Last lecture: logistic regression
touched overfitting

Today: overfitting & regularization

simplest regression

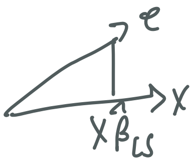
1		
⋮		
i	x_i	y_i
⋮		
n		

model: $y_i = x_i\beta + e_i \quad e_i \sim N(0,\sigma^2)$ iid

hypothesis testing: $H_0: \beta = 0$ simpler model
$H_1: \beta \neq 0$ more complex

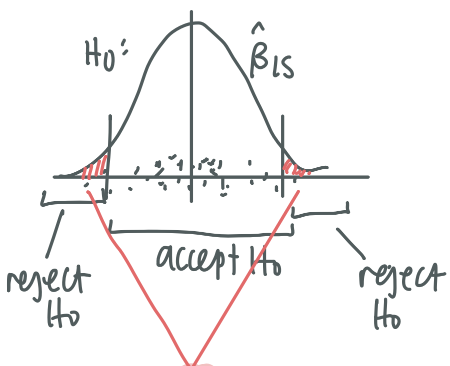
$$\hat\beta_{LS} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \overset{H_0}{=} \frac{\sum_{i=1}^n x_i e_i}{\sum_{i=1}^n x_i^2} \neq 0$$

there is always a certain amount of overfitting in $\hat\beta$ even if true $\beta = 0$



overfitting: model absorbs noise

ways to avoid overfitting
1. hypothesis testing



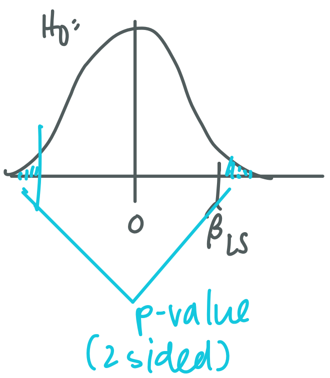
let t = threshold

$$\hat\beta_{\substack{\text{hypothesis}\\\text{testing}}} = \begin{cases} 0 & |\hat\beta_{LS}| < t \\ \hat\beta_{LS} & |\hat\beta_{LS}| > 0 \end{cases}$$

hard thresholding

error: α, usually taken as 5% (2.5% in each tail)

p-value: how extreme $\hat\beta_{LS}$ is relative to H_0



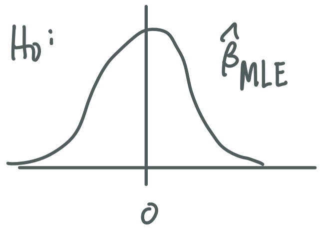
logistic regression

$H_0: \beta = 0 \Rightarrow s_i = x_i\beta = 0$

$y_i \sim \text{Bernoulli}(\frac{1}{2})$

nothing more than a coin flip

$$p_i = \frac{e^{s_i}}{1+e^{s_i}} = \frac{1}{2}$$



$\hat{\beta}_{MLE} \neq 0$

max likelihood estimate

noise in the logistic regression classification scenario is overinterpretation of coin flipping, as opposed to gaussian noise observed in linear regression

② Regularization, ML treatment

① L2 → ridge regression   keeps all covariates

simplest regression

$$L(\beta) = \underbrace{\sum_{i=1}^{n}(y_i - x_i\beta)^2}_{\text{ls term}} + \underbrace{\lambda\beta^2}_{\text{penalty term}}$$

penalty term — drags estimate of β towards 0 to help avoiding overfitting

$$L'(\beta) = -2\sum_{i=1}^{n}(y_i - x_i\beta)x_i + 2\lambda\beta$$

univariate scenario

$$= \sum_{i=1}^{n} y_i x_i + \sum_{i=1}^{n} x_i^2\beta + \lambda\beta = 0$$

$$\hat{\beta}_{ridge} = \frac{\sum x_i y_i}{\sum x_i^2 + \lambda}$$

shrinkage estimator;

interpolation btwn H_0 & H_1, as opposed to hard thresholding

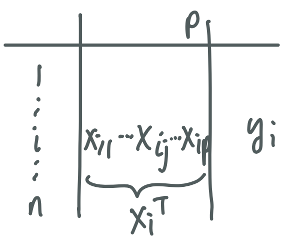
under H_0: $y_i = e_i$

adding λ will help w/ not absorbing as much noise (however, tradeoff is introduction of some bias)

$$\downarrow \lambda \to 0$$

$$\hat{\beta}_{LS} = \frac{\sum x_i y_i}{\sum x_i^2}$$

multivariate scenario



$$L(\beta) = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\|\beta\|_{L2}^2 \rightarrow \sum_{j=1}^{p}\beta_j^2$$

$$L'(\beta) = -2\sum_{i=1}^{n} x_i(y_i - x_i^T\beta) + 2\lambda\beta = 0$$

$$= -\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} x_i x_i^T\beta + \lambda\beta = 0$$

$$\Big(\underbrace{\boxed{\sum_{i=1}^{n} x_i x_i^T}}_{}^{p\times p} + \lambda I\Big)\beta = \sum_{i=1}^{n} x_i y_i$$

$$\hat{\beta}_{ridge} = \left(\sum x_i x_i^T + \lambda I\right)^{-1}\left(\sum x_i y_i\right)$$

$$= (\underline{\underline{X}}^T\underline{\underline{X}} + \lambda \underline{\underline{I}})^{-1}\underline{\underline{X}}^T Y$$

raw input	features p	
X_i	$\sin X_i, \cos X_i \cdots$	y_i
n	$p >> n$	

$y_i = \beta_{1,1}\sin X_i + \beta_{1,2}\cos X_i + \beta_{2,1}\sin 2X_i + \beta_{2,2}\cos 2X_i + \ldots + e_i$

$|\beta|^2_{\ell 2}$ small → smaller curve

otherwise, training error = 0 → model absorbs too much noise

② L1 regularization → Lasso

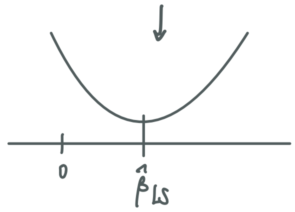
(least absolute shrinkage selection operator)

simplest regression

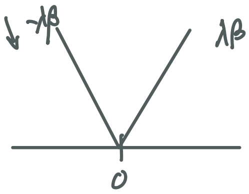
i	X_i	y_i

$$L(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda|\beta|$$

shrinkage + select



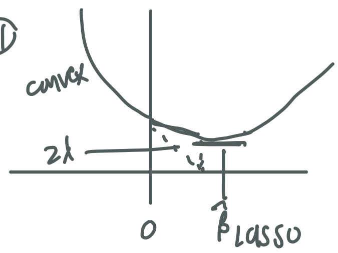
smooth, differentiable @ 0



sharp, not differentiable @ 0

Three scenarios:

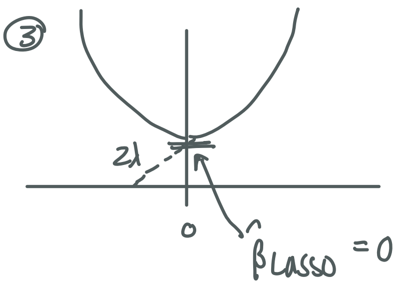
①



change in slope 2λ

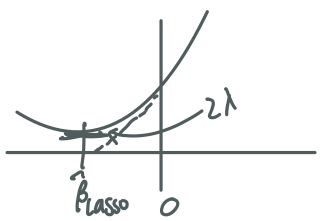
$\beta > 0$

③



accept H_0 in this scenario

②



change in slope 2λ

scenario ① $\hat{\beta}_{Lasso} > 0$ ⟶ $\hat{\beta}_{LS} > \frac{\lambda}{\Sigma X_i^2}$ for $\hat{\beta}_{Lasso} > 0$

$$L'(\beta) = -\Sigma(y_i - X_i\beta)X_i + \lambda = 0$$

$$= -\Sigma X_i y_i + \Sigma X_i^2\beta + \lambda = 0$$

$$\hat{\beta}_{Lasso} = \frac{\Sigma X_i y_i - \lambda}{\Sigma X_i^2} = \underbrace{\boxed{\frac{\Sigma X_i y_i}{\Sigma X_i^2}}}_{\hat{\beta}_{LS}} - \frac{\lambda}{\Sigma X_i^2} = \hat{\beta}_{LS} - \frac{\lambda}{\Sigma X_i^2}$$ shrinkage

scenario 2 $\quad \hat{\beta}_{Lasso} < 0 \longrightarrow \hat{\beta}_{LS} < \frac{-\lambda}{\sum x_i^2}$ for $\hat{\beta}_{Lasso} < 0$

$$L'(\beta) = -\sum (y_i - x_i\beta)x_i - \lambda = 0$$

$$= -\sum x_i y_i + \sum x_i^2 \beta - \lambda = 0$$

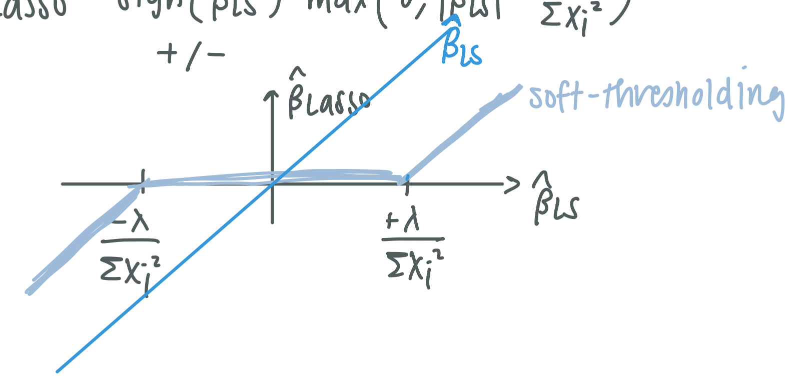
$$\hat{\beta}_{Lasso} = \frac{\sum x_i y_i + \lambda}{\sum x_i^2} = \underbrace{\boxed{\frac{\sum x_i y_i}{\sum x_i^2}}}_{\hat{\beta}_{LS}} + \frac{\lambda}{\sum x_i^2} = \hat{\beta}_{LS} + \frac{\lambda}{\sum x_i^2}$$

$$\frac{-\lambda}{\sum x_i^2} < \hat{\beta}_{LS} < \frac{\lambda}{\sum x_i^2}$$

$$\hat{\beta}_{Lasso} = 0$$

selection in scenario 3

$$\hat{\beta}_{Lasso} = \underset{+/-}{\text{sign}(\hat{\beta}_{LS})}\ \max\left(0, |\hat{\beta}_{LS}| - \frac{\lambda}{\sum x_i^2}\right)$$



multivariate scenario

$$L(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda \|\beta\|_{\ell 1} \searrow \sum_{j=1}^{p}|\beta_j|$$

conducts selection for covariate in final model as well as minimizing the βs

$$= \frac{1}{2}\Big|\underset{n\times 1}{Y} - \underset{n\times 1}{\sum_{j=1}^{p} x_j\beta_j}\Big|^2 + \lambda\|\beta\|_{\ell 1}$$

for k in 1 to p
fix β_j, $j \neq k$
solve β_k

coordinate descent

$$L(\beta_k) = \frac{1}{2}\Big|\underbrace{Y - \sum_{\substack{j=1 \\ j\neq k}}^{p} x_j\beta_j}_{\hat{Y}} - X_k\beta_k\Big|^2 + \lambda\sum_{j\neq k}|\beta_k| + \lambda|\beta_k|$$

$$|\hat{Y} - X_k\beta_k|^2 + \lambda|\beta_k|$$

start from big λ [βs starts at 0]
gradually reduce λ [βs increase]