Lecture 13 - Ridge Regression

Variable Selection

→ All procedures we've looked at allow for some set of p features (regr. or class.)

→ May want to: select a "best" set of features

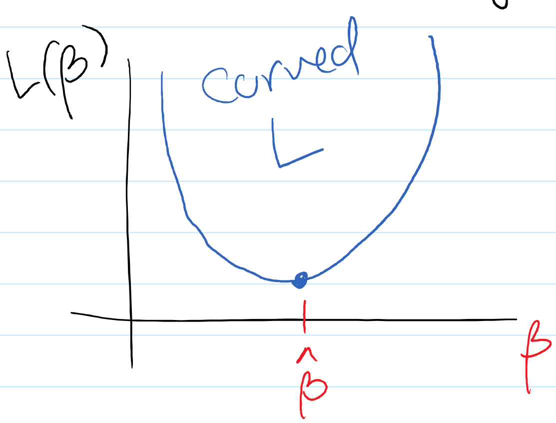
1. prediction accuracy (bias ↑ and var ↓ by p ↓) ⊛
2. interpretation
3. model may be <u>ill-conditioned</u> ($p \gg N$)

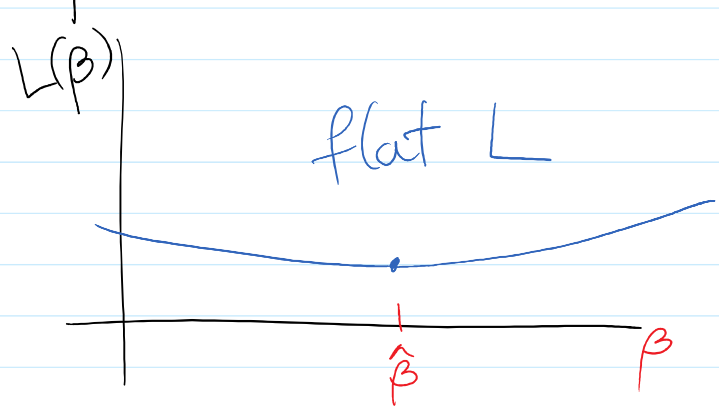
<u>note</u>: apply to class. and regr

Review of LS Regr.

→ $\hat{\beta}$ comes from solving $(X^TX)\beta = X^TY$

so stability of $\hat{\beta}$ depends on $K(X^TX)$





$K(X^TX)$ small $\qquad$ $K(X^TX)$ large

<u>Simple illustration:</u>

Consider $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$

<u>Q</u>: what happens if $X_1 \approx X_2$.

<u>our model:</u>

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$$
$$= \beta_0 + \underbrace{(\beta_1 + \beta_2)}_{} X_1 + \cdots$$

if $\hat{\beta}_1 = 5$ and $\hat{\beta}_2 = 10$

then just as good of a fit w/ $\hat{\beta}_1 = 7$, $\hat{\beta}_2 = 8$

b/c $\hat{\beta}_1 + \hat{\beta}_2 = 15$ still.

Generally any $\hat{\beta}_1 + \hat{\beta}_2 = 15$ gives same fit

eg. $\hat{\beta}_1 = -1000$, $\hat{\beta}_2 = 1015$

So when $K(X^TX)$ we tend to have problems when $\hat{\beta}$s run off to $\pm\infty$.

Variable Selection: choose X_1 and exclude X_2 from our fit. Fix $K(X^TX)$

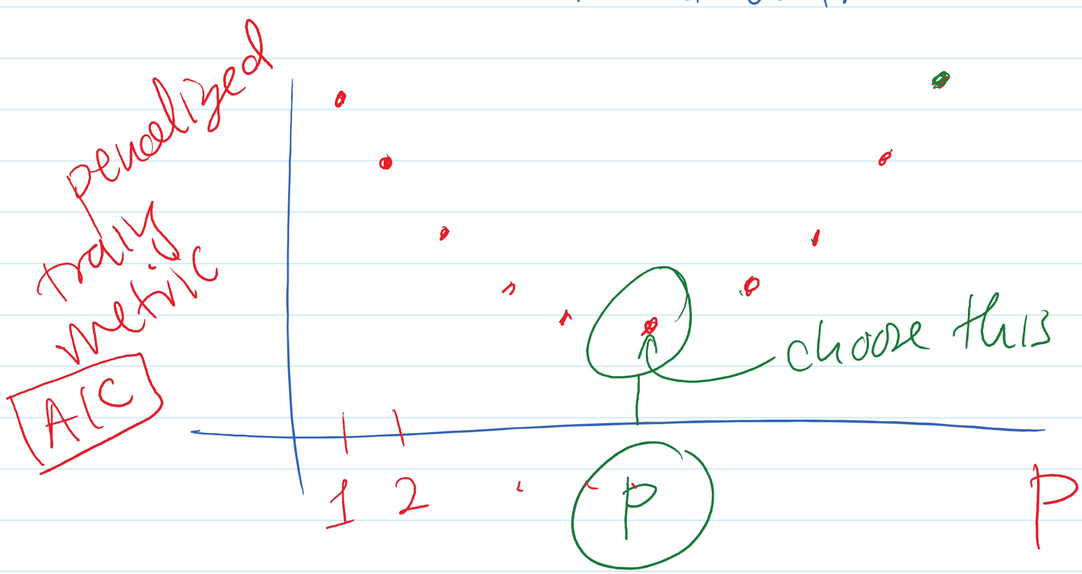
Idea: try a bunch of models w/ different subsets of variables and choose best model.

Way 1: penalize training metric by P and use this to choose P

$\boxed{\text{Classic very fast}}$

ex. forward stepwise regression
→ add "best" variables 1 at a time

ex. backwards stepwise
→ start w/ all, remove "worst" var. 1 at a time



way 2: X-validation
(train/validate/test)

[modern
slower]

Note: typical testing all possible subsets of vars is not feasible.

We not want to test all possible subsets

Remember: how well does our model building procedure work?

① testing all possible subsets is a very flexible building procedure
(high flex - low bias high var)

② using a constrained/simpler search gives a less flexible model building procedure
(low flex - higher bias lower var)

If our vars are standardized and I get $\hat{\beta}$s order by size (abs)

$$|\hat{\beta}|_{(1)}, |\hat{\beta}|_{(2)}, \Big| \ldots, |\hat{\beta}|_{(p)}$$

$|\hat{\beta}|_{(1)}, |\hat{\beta}|_{(2)}, | \dots, |\hat{\beta}|_{(p)}$

largest $\qquad t \qquad$ smallest

$\qquad\qquad$ threshold

Do var selection by incl. only covs w/

$$|\hat{\beta}_{(i)}| \geq t$$

Hard-Thresholding

$$\hat{\beta}_i^{(HS)} = \begin{cases} \hat{\beta}_i & \text{if } |\hat{\beta}_i| \geq t \\ 0 & \text{else} \end{cases}$$

Ridge Regression:

Q: can we deal w/ ill conditioning in a "continuous" way?

Recall: For OLS (ordinary Least-Squares)

$$L(\beta) = \|y - X\beta\|^2 \text{ and}$$

$$\underbrace{\hat{\beta}^{(OLS)} = \arg\min_{\beta} L(\beta)}$$

Ridge Regression: slightly alter OLS to penalize if β is too large.

Interp 1 of Ridge

$$\hat{\beta}^{(Ridge)} = \underset{\beta}{\arg\min}\ L(\beta) + \lambda \overbrace{\|\beta\|^2}^{\text{size of } \beta}$$

$\lambda \geq 0$ penalty parameter

typically we don't penalize β_0

$\|\beta\|^2 = \sum_{j=1}^{p} \beta_j^2$

λ = how much we penalize big βs

$\lambda = 0$ no penalty

$\hat{\beta}^{(ridge)} = \hat{\beta}^{(OLS)}$

$\lambda \to \infty$, $\hat{\beta} \to 0$

(Secret: choose λ by X-validation)

Ridge Interp 2

Recall Calc III (Lagrange Multipliers)

(*) $\min_x f(x)$ s.t. $g(x) \leq t$ [constrained optim]

Solved w/ Lagrange multipliers

$$\mathcal{L}(\lambda, x) = f(x) + \lambda(g(x) - t)$$

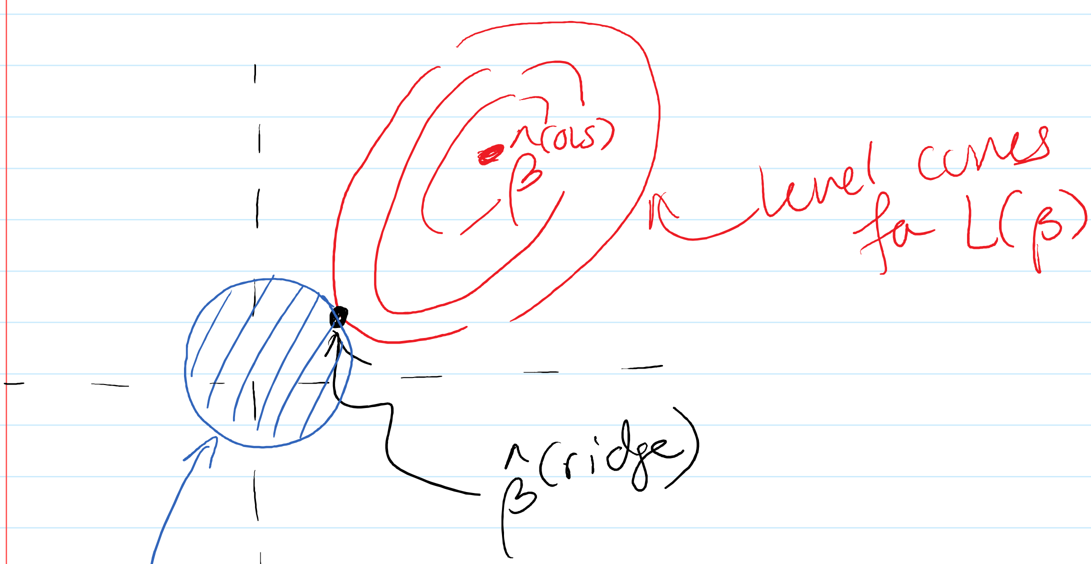
and (*) equiv to

$\min \mathcal{L}(\lambda, x)$ for some λ corresp. to t

Ridge equiv. formulation

$$\hat{\beta}^{(ridge)} = \underset{\beta}{\operatorname{argmin}}\, L(\beta) \text{ s.t. } \|\beta\|^2 \le t$$

corresp. to λ



Interp 3 Note that $L(\beta) = \|y - X\beta\|^2$

Interp 3 Note that $L(\beta) = \|y - X\beta\|^2$ and $\|\beta\|^2$ are both quadratic, and differentiable.

For OLS: this means closed form soln

$$\hat{\beta}^{(OLS)} = (X^TX)^{-1}X^Ty \qquad \left[(X^TX)\beta = X^Ty\right]$$

Ridge: also has a closed form soln

$$\hat{\beta}^{(ridge)} = (X^TX + \lambda I)^{-1}X^Ty \qquad \left[(X^TX + \lambda I)\beta = X^Ty\right]$$

replaced X^TX w/ $X^TX + \lambda I$

For OLS: conditioning depended on $K(X^TX)$

Ridge: conditioning depends on $K(X^TX + \lambda I)$

X^TX may not be well cond.

but $X^TX + \lambda I$ <u>may</u> be better conditioned

Decompose $X = UDV^T$. (SVD)

OLS

$$\to X^TX = VDU^TUDV^T = VD^2V^T$$

$$\to (X^TX)^{-1} = VD^{-2}V^T$$

$$\to (X^TX)^{-1}X^T = VD^{-2}V^TVDU^T$$
$$= VD^{-1}U^T$$

$$\to \hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY = UDV^TVD^{-1}U^TY$$
$$= UU^TY$$
$$= \sum_j \underbrace{u_j u_j^T Y}_{\text{sum of } Y \text{ proj. onto } u_j}$$

cols of U basis of col(X)

Re-do for Ridge

$$\hat{y}^{(ridge)} = X\hat{\beta}^{(ridge)} = X(X^TX+\lambda I)^{-1}X^TY$$

$(AB)^{-1} = B^{-1}A^{-1}$

$$= X(VD^2V^T+\lambda I)^{-1}X^TY$$
$$= UDV^T(VD^2V^T+\lambda I)^{-1}VDU^TY$$
$$= UD(V^T(VD^2V^T+\lambda I)V)^{-1}DU^TY$$
$$= UD\underbrace{(D^2+\lambda I)^{-1}}_{\text{diagonal}}DU^TY$$

diagonal

$D = \mathrm{diag}(\sigma_i)$

$\mathrm{diag}(\sigma_i^2 + \lambda)$

inverse $\mathrm{diag}\left(\frac{1}{\sigma_i^2+\lambda}\right)$

$$= U \,\mathrm{diag}\left(\frac{\sigma_i^2}{\sigma_i^2+\lambda}\right) U^T Y$$

$$\boxed{= \sum_j \left(\frac{\sigma_i^2}{\lambda+\sigma_i^2}\right) u_j u_j^T Y}$$

weighted sum of proj. of Y onto u_j

$$\frac{\sigma_i^2}{\lambda+\sigma_i^2} \le 1 \qquad \text{(Shinkage)}$$

Shrink directions w/ lower σ_i more.

Var sel