

1. Lasso

Lasso
Ridge

Why We Prefer Sparsity

• 如果维度太高，计算量也变得很高 •
 • 在稀疏性条件下，计算量只依赖非0项的个数
 • 提高可解释性

$P \approx N < D$
 D: # of dimensions
 N: # of samples

1.1 特征选择

Option1: Exhaustive Search : "all subsets"

i.e.	$A = \{f_1, f_2, f_3, f_4\}$		
	$\{f_1\} \rightarrow acc_1$	$\{f_2, f_3\} \rightarrow acc_8$	$\{f_1, f_2, f_3, f_4\} \rightarrow acc_{15}$
	$\{f_2\} \rightarrow acc_2$	$\{f_2, f_4\} \rightarrow acc_9$	$acc_1 = acc_{15}$
	$\{f_3\} \rightarrow acc_3$	$\{f_3, f_4\} \rightarrow acc_{10}$	相比 acc_{15} 是不高
	$\{f_4\} \rightarrow acc_4$	$\{f_1, f_2, f_3\} \rightarrow acc_{11}$	$\{f_3, f_4\}$
	$\{f_1, f_2\} \rightarrow acc_5$	$\{f_1, f_2, f_4\} \rightarrow acc_{12}$	<u>powerset: $2^n - 1$</u>
	$\{f_1, f_3\} \rightarrow acc_6$	$\{f_1, f_3, f_4\} \rightarrow acc_{13}$	$n: \# \text{ of features}$
	$\{f_1, f_4\} \rightarrow acc_7$	$\{f_1, f_2, f_3, f_4\} \rightarrow acc_{14}$	i.e.; $n=4 \Rightarrow 2^4 - 1 = 15$
			如果 $n=100$; $2^{100} - 1 \gg 0$

Option2: Greedy Approaches

Forward Stepwise & Backward Stepwise

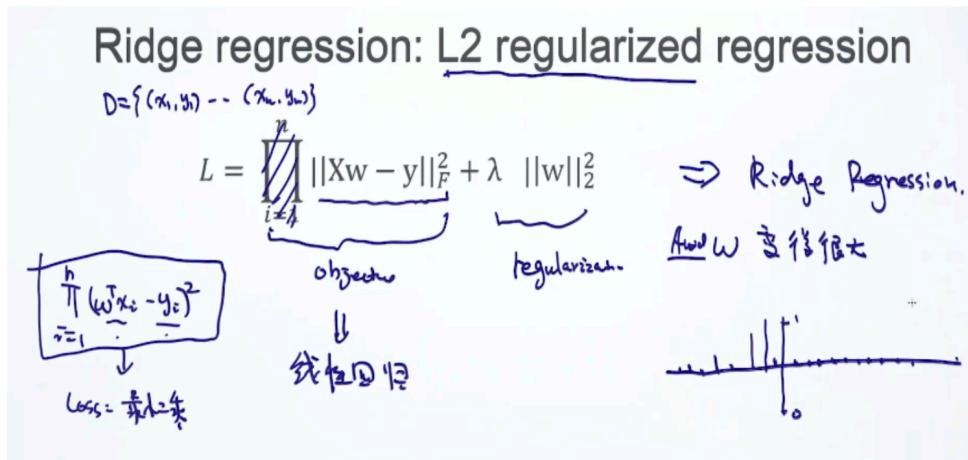
Forward Stepwise (feature Selection)			
$A = \{f_1, f_2, f_3, f_4, f_5\} \quad A =5$			
① best_feature_set = {}	② $A - \text{best_feature_set} = \{f_1, f_2, f_3, f_4\}$	$[f_2, f_3, f_4] \rightarrow acc_3$	
② for f in A :	$\{f_2, f_1\} \rightarrow acc_1$	Terminate.	
$f_1 \rightarrow acc_1$ $\{f_2 \rightarrow acc_2\}$ $f_3 \rightarrow acc_3$ $f_4 \rightarrow acc_4$ $f_5 \rightarrow acc_5$	$\{f_2, f_3\} \rightarrow acc_2$ $\{f_2, f_4\} \rightarrow acc_3$ $\{f_2, f_5\} \rightarrow acc_4$ $\{f_3, f_4\} \rightarrow acc_5$	$\{f_2, f_3\}$	$\{f_2, f_3\} \rightarrow acc_3$
$best_feature_set = \{f_1\}$	$\{f_2, f_3, f_4\} \rightarrow acc_3$	$\{f_2, f_3\}$	$\{f_2, f_3\} \rightarrow acc_3$
\downarrow	$\{f_2, f_3, f_4, f_5\} \rightarrow acc_4$	$\{f_2, f_3, f_4\}$	$\{f_2, f_3, f_4\} \rightarrow acc_4$
a_{83}	$\{f_2, f_3, f_4, f_5\} \rightarrow acc_5$	$\{f_2, f_3, f_4, f_5\}$	$\{f_2, f_3, f_4, f_5\} \rightarrow acc_5$

只能是局部最优解，因为我们只能看到目前最好的情况。

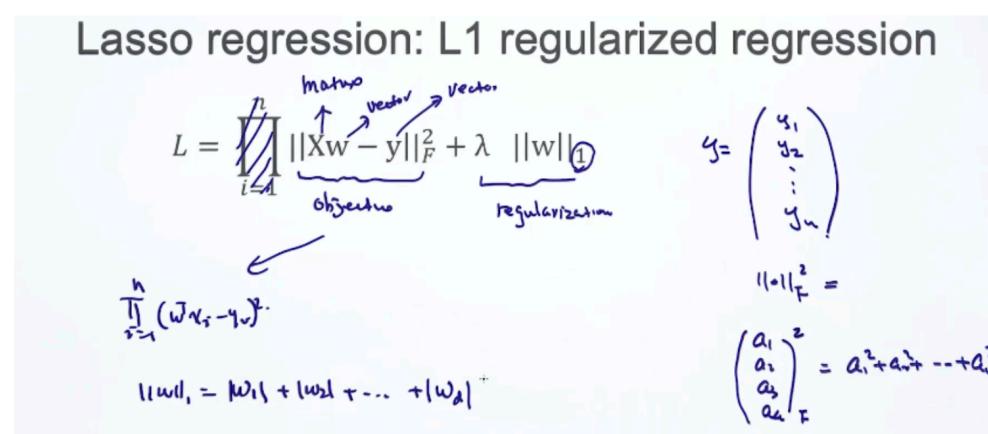
Backward Stepwise			
$A = \{f_1, f_2, f_3, f_4, f_5\}$			
best_feature_set = $\{f_1, f_2, f_3, f_4, f_5\} : 0.83$			
① $\{f_1, f_2, f_3, f_4\} : f_5 : 0.87$	② $\{f_3, f_4, f_5\} : 0.90$	Terminate	
$\{f_1, f_2, f_3, f_5\} : f_4 : 0.86$	$\{f_1, f_4, f_5\} : 0.89$	best_feature_set*	
$\{f_1, f_2, f_4, f_5\} : f_3 : 0.85$	$\{f_2, f_3, f_5\} : 0.90$	= $\{f_1, f_3, f_4\}$	
$\{f_1, f_3, f_4, f_5\} : f_2 : 0.88$	$\{f_1, f_3, f_5\} : 0.91$		
$\{f_2, f_3, f_4, f_5\} : f_1 : 0.87$	best_feature_set = $\{f_1, f_3, f_5\}$		
best_feature_set = $\{f_1, f_3, f_4, f_5\}$	③ $\{f_1, f_3\} : 0.88$		
	$\{f_2, f_3\} : 0.88$		
	$\{f_1, f_3\} : \boxed{0.89} < \boxed{0.91}$		

Option3: via Regularization - A Principled Way (Lasso)

1.2 Lasso regression



可以避免 w 变得很大。但是不会让 w 变零，也就是不会稀疏。
不具备特征选择的功能。



1.3 optimize Lasso objective

$$L = \prod_{l=1}^n ||Xw - y||_F^2 + \lambda \boxed{||w||_1}$$

$\frac{\partial ||w||_1^2}{\partial w} = 2 \cdot w$

$d = 4 \text{ at dimension}$

Q: $|w_j|$ 对 w_j 梯度是多少?

$\frac{\partial ||w||_1}{\partial w_j} = \frac{\partial |w_j|}{\partial w_j}$

2. Coordinate Descent

2.1

函数有很多维度，寻求最优解的时候每次只考虑一个维度。

Goal: minimize some function g

$$\underline{g(w) = g(w_1, w_2, \dots, w_D)}$$

$$g(w_1, \underline{w_2 \dots w_D}) \equiv g(w_1) \rightarrow \underset{w_1}{\text{argmin}}$$

Coordinate descent: $\frac{\partial}{\partial w_1} g(w)$

ascent: $\frac{\partial}{\partial w_1} g(w)$

$\begin{array}{ll} ① t=1 & w_2 \\ \hat{w}_2 = \underset{w_2}{\text{argmin}} g(w_2) & \\ ② t=2 & w_3 \\ \hat{w}_3 = \underset{w_3}{\text{argmin}} g(w_3) & \\ \vdots & \vdots \\ ④ t=3 & w_4 \\ \hat{w}_4 = \underset{w_4}{\text{argmin}} g(w_4) & \\ \vdots & \vdots \end{array}$

$\Rightarrow \hat{w}_1 =$

argmin $g(w_1)$

Random
 w_1
 w_2
 w_3
 w_4
 w_5
 w_6
 \vdots

怎么选择下一个 coordinate?

- ① 依次选择 dimension / coordinates
- ② random selection

不需要设定 step-size!

lasso regression
对于 lasso objective, 它会收敛

2.2 coordinate descent for Lasso

加进 $\log \lambda$ 的 Coordinate Descent for LASSO (1)

$$\tilde{L} = \underbrace{\sum_{i=1}^n \left(\sum_{j=1}^d w_j x_{ij} + b - y_i \right)^2}_{\text{RSS}} + \underbrace{\lambda \sum_{j=1}^d |w_j|}_{L_1 - \text{正则化}}$$

x_{ij} : 第 i 个样本的第 j 个特征

$$\begin{aligned} \frac{\partial \tilde{L}}{\partial w_l} &= 2 \sum_{i=1}^n \left(\sum_{j \neq l}^d w_j x_{ij} + b - y_i \right) \cdot x_{il} + \frac{\partial |w_l|}{\partial w_l} \\ &= 2 \sum_{i=1}^n \left(\sum_{j \neq l}^d w_j x_{ij} + b - y_i + w_l x_{il} \right) \cdot x_{il} + \frac{\partial |w_l|}{\partial w_l} \\ &= 2 \sum_{i=1}^n \left(\sum_{j \neq l}^d w_j x_{ij} + b - y_i \right) x_{il} + 2 \sum_{i=1}^n w_l x_{il}^2 + \frac{\partial |w_l|}{\partial w_l} \\ &= 2 \sum_{i=1}^n \left(\sum_{j \neq l}^d w_j x_{ij} + b - y_i \right) x_{il} + 2w_l \cdot \sum_{i=1}^n x_{il}^2 + \frac{\partial |w_l|}{\partial w_l}. \end{aligned}$$

$C_{ll} + \lambda \sum_{j \neq l}^d w_j x_{il} + \lambda w_l$

$$\begin{aligned} \frac{\partial \tilde{L}}{\partial w_l} &= 2 \cdot \sum_{i=1}^n \left(\sum_{j \neq l}^d w_j x_{ij} + b - y_i \right) x_{il} + 2w_l \sum_{i=1}^n x_{il}^2 + \lambda \frac{\partial |w_l|}{\partial w_l} \quad \alpha_l > 0 \\ &= C_l + w_l \cdot \alpha_l + \frac{\partial |w_l|}{\partial w_l} \end{aligned}$$

$$\frac{\partial \tilde{L}}{\partial w_l} = \begin{cases} C_l + \alpha_l \cdot w_l + \lambda & w_l > 0 \quad ① \\ [C_l - \lambda, C_l + \lambda] & w_l = 0 \quad ② \\ C_l + \alpha_l \cdot w_l - \lambda & w_l < 0 \quad ③ \end{cases}$$

$\begin{array}{l} ① \Rightarrow C_l + \alpha_l w_l + \lambda = 0 \\ w_l = \frac{-C_l - \lambda}{\alpha_l} > 0 \\ -C_l - \lambda > 0 \Rightarrow C_l < -\lambda \end{array}$
 $\begin{array}{l} ② \Rightarrow C_l + \alpha_l w_l - \lambda = 0 \\ w_l = \frac{\lambda - C_l}{\alpha_l} < 0 \\ \lambda - C_l < 0 \Rightarrow C_l > \lambda \end{array}$

对于 w_l 来更新：

$$\hat{w}_l = \begin{cases} \frac{-C_l - \lambda}{\alpha_l}, & \text{if } C_l < -\lambda \\ 0, & \text{if } -\lambda \leq C_l \leq \lambda \\ \frac{\lambda - C_l}{\alpha_l}, & \text{if } C_l > \lambda \end{cases}$$

Sparsity

while not converged
l ← select coordinate



其中在更新的时候，强行将 w 设置为 0，这就是会稀疏的原因。

2.3 other Lasso solvers

Classically: Least angle regression (**LARS**) [Efron et al. '04]

Then: **Coordinate descent** algorithm [Fu '98, Friedman, Hastie, & Tibshirani '08]

Now:

Parallel CD (e.g., Shotgun, [Bradley et al. '11])

Other parallel learning approaches for linear models:

- Parallel stochastic gradient descent (**SGD**) (e.g., Hogwild! [Niu et al. '11])
- Parallel independent solutions then **averaging** [Zhang et al. '12]

Alternating directions method of multipliers (**ADMM**) [Boyd et al. '11]