ordinary least-squares regression

$Y = X\beta + \varepsilon$

Y is $n \times 1$ vector of results

β is a $d \times 1$ vector of regression coefficients

X is $n \times d$ matrix of predictors

ε is $n \times 1$ vector of errors

$$\hat{\beta}_{OLS} = \min \sum_{i=1}^{n} (Y_i - X_i \beta)^2$$

$$= (Y - X\beta)^T (Y - X\beta)$$

dimension of β starts to get large,

or we start to see large correlation in $X^T X$

$(X^T X)^{-1}$ is ill-conditioned

our estimates of β start to get very noisy and don't work well at prediction

$\hat{\beta} \sim N(\beta, (X^T X)^{-1} \sigma^2)$ σ^2 is the variance of the residuals

$\beta = (0.5, 0.5)$

$\hat{\beta} = (3.7, -2.4)$

minimize $(Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta$
β

↳ $\sum \beta_i^2$

How do we pick λ?

$(\vec{X}_1, Y_1), \ldots, (\vec{X}_n, Y_n)$

↳ split my data into a test set and a train set

test set is about 10-20% of the data

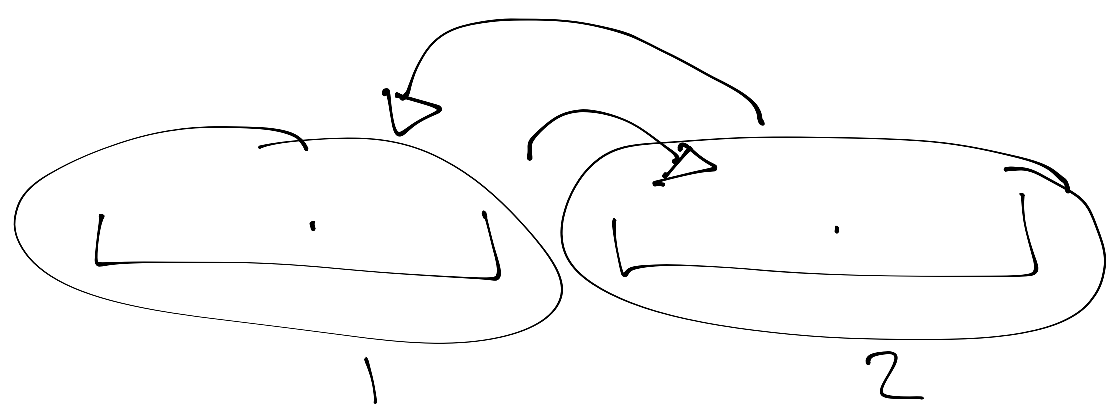
↳ look at a large grid of values of λ [.01, .02, .04, .08, .16]

↳ cross-validation (k-folds)

↳ randomly partition my train set into k parts folds

↳ going to my model for each λ in my grid k times.

5-fold cross-validation



↳ choose the lambda that gives me the best out of sample performance

$\lambda = [1, 2]$

fold 1

fold 2

m_{ij} = model fit using lamba[i] fold[j]

m_{11} m_{12}

m_{21} m_{22}

predict

fit m_{11} on fold 1 $\rightarrow$ fold 2

$\mathcal{E}_{11}$ is going to be prediction on fold 2 of m_{11}

avg prediction error $\lambda = 1$

$$\left(\frac{\mathcal{E}_{11} + \mathcal{E}_{12}}{\#\text{ of samples}} \right.$$

$$\lambda = 2$$

$$\left\{ \frac{\mathcal{E}_{21} + \mathcal{E}_{22}}{\#\text{ of samples}} \right.$$

$\rightarrow$ let's $\lambda = 1$ performs better, then I fit the model using $\lambda = 1$ on the entire training set, and validate it's performance on the test set

$$\hat{y}_i = X_i \hat{\beta}$$

$$R^2 = 1 - \frac{\Sigma (y_i - \hat{y})^2}{\Sigma (y_i - \bar{y})^2}$$

general question is how to choose k?

sklearn default set 5

ridge regression

$$\underset{\beta}{\text{minimize}} \quad (Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta$$

$\hookrightarrow \sum \beta_i^2$

LASSO

$$\underset{\beta}{\text{minimize}} \quad (Y - X\beta)^T (Y - X\beta) + \lambda \sum_{i=1}^{n} |\beta_i|$$

$\Rightarrow$ not a major difference th predictive abilities of LASSO/ridge

$\Rightarrow$ lasso is slower to fit

$\Rightarrow$ lasso can perform variable selection

$\lambda = \text{BIG}$