Variable Selection LASSO: Sparse Regression

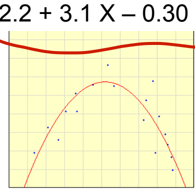
Machine Learning – CSE446
Carlos Guestrin
University of Washington

April 10, 2013

Regularization in Linear Regression

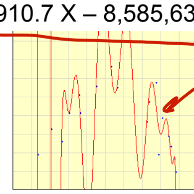
- Overfitting usually leads to very large parameter choices, e.g.:

$-2.2 + 3.1\,X - 0.30\,X^2$



$-1.1 + 4{,}700{,}910.7\,X - 8{,}585{,}638.4\,X^2 + \ldots$

penalty for large weights



over fitting

- ***Regularized*** or ***penalized*** regression aims to impose a "complexity" penalty by penalizing large weights
  - "Shrinkage" method

L_2 regularization → penalizes towards smoother functions

Variable Selection

- Ridge regression: Penalizes large weights
- What if we want to perform "feature selection"?
  - E.g., Which regions of the brain are important for word prediction?
  - Can't simply choose features with largest coefficients in ridge solution
  - Computationally intractable to perform "all subsets" regression

    2^k subsets to explore
- Try new penalty: Penalize non-zero weights
  - Regularization penalty: $\|w\|_1 = \sum_i |w_i|$
  - Leads to sparse solutions $\Rightarrow$ many $w_i = 0$
  - Just like ridge regression, solution is indexed by a continuous param λ
  - This simple approach has changed statistics, machine learning & electrical engineering

LASSO Regression

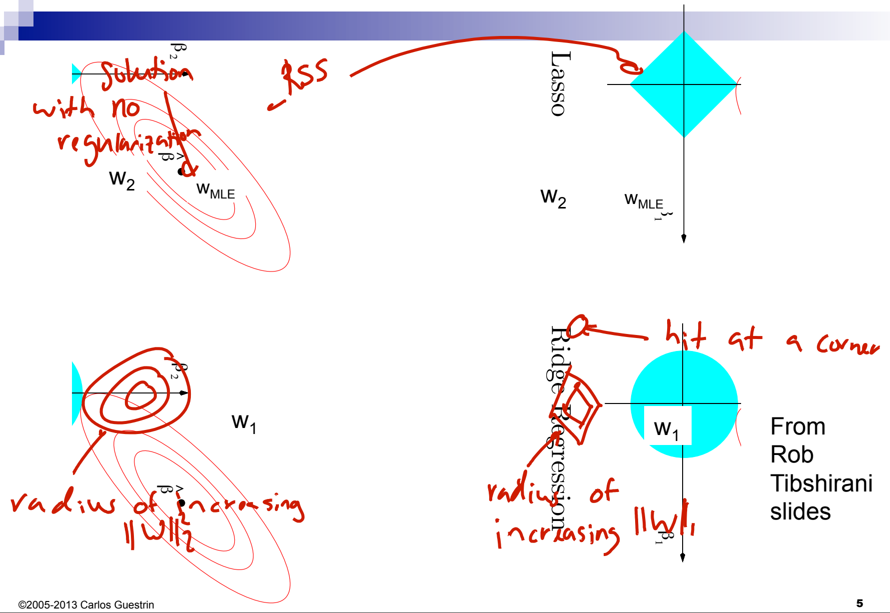
$\lambda > 0$

- **LASSO:** least absolute shrinkage and selection operator
- New objective:

$$\min_w \sum_{j=1}^{N} \left(t(x_j) - \left(w_0 + \sum_i w_i h_i(x_j) \right) \right)^2 + \lambda \sum_{i=1}^{k} |w_i|$$

don't regularize w_0

Geometric Intuition for Sparsity



From Rob Tibshirani slides

Optimizing the LASSO Objective

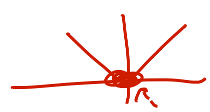
- LASSO solution:

$$\hat{\mathbf{w}}_{LASSO} = \arg\min_{w} \sum_{j=1}^{N} \left(t(x_j) - (w_0 + \sum_{i=1}^{k} w_i h_i(x_j)) \right)^2 + \lambda \sum_{i=1}^{k} |w_i|$$

From quarter thus far: take derivative and set to ∅

But:

1. Derivative of $|w|$? ...



2. Even if you could take derivative, no closed form solution.

Coordinate Descent

$F(w_0, w_1, \ldots, w_k)$

- Given a function $F(w)$
  - Want to find minimum
    $\hat{w} = \arg\min_w F(w)$
- Often, hard to find minimum for all coordinates, but easy for one coordinate
  1-d optimization often much easier
- Coordinate descent: initialize $w = 0$
  while not converged:
    Pick a coordinate ℓ
    $\hat{w}_\ell \leftarrow \arg\min_\omega F(w_0, w_1, \ldots, w_{\ell-1}, \omega, w_{\ell+1}, \ldots w_k)$
- How do we pick next coordinate?
  round robin, randomly, "smartly"
- Super useful approach for *many* problems
  - Converges to optimum in some cases, such as LASSO ✓

Because of:
· convexity &
- separability of non-smooth terms

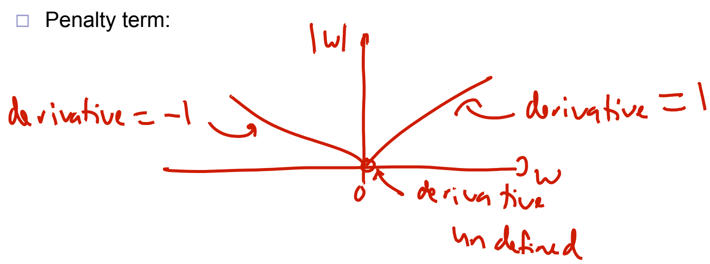
Optimizing LASSO Objective One Coordinate at a Time

$$\overbrace{\sum_{j=1}^{N}\left(t(x_j) - (w_0 + \sum_{i=1}^{k} w_i h_i(x_j))\right)^2}^{RSS(w)} + \overbrace{\lambda \sum_{i=1}^{k} |w_i|}^{Reg}$$

- Taking the derivative: For now, only $\ell \in \{1, \ldots, k\}$, deal with w_0 later
  - Residual sum of squares (RSS):

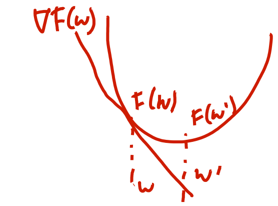
$$\frac{\partial}{\partial w_\ell} RSS(\mathbf{w}) = -2 \sum_{j=1}^{N} h_\ell(x_j) \left(t(x_j) - (w_0 + \sum_{i=1}^{k} w_i h_i(x_j)) \right)$$

  - Penalty term:



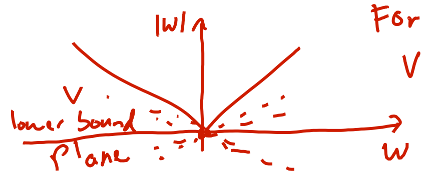
Subgradients of Convex Functions

- Gradients lower bound convex functions:



$F(w') \geq F(w) + \nabla F(w)^T (w'-w)$

- Gradients are unique at **w** iff function differentiable at **w**
- Subgradients: Generalize gradients to non-differentiable points:
  - Any plane that lower bounds function:



For $|w|$ at 0: $v \in [-1,1]$

$v \in \partial_w F(w)$ subgradient iff $F(w') \geq F(w) + v^T(w'-w)$

Taking the Subgradient

$$\sum_{j=1}^{N}\left(t(x_j) - (w_0 + \sum_{i=1}^{k} w_i h_i(x_j))\right)^2 + \lambda \sum_{i=1}^{k} |w_i|$$

- Gradient of RSS term:

$$a_\ell = 2\sum_{j=1}^{N}(h_\ell(\mathbf{x}_j))^2$$

$$\frac{\partial}{\partial w_\ell} RSS(\mathbf{w}) = a_\ell w_\ell - c_\ell$$

$$c_\ell = 2\sum_{j=1}^{N} h_\ell(\mathbf{x}_j)\left(t(\mathbf{x}_j) - (w_0 + \sum_{i \neq \ell} w_i h_i(\mathbf{x}_j))\right)$$

  - If no penalty: $\frac{\partial}{\partial w_\ell} RSS(w) = 0 \Rightarrow w_\ell = c_\ell / a_\ell$
- Subgradient of full objective:

$$\partial_{w_\ell} F(w) = a_\ell w_\ell - c_\ell + \lambda \underbrace{\partial_{w_\ell} |w_\ell|}$$

$$\partial_{w_\ell}|w_\ell| = \begin{cases} -1 & \text{if } w_\ell < 0 \\ [-1,1] & \text{if } w_\ell = 0 \\ 1 & \text{if } w_\ell > 0 \end{cases}$$

$$= \begin{cases} a_\ell w_\ell - c_\ell - \lambda & \text{if } w_\ell < 0 \\ [-c_\ell - \lambda, -c_\ell + \lambda] & \text{if } w_\ell = 0 \\ a_\ell w_\ell - c_\ell + \lambda & \text{if } w_\ell > 0 \end{cases}$$

Setting Subgradient to 0

$a_\ell > 0$

$$0 = \partial_{w_\ell} F(\mathbf{w}) = \begin{cases} a_\ell w_\ell - c_\ell - \lambda & w_\ell < 0 \\ [-c_\ell - \lambda, -c_\ell + \lambda] & w_\ell = 0 \\ a_\ell w_\ell - c_\ell + \lambda & w_\ell > 0 \end{cases}$$

when $w_\ell < 0$? $\Rightarrow$ $a_\ell w_\ell - c_\ell - \lambda = 0$

$\Rightarrow$ $w_\ell = \frac{c_\ell + \lambda}{a_\ell} < 0$ $\Rightarrow$ $c_\ell < -\lambda$

when $w_\ell > 0$? $\Rightarrow$ $a_\ell w_\ell - c_\ell + \lambda = 0$

$\Rightarrow$ $w_\ell = \frac{c_\ell - \lambda}{a_\ell} > 0$ $\Rightarrow$ $c_\ell > \lambda$

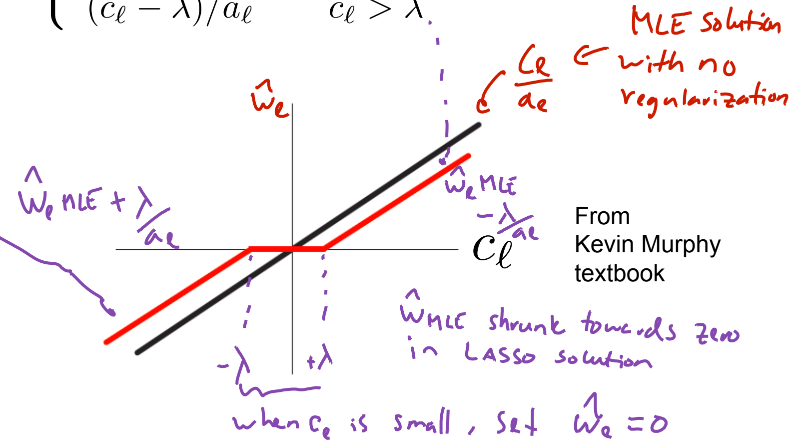
when $w_\ell = 0$? $\Rightarrow$ $0 \in [-c_\ell - \lambda, -c_\ell + \lambda]$

$\Rightarrow$ $-\lambda < c_\ell < \lambda$

Soft Thresholding

$a_\ell > 0$

$$\hat{w}_\ell = \begin{cases} (c_\ell + \lambda)/a_\ell & c_\ell < -\lambda \\ 0 & c_\ell \in [-\lambda, \lambda] \\ (c_\ell - \lambda)/a_\ell & c_\ell > \lambda \end{cases}$$



From Kevin Murphy textbook

Coordinate Descent for LASSO (aka Shooting Algorithm)

- Repeat until convergence
  - Pick a coordinate l at (random or sequentially)

    round robin 2
    - Set:

$$\hat{w}_\ell = \begin{cases} (c_\ell + \lambda)/a_\ell & c_\ell < -\lambda \\ 0 & c_\ell \in [-\lambda, \lambda] \\ (c_\ell - \lambda)/a_\ell & c_\ell > \lambda \end{cases}$$

    - Where:

$$a_\ell = 2\sum_{j=1}^{N}(h_\ell(\mathbf{x}_j))^2$$

$$c_\ell = 2\sum_{j=1}^{N} h_\ell(\mathbf{x}_j)\left(t(\mathbf{x}_j) - (w_0 + \sum_{i \neq \ell} w_i h_i(\mathbf{x}_j))\right)$$

w_0 ??

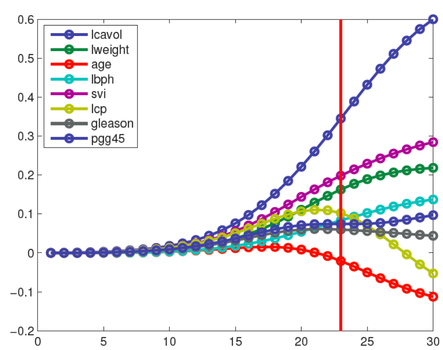
no regularization.

$w_0 = c_0 / a_0$

$\Rightarrow w_0 = \frac{1}{N} \sum_{j=1}^{N} \left(t(x_j) - \sum_{i=1}^{k} w_i h_i(x_j)\right)$

  - For convergence rates, see Shalev-Shwartz and Tewari 2009
- Other common technique = LARS
  - Least angle regression and shrinkage, Efron et al. 2004

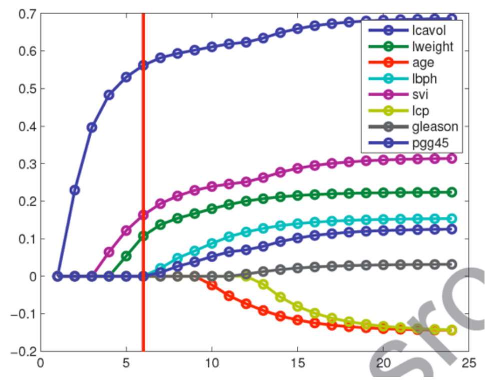
Recall: *Ridge Coefficient Path*



From
Kevin Murphy
textbook

- Typical approach: select λ using cross validation

Now: *LASSO Coefficient Path*



From
Kevin Murphy
textbook

LASSO Example

Term	Least Squares	Ridge	Lasso
Intercept	2.465	2.452	2.468
lcavol	0.680	0.420	0.533
lweight	0.263	0.238	0.169
age	−0.141	−0.046	
lbph	0.210	0.162	0.002
svi	0.305	0.227	0.094
lcp	−0.288	0.000	
gleason	−0.021	0.040	
pgg45	0.267	0.133	

From
Rob
Tibshirani
slides

What you need to know

- Variable Selection: find a sparse solution to learning problem
- L_1 regularization is one way to do variable selection
  - Applies beyond regressions
  - Hundreds of other approaches out there
- LASSO objective non-differentiable, but convex ➔ Use subgradient
- No closed-form solution for minimization ➔ Use coordinate descent
- Shooting algorithm is very simple approach for solving LASSO