ECE595 / STAT598: Machine Learning I
Lecture 2.2: Regularization - LASSO Regression

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

Outline

Mathematical Background

- Lecture 1: Linear regression: A basic data analytic tool
- Lecture 2: Regularization: Constraining the solution
- Lecture 3: Kernel Method: Enabling nonlinearity

Lecture 2: Regularization

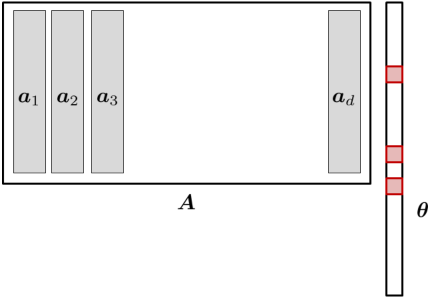
- Ridge Regression
  - Regularization
  - Parameter
- LASSO Regression
  - Sparsity
  - Algorithm
  - Application

LASSO Regression

- An alternative to the Ridge Regression is **L**east **A**bsolute **S**hrinkage and **S**election **O**perator (**LASSO**)
- The loss function is

$$J(\boldsymbol{\theta}) = \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|_1$$

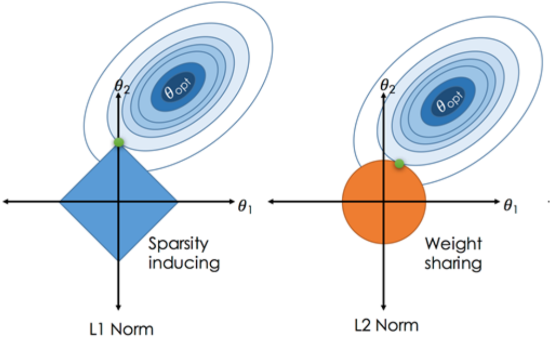
- Intuition behind LASSO: Many features are not active.



Interpreting the LASSO Solution

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \quad \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|_1$$

- $\|\boldsymbol{\theta}\|_1$ promotes sparsity of $\boldsymbol{\theta}$. It is the nearest convex approximation to $\|\boldsymbol{\theta}\|_0$, which is the number of non-zeros.
- The difference between ℓ_2 and ℓ_1 [1]:



[1] Figure source: `http://www.ds100.org/`

Why are Sparse Models Useful?

non-zeros = 33.51%

13.58%

1.21%

- Images are sparse in transform domains, e.g., Fourier and wavelet.
- Intuition: There are more low frequency components and less high frequency components.
- Examples above: $\boldsymbol{A}$ is the wavelet basis matrix. $\boldsymbol{\theta}$ are the wavelet coefficients.
- We can truncate the wavelet coefficients and retain a good image.
- Many image compression schemes are based on this, e.g., JPEG, JPEG2000.

LASSO for Image Reconstruction

Image inpainting via KSVD **dictionary-learning** [2]

- $\boldsymbol{y}$ = image with missing pixels. $\boldsymbol{A}$ = a matrix storing a set of trained feature vectors (called dictionary atoms). $\boldsymbol{\theta}$ = coefficients.
- minimize $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_1$.
- KSVD = k-means + Singular Value Decomposition (SVD): A method to train the feature vectors that demonstrate sparse representations.

[2] Figure is taken from Mairal, Elad, Sapiro, IEEE T-IP 2008
`https://ieeexplore.ieee.org/document/4392496`

Shrinkage Operator

The LASSO problem can be solved using a shrinkage operator. Consider a simplified problem (with $\boldsymbol{A} = \boldsymbol{I}$)

$$
\begin{aligned}
J(\boldsymbol{\theta}) &= \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_1 \\
&= \sum_{j=1}^{d} \left\{ \frac{1}{2}(y_j - \theta_j)^2 + \lambda|\theta_j|_1 \right\}
\end{aligned}
$$

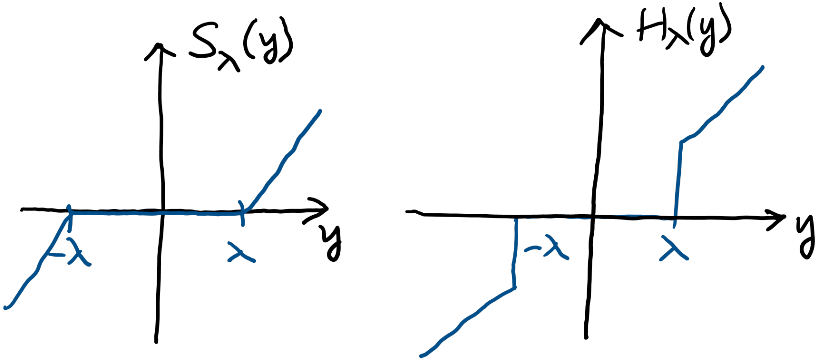
Since the loss is **separable**, the ,optimization is solved when each individual term is minimized. The individual problem

$$
\begin{aligned}
\widehat{\theta} &= \underset{\theta}{\text{argmin}} \left\{ \frac{1}{2}(y - \theta)^2 + \lambda|\theta| \right\} \\
&= \max(|y| - \lambda, 0)\text{sign}(y) \\
&\overset{\text{def}}{=} \mathcal{S}_\lambda(y).
\end{aligned}
$$

Proof: See Appendix.

Shrinkage VS Hard Threshold

- The shrinkage operator looks as follows.
- Any number between $[-\lambda, \lambda]$ is "shrink" to zero.
- Try compare with the hard threshold operator $\mathcal{H}_\lambda(y) = y \cdot \mathbf{1}\{|y| \geq \lambda\}$



Algorithms to Solve LASSO Regression

In general, the LASSO problem requires iterative algorithms:

- ISTA Algorithm (Daubechies et al. 2004)
  - For $k = 1, 2, \ldots$
  - $\boldsymbol{v}^k = \boldsymbol{\theta}^k - 2\gamma \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{\theta}^k - \boldsymbol{y})$.
  - $\boldsymbol{\theta}^{k+1} = \max(|\boldsymbol{v}^k| - \lambda, 0)\text{sign}(\boldsymbol{v}^k)$.
- FISTA Algorithm (Beck-Teboulle 2008)
  - For $k = 1, 2, \ldots$
  - $\boldsymbol{v}^k = \boldsymbol{\theta}^k - 2\gamma \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{\theta}^k - \boldsymbol{y})$.
  - $\boldsymbol{z}^k = \max(|\boldsymbol{v}^k| - \lambda, 0)\text{sign}(\boldsymbol{v}^k)$.
  - $\boldsymbol{\theta}^{k+1} = \alpha_k \boldsymbol{\theta}^k + (1 - \alpha_k)\boldsymbol{z}^k$.
- ADMM Algorithm (Eckstein-Bertsekas 1992, Boyd et al. 2011)
  - For $k = 1, 2, \ldots$
  - $\boldsymbol{\theta}^{k+1} = (\boldsymbol{A}^T\boldsymbol{A} + \rho\boldsymbol{I})^{-1}(\boldsymbol{A}^T\boldsymbol{y} + \rho\boldsymbol{z}^k - \boldsymbol{u}^k)$
  - $\boldsymbol{z}^{k+1} = \max(|\boldsymbol{\theta}^{k+1} + \boldsymbol{u}^k/\rho| - \lambda/\rho, 0)\text{sign}(\boldsymbol{\theta}^{k+1} + \boldsymbol{u}^k/\rho)$
  - $\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \rho(\boldsymbol{\theta}^{k+1} - \boldsymbol{z}^{k+1})$
- And many others.

Example: Crime Rate Data

city	funding	hs	not-hs	college	college4	crime rate
1	40	74	11	31	20	478
2	32	72	11	43	18	494
3	57	70	18	16	16	643
4	31	71	11	25	19	341
5	67	72	9	29	24	773
⋮	⋮	⋮	⋮	⋮		
50	66	67	26	18	16	940

https://web.stanford.edu/~hastie/StatLearnSparsity/data.html

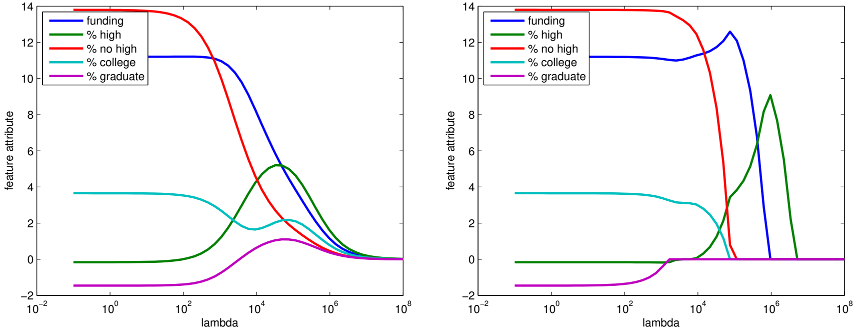
Consider the following two optimizations

$$\widehat{\boldsymbol{\theta}}_1(\lambda) = \underset{\boldsymbol{\theta}}{\text{argmin}} \quad J_1(\boldsymbol{\theta}) \overset{\text{def}}{=} \|\boldsymbol{A\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|_1,$$

$$\widehat{\boldsymbol{\theta}}_2(\lambda) = \underset{\boldsymbol{\theta}}{\text{argmin}} \quad J_2(\boldsymbol{\theta}) \overset{\text{def}}{=} \|\boldsymbol{A\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|^2.$$

Comparison between ℓ-1 and ℓ-2 norm

- Plot $\widehat{\boldsymbol{\theta}}_1(\lambda)$ and $\widehat{\boldsymbol{\theta}}_2(\lambda)$ vs. λ.
- LASSO tells us which factor appears first.
- If we are allowed to use only one feature, then % high is the one.
- Two features, then % high + funding.



Ridge LASSO

Pros and Cons

Ridge Regression

- (+) Analytic solution, because the loss function is differentiable.
- (+) As such, a lot of well-established theoretical guarantees.
- (+) Algorithm is simple, just one equation.
- (-) Limited interpretability, since the solution is usually a dense vector.
- (-) Does not reflect the nature of certain problems, e.g., sparsity.

LASSO

- (+) Proven applications in many domains, e.g., images and speeches.
- (+) Echoes particularly well in modern deep learning where parameter space is huge.
- (+) Increasing number of theoretical guarantees for special matrices.
- (+) Algorithms are available.
- (-) No closed-form solution. Algorithms are iterative.

Reading List

Ridge Regression

- Stanford CS 229 Note on Linear Algebra
  http://cs229.stanford.edu/section/cs229-linalg.pdf
- Lecture Note on Ridge Regression
  https://arxiv.org/pdf/1509.09169.pdf
- Theobald, C. M. (1974). Generalizations of mean square error applied to ridge regression. Journal of the Royal Statistical Society. Series B (Methodological), 36(1), 103-106.

LASSO Regression

- ECE/STAT 695 (Lecture 1)
  https://engineering.purdue.edu/ChanGroup/ECE695.html
- Statistical Learning with Sparsity (Chapter 2)
  https://web.stanford.edu/~hastie/StatLearnSparsity/
- Elements of Statistical Learning (Chapter 3.4)
  https://web.stanford.edu/~hastie/ElemStatLearn/