\|w\|_2^2$

$x_1^T w = d_1$, $x_2 \approx x_1$

$x_2^T w = d_2$



Previously, we had $A\underline{w} = \underline{d} + \underline{\varepsilon}$

$\downarrow$

$\min_{\underline{w}} \|A\underline{w} - \underline{d}\|^2$ for minimizing the squared error

Problem: there are multiple $\underline{w}$ to let $\|A\underline{w} - \underline{d}\|^2 = 0$, we cannot choose a single solution to the least-squared error problem

$\downarrow$

Solution: Connect ε with the size of $\underline{w}$ (不再考虑 d 中的 noise ⟶ 假设 x 里有 noise)

$\underline{\varepsilon}\underline{w}$ as the variance term, if $\underline{w} \uparrow$, the $\underline{\varepsilon}\underline{w} \uparrow$

$\|\varepsilon^T \underline{w}\|^2 = \|\underline{w}\|^2 \|\underline{\varepsilon}\|^2 \cos^2\theta$

$\therefore$ we need to minimize $\|\underline{w}\|^2$

$\searrow$

Among the $\underline{w}$ that let the least-squared error be the minimum, choose one that is the smallest in its norm

Choose $\underline{w}$ to balance bias and variance

Consider $(\underline{x}_i + \underline{\varepsilon})^T \underline{w} - d_i = \underbrace{\underline{x}_i^T \underline{w} - d_i}_{\text{bias}} + \underline{\varepsilon}^T \underline{w}$ (variance term)

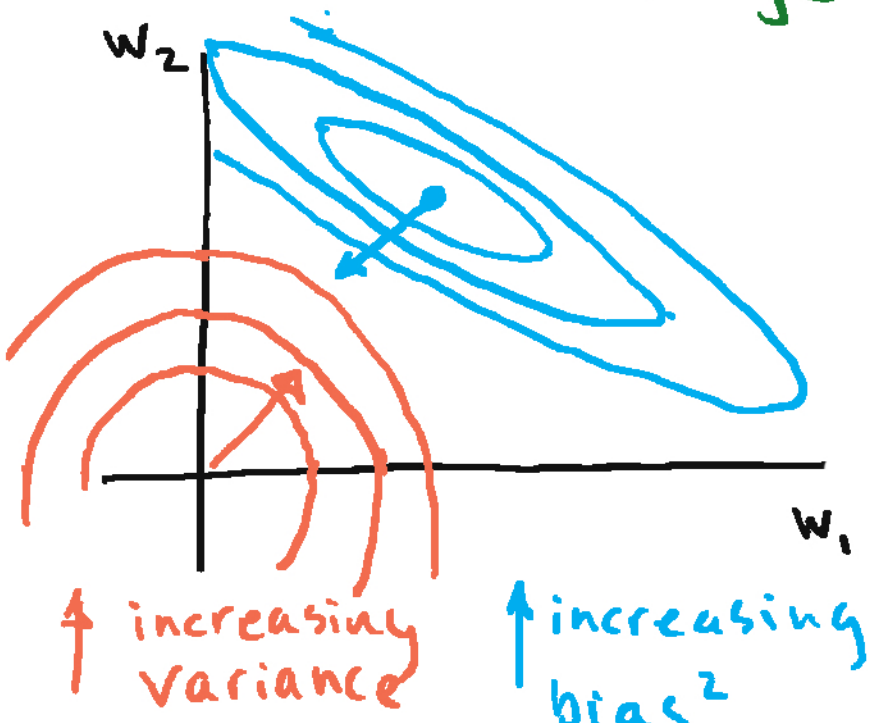
$$\min_{\underline{w}} \underbrace{\|\underline{A}\underline{w} - \underline{d}\|_2^2}_{\text{bias}^2} + \underbrace{\lambda \|\underline{w}\|_2^2}_{\text{variance}}$$ (山坡函数)

ridge regression

$\|\underline{w}\|_2^2$ - regularizer

λ - regularization parameter

(Tikhonov regularized)



Solution:

$$\|\underline{a}\|_2^2 + \|\underline{b}\|_2^2 = \underline{a}^T\underline{a} + \underline{b}^T\underline{b} = \left\| \begin{bmatrix} \underline{a} \\ \underline{b} \end{bmatrix} \right\|_2^2$$

$$\min_{\underline{w}} \left\| \begin{bmatrix} \underline{A}\underline{w} - \underline{d} \\ \lambda^{1/2}\underline{w} \end{bmatrix} \right\|_2^2 \rightarrow \min_{\underline{w}} \left\| \underbrace{\begin{bmatrix} \underline{A} \\ \lambda^{1/2}\underline{I} \end{bmatrix}}_{\tilde{\underline{A}}} \underline{w} - \underbrace{\begin{bmatrix} \underline{d} \\ 0 \end{bmatrix}}_{\tilde{\underline{d}}} \right\|_2^2$$

用 least-square 的套

$$\underline{w} = (\tilde{\underline{A}}^T\tilde{\underline{A}})^{-1}\tilde{\underline{A}}^T\tilde{\underline{d}} = \left(\begin{bmatrix} \underline{A}^T & \lambda^{1/2}\underline{I} \end{bmatrix} \begin{bmatrix} \underline{A} \\ \lambda^{1/2}\underline{I} \end{bmatrix} \right)^{-1} \underline{A}^T\underline{d}$$

$$= (\underline{A}^T\underline{A} + \lambda \underline{I})^{-1}\underline{A}^T\underline{d}$$

$\underline{A}^T\underline{A} + \lambda\underline{I} > 0$ $(\lambda > 0)$ → inverse exists

☆ Full rank matrix (Every column has 1 differently)

Ridge regression results in a robust solution5

Suppose $\underline{d} = \underline{A}\underline{w} + \underline{n}$ $\underline{n}$: model error or label error

Least Squares: $\underline{w} = (\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{d} = (\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{A}\underline{w} + (\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{n}$

$= \underline{w} + (\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{n}$ ← variance

Ridge: $\underline{w} = (\underline{A}^T\underline{A}+\lambda\underline{I})^{-1}\underline{A}^T\underline{d} = (\underline{A}^T\underline{A}+\lambda\underline{I})^{-1}\underline{A}^T\underline{A}\underline{w} + (\underline{A}^T\underline{A}+\lambda\underline{I})^{-1}\underline{A}^T\underline{n}$

bias variance

LS is problematic (large variance term) when $\underline{A}$ is nearly low rank

Example:

$$\underline{A} = \frac{1}{2}\begin{bmatrix} 1 & 1 & .001 \\ 1 & -1 & .001 \\ 1 & 1 & -.001 \\ 1 & -1 & -.001 \end{bmatrix}$$

or thog cols, rank$(\underline{A})$=3

$$\underline{A}^T\underline{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10^{-6} \end{bmatrix} \rightarrow (\underline{A}^T\underline{A})^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10^{6} \end{bmatrix} \rightarrow \text{large variance!}$$

$$(\underline{A}^T\underline{A}+\lambda\underline{I})^{-1} = \begin{bmatrix} \frac{1}{1+\lambda} & 0 & 0 \\ 0 & \frac{1}{1+\lambda} & 0 \\ 0 & 0 & \frac{1}{10^{-6}+\lambda} \end{bmatrix} \approx \begin{bmatrix} \frac{1}{1+\lambda} & 0 & 0 \\ 0 & \frac{1}{1+\lambda} & 0 \\ 0 & 0 & \frac{1}{\lambda} \end{bmatrix}$$

→ λ controls variance

Use Cross validation to choose λ

1) Choose $\Lambda = \{\lambda_1, \lambda_2, \cdots \lambda_m\}$ (often logarithmic spacing)

2) Split data: training (t) and validation (v) sets

$$\underline{A}^{(t)}, \underline{d}^{(t)}, \underline{A}^{(v)}, \underline{d}^{(v)}$$

3) Compute $w_{\lambda_i} = \left(\underline{A}^{(t)T}\underline{A}^{(t)} + \lambda_i \underline{I}\right)^{-1} \underline{A}^{(t)T}\underline{d}^{(t)}$

4) Compute $e^2(\lambda_i) = \| \underline{A}^{(v)} \underline{w}_{\lambda_i} - \underline{d}^{(v)} \|_2^2$

(average over different training/validation partitions)

5) Choose λ_i that minimizes $e^2(\lambda_i)$