Day 6, 09/25/2018

Statistical regularization: ridge regression

$y = X\beta + \varepsilon$, $Var(\varepsilon) = \sigma^2 I$

Recall: Normal equations $\hat{\beta} = \arg\min_{\beta} (Y - X\beta)^T (Y - X\beta)$

Analytical solution: $\frac{dJ(\beta)}{d\beta} = \frac{d}{d\beta}(Y - X\beta)^T(Y - X\beta) \overset{\text{set to}}{=} 0$

$X^T X\beta - X^T Y = 0 \Rightarrow \hat{\beta}_{ls} = (X^T X)^{-1} X^T Y$

($(X^T X)^{-1}$: singular if $N < p$ or collinear columns)

Properties of $\hat{\beta}$: $E\{\hat{\beta}_{ls}\} = E\{(X^T X)^{-1} X^T (X\beta + \varepsilon)\} = \beta$ (over all data)

$Var\{\hat{\beta}_{ls}\} = Var\{(X^T X)^{-1} X^T (X\beta + \varepsilon)\} = \sigma^2 (X^T X)^{-1} X^T I X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$

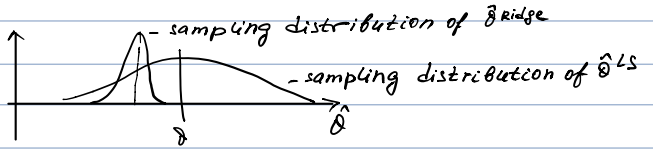
I.e.: - $\hat{\beta}$ is a linear combination of Y, is unbiased; $Var = \sigma^2 (X^T X)^{-1}$

- can show that has min variance among all linear unbiased estimators (Gauss-Markov theorem)

Can improve the estimation when $X^T X$ is (nearly) singular: define $\hat{\beta}_{Ridge} = (X^T X + \lambda I)^{-1} X^T Y$ (pre-defined constant $\lambda > 0$) - biased, but smaller variance } can invert

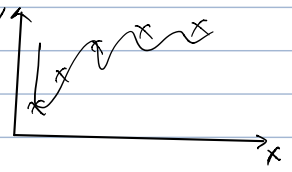
$E\{\hat{\beta}_{Ridge}\} = E\{(X^T X + \lambda I)^{-1} X^T (X\beta + \varepsilon)\} = (X^T X + \lambda I)^{-1} X^T X\beta + 0 \neq \beta$

$Var\{\hat{\beta}_{Ridge}\} = Var\{(X^T X + \lambda I)^{-1} X^T (X\beta + \varepsilon)\} = (X^T X + \lambda I)^{-1} X^T X \underbrace{(X^T X + \lambda I)^{-1}}_{\text{symmetric}} < \hat{\beta}_{LS}$



- the coefficients are smaller in absolute value $\Rightarrow$ **shrinkage estimates**

Intuition: polynomial regression



$h_\beta(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$

If we shrink β_3 and β_4 closer to 0 $\Rightarrow$ $h_\beta(x)$ is closer to quadratic

The same solution can be obtained using (1) penalized estimation (2) constrained optimization

(1) Penalized estimation:

$\hat{\beta} = \arg\min_{\beta} \underbrace{(Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta}_{\text{New loss function } J_\lambda(\beta)}$ — minimize SSE, penalize large $\hat{\beta}_j$

$\beta^T\beta = \|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$

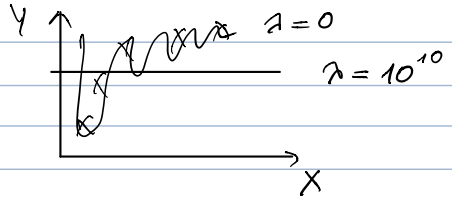
λ is a hyperparameter

Solution as function of β:

$\frac{d}{d\beta} J_\lambda(\beta) = \frac{d}{d\beta}\left[Y^T Y - \underbrace{(X\beta)^T Y}_{\beta^T X^T Y} - Y^T X\beta + \underbrace{(X\beta)^T (X\beta)}_{\beta^T X^T X \beta} + \lambda \beta^T \beta \right]$

$= -2X^T Y + 2X^T X\beta + 2\lambda\beta \overset{\text{set to}}{=} 0$

$\hat{\beta}_{Ridge} = (X^T X + \lambda I)^{-1} X^T Y$



(2) Constrained optimization (See CB Appendix E for reference)

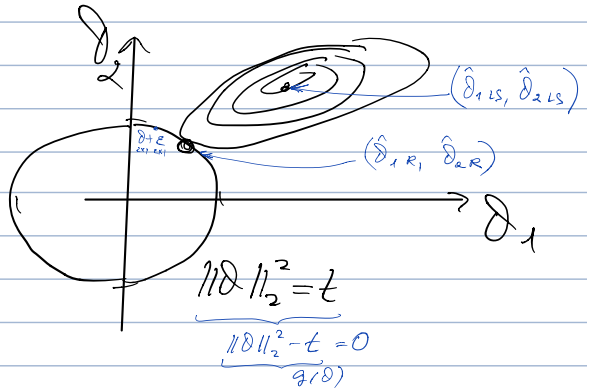
"budget"

Goal: find the stationary point of a function subject to a constraint

$$\min_{\theta} J(\theta), \quad \text{such that } \underbrace{\|\theta\|_2^2}_{g(\theta)} = t, \quad \text{equivalently, } \|\theta\|_2^2 - t = 0$$

θ: $p \times 1$; $\|\theta\|_2^2$: $p \times 1$; t: pre-defined constant $\neq 0$

At any point on the constraint surface $g(\theta)$,
$\nabla g(\theta) \perp g(\theta)$ (i.e., gradient is $\perp$ to surface)



From Taylor approximation around θ:

$$g(\theta+\varepsilon) \approx g(\theta) + (\theta+\varepsilon-\theta)^T \nabla g(\theta) = 0$$

- $g(\theta+\varepsilon)$: another point on the surface
- $g(\theta)$: on the surface, $= 0$
- ε: on the surface

$\Rightarrow$ for small ε, $\nabla g(\theta)$ is orthogonal to the surface

Next, look for point $\hat{\theta}$ on the surface, to minimize $J(\theta)$
At $\hat{\theta}$, $\nabla J(\theta)$ is also orthogonal to the surface (i.e, cannot decrease $J(\theta)$ by moving along the surface)

$\Rightarrow$ $\nabla J(\theta)$ and $\nabla g(\theta)$ are (anti) parallel vectors $\Rightarrow$ there exists λ s.t.

$$\nabla J(\theta) + \lambda \nabla g(\theta) = 0$$

λ: Lagrange multiplier, $\lambda \neq 0$ (any sign)

Lagrangian function: $L(\theta, \lambda) = J(\theta) + \lambda g(\theta)$

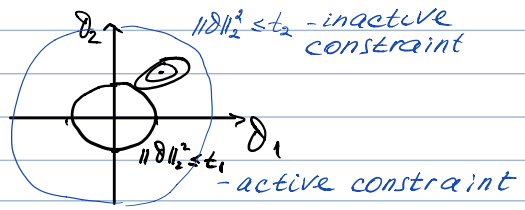
Stationary point: $\frac{dL(\theta,\lambda)}{d\theta} = \nabla J(\theta) + \lambda \nabla g(\theta) \overset{\text{set to}}{=} 0$ and $\frac{dL(\theta,\lambda)}{d\lambda} = g(\theta) \overset{\text{set to}}{=} 0$

- $\frac{dL(\theta,\lambda)}{d\theta} = 0$: p equations
- $\frac{dL(\theta,\lambda)}{d\lambda} = 0$: checks the constraint

Extension to inexact (i.e., inequality) constraints:

$$\min_{\theta} J(\theta), \quad \text{such that } \underbrace{\|\theta\|_2^2}_{g(\theta)} \leq t, \quad \text{or } \|\theta\|_2^2 - t \leq 0$$

θ: $p \times 1$



Inactive constraint:
- stationary point solves $\nabla J(\theta) = 0$ (i.e., $\lambda = 0$)

Active constraint:
- $L(\theta, \lambda) = J(\theta) + \lambda g(\theta)$, $\lambda > 0$
  - sign is important
  - gradient is oriented away from $g(\theta) \leq 0$ (anti-parallel vectors)
  - $\nabla J(\theta) = -\lambda \nabla g(\theta)$ for $\lambda > 0$

$\Rightarrow$

Solution of $\min_{\theta} J(\theta)$ s.t. $\underbrace{\|\theta\|_2^2 - t}_{g(t)} \leq 0$ is obtained by

minimizing $L(\theta, \lambda)$ subject to constraints

constraint:	$\|\theta\|_2^2 - t \leq 0$	Karush-Kuhn-Tucker (KKT) conditions
direction:	$\lambda \geq 0$	
active / inactive	$\lambda(\|\theta\|_2^2 - t) = 0$	i.e., optimize first; then check that the solution satisfies the KKT conditions

$\lambda = 0$ if inactive; $= 0$ at the optimum if active

Connection between (1) and (2):

(1) $\min_{\theta} J(\theta) + \lambda \|\theta\|_2^2$

(2) $\min_{\theta} J(\theta)$, s.t. $\|\theta\|_2^2 - t \le 0 \rightarrow L(\theta, \alpha) = J(\theta) + \alpha(\|\theta\|_2^2 - t)$

rename λ to α for the purpose of comparing two approaches

different from (1)

Conditions for optimum:

(1) $\nabla[J(\theta) + \lambda \|\theta\|_2^2] = 0$

(2) $\nabla[J(\theta) + \alpha(\|\theta\|_2^2 - t)] = 0$ and $\alpha(\|\theta\|_2^2 - t) = 0$ (KKT)

Let's say that solution of (1) is $\hat{\theta}(\lambda)$

Then (2) has the same solution wrt θ if $t = \|\hat{\theta}(\lambda)\|_2^2$

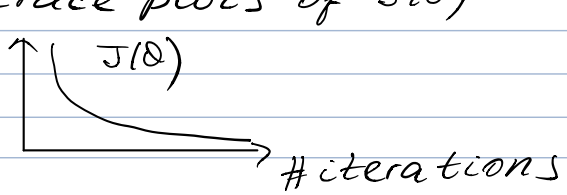
(i.e., λ and t are connected but not the same)

- $\nabla[J(\theta) + \lambda \|\theta\|_2^2]$ and $\nabla[J(\theta) + \alpha(\|\theta\|_2^2 - t)]$ are the same wrt $\theta \Rightarrow$ same optimum $\hat{\theta}(\lambda)$
- if $t = \|\hat{\theta}(\lambda)\|_2^2$, then the KKT condition verifies

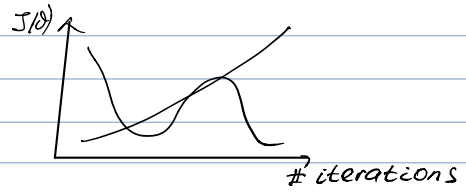
Parameter estimate by gradient descent: new loss $J(\theta) + \lambda \|\theta\|_2^2$ (replace the old loss function)

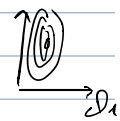
Practical considerations:

- scaling features (parameter values are comparable)
- centering features (no intercept)
- trace plots of $J(\theta)$

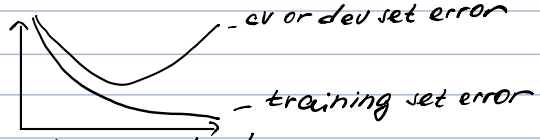


vs





Q: how to select λ? Try a range of λ; use dev set or cross-valid. to find the optimal point

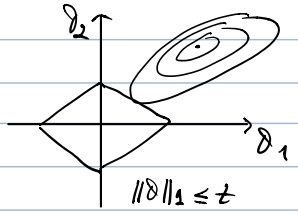


model complexity (# params)

	Low	High
λ	High	Low
Bias	High	Low
Variance	Low	High

$N \uparrow$	$\rightarrow$	fix high variance
$p \downarrow$	$\rightarrow$	— " —
$p \uparrow$	$\rightarrow$	fix high bias
$\lambda \downarrow$	$\rightarrow$	fix high bias
$\lambda \uparrow$	$\rightarrow$	fix high variance

Modifications to penalized least squares:

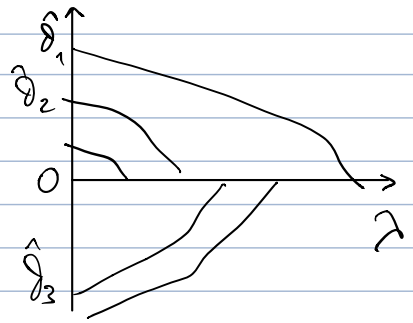


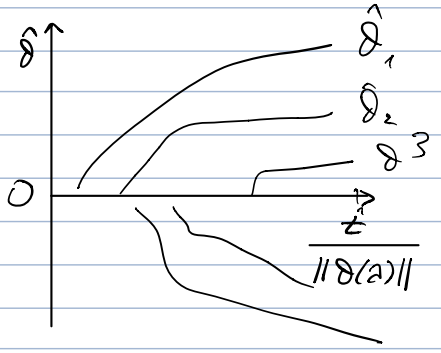
Lasso: $\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\ J(\theta) + \lambda \underbrace{\|\theta\|_1}_{\sum_{j=1}^{p} |\theta_j|}$

Elastic net: $\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\ J(\theta) + \lambda \sum_{j=1}^{p} \left[\alpha \theta_j^2 + (1-\alpha)|\theta_j|\right]$

- Solutions are non-linear in Y, no closed-form expression
- Efficient algos exist
- Penalized optimization, while setting small $\hat{\theta}_j = 0$

Standardize predictors to have mean 0 and sd 1





Option for methods / implementation project:
implement Lasso
(HT Algo 3.2 and 3.2a)