COMS 4721: Machine Learning for Data Science
Lecture 6, 2/2/2017

Prof. John Paisley

Department of Electrical Engineering
& Data Science Institute

Columbia University

UNDERDETERMINED LINEAR EQUATIONS

We now consider the regression problem $y = Xw$ where $X \in \mathbb{R}^{n\times d}$ is “fat” (i.e., $d \gg n$). This is called an “underdetermined” problem.

- There are more dimensions than observations.
- w now has an infinite number of solutions satisfying $y = Xw$.

$$\begin{bmatrix} y \end{bmatrix} = \begin{bmatrix} \qquad\qquad X \qquad\qquad \end{bmatrix} \begin{bmatrix} \\ w \\ \\ \\ \end{bmatrix}$$

These sorts of high-dimensional problems often come up:

- In gene analysis there are 1000’s of genes but only 100’s of subjects.
- Images can have millions of pixels.
- Even polynomial regression can quickly lead to this scenario.

Minimum ℓ_2 regression

One solution (least norm)

One possible solution to the underdetermined problem is

$$w_{\mathrm{ln}} = X^T(XX^T)^{-1}y \quad \Rightarrow \quad Xw_{\mathrm{ln}} = XX^T(XX^T)^{-1}y = y.$$

We can construct another solution by adding to w_{ln} a vector $\delta \in \mathbb{R}^d$ that is in the *null space* $\mathcal{N}$ of X:

$$\delta \in \mathcal{N}(X) \quad \Rightarrow \quad X\delta = 0 \text{ and } \delta \neq 0$$

and so $X(w_{\mathrm{ln}} + \delta) = Xw_{\mathrm{ln}} + X\delta = y + 0$.

In fact, there are an infinite number of possible δ, because $d > n$.

We can show that w_{ln} is the solution with smallest ℓ_2 norm. We will use the proof of this fact as an excuse to introduce two general concepts.

TOOLS: ANALYSIS

We can use *analysis* to prove that w_{ln} satisfies the optimization problem

$$w_{\text{ln}} = \arg\min_w \ \|w\|^2 \quad \text{subject to} \quad Xw = y.$$

(Think of mathematical analysis as the use of inequalities to prove things.)

Proof: Let w be another solution to $Xw = y$, and so $X(w - w_{\text{ln}}) = 0$. Also,

$$\begin{aligned}(w - w_{\text{ln}})^T w_{\text{ln}} &= (w - w_{\text{ln}})^T X^T (XX^T)^{-1} y \\ &= \underbrace{(X(w - w_{\text{ln}}))}_{=\,0}{}^T (XX^T)^{-1} y \ = \ 0\end{aligned}$$

As a result, the vector $w - w_{\text{ln}}$ is *orthogonal* to w_{ln}. It follows that

$$\|w\|^2 = \|w - w_{\text{ln}} + w_{\text{ln}}\|^2 = \|w - w_{\text{ln}}\|^2 + \|w_{\text{ln}}\|^2 + 2\underbrace{(w - w_{\text{ln}})^T w_{\text{ln}}}_{=\,0} > \|w_{\text{ln}}\|^2$$

TOOLS: LAGRANGE MULTIPLIERS

Instead of starting from the solution, start from the problem,

$$w_{\text{ln}} = \arg\min_{w} \; w^T w \quad \text{subject to} \quad Xw = y.$$

- Introduce Lagrange multipliers: $\mathcal{L}(w, \eta) = w^T w + \eta^T (Xw - y)$.
- Maximize over η, minimize over w. If $Xw \neq y$, we can get $\mathcal{L} = +\infty$, so η effectively *forces* this equality.
- The optimal conditions are

$$\nabla_w \mathcal{L} = 2w + X^T \eta = 0, \qquad \nabla_\eta \mathcal{L} = Xw - y = 0.$$

We have everything necessary to find the solution:

1. From first condition we know: $w = -X^T \eta / 2$
2. Plug #1 into second condition to find: $\eta = -2(XX^T)^{-1} y$
3. Plug #2 back into #1 to find the solution: $w_{\text{ln}} = X^T (XX^T)^{-1} y$

SPARSE ℓ_1 REGRESSION

LS AND RR IN HIGH DIMENSIONS

LS and RR not suited for high-dimensional data

- Modern problems: Many dimensions/features/predictors
- Only a few of these may be important or relevant for predicting y
- Therefore, we need some form of “feature selection”

- Some drawbacks of LS and RR in high dimensions are:
  - They weight all dimensions without favoring subsets of dimensions
  - The unknown “important” dimensions are averaged with irrelevant ones
  - They generalize poorly to new data, weights may not be interpretable

Regression with Penalties

Penalty terms

Recall: General ℓ_2-penalized regression is of the form

$$\mathcal{L} = \sum_{i=1}^{n} (y_i - f(x_i; w))^2 + \lambda \|w\|^2$$

We've referred to the term $\|w\|^2$ as a *penalty term* and used $f(x_i; w) = x_i^T w$.

Penalized fitting

The general structure of the optimization problem is

$$\text{total cost} \;=\; \text{goodness-of-fit term} \;+\; \text{penalty term}$$

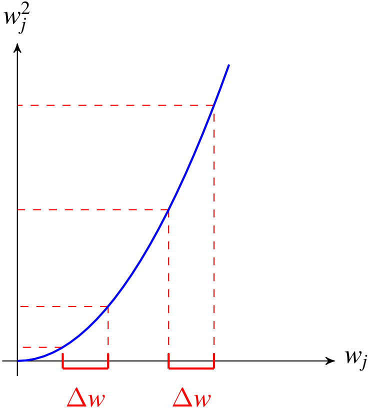
- Goodness-of-fit measures how well our model f approximates the data.
- Penalty term makes the solutions we don't want more "expensive".

What kind of solutions does the choice $\|w\|^2$ favor or discourage?

Quadratic Penalties

Intuitions

- Quadratic penalty: Reduction in cost depends on $|w_j|$.
- Suppose we reduce w_j by Δw. The effect on $\mathcal{L}$ depends on the starting point of w_j.
- We penalize larger values much more than smaller ones.
- Consequence: We will favor vectors w whose entries are of similar size, preferably small.



SPARSITY

Our setting

- Regression problem with n data points $x \in \mathbb{R}^d$, $d \gg n$.
- Model data using a linear function, $y \approx f(x, w) = x^T w$.
- Goal: Select a small subset of the d dimensions and switch off the rest. This is sometimes referred to as "feature selection".

What does it mean to "switch off" a dimension?

- Each entry of w corresponds to a dimension of the data x.
- If $w_k = 0$, the prediction does not depend on the kth dimension,

$$f(x, w) = x^T w = w_1 x_1 + \cdots + 0 \cdot x_k + \cdots + w_d x_d ,$$

- Feature selection: Find a vector w that (1) predicts well, and (2) has only a small number of non-zero entries.
- A w for which most dimensions $= 0$ is called a *sparse* solution.

Sparsity and Penalties

Penalty goal

Find a penalty term which encourages sparse solutions.

Quadratic penalty vs sparsity

- Suppose using least squares w_k is large, all other w_j are very small
- Sparsity: The penalty should keep w_k large, and push other w_j to zero
- Quadratic penalty: Will favor entries w_j which all have similar size, and so it will push w_k towards a smaller value.

A solution

Sparsity can be achieved using a *linear* penalty term.

LASSO

Sparse regression

One penalty that encourages a sparse solution is known as the "LASSO."

LASSO: Least Absolute Shrinkage and Selection Operator

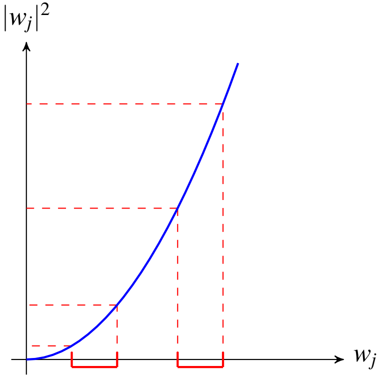
With the LASSO, we replace the ℓ_2 penalty with an ℓ_1 penalty:

$$w_{\text{lasso}} = \arg\min_{w} \|y - Xw\|_2^2 + \lambda\|w\|_1$$

where $\|w\|_1 = \sum_{j=1}^{d} |w_j|$. This is also called ℓ_1-regularized regression.

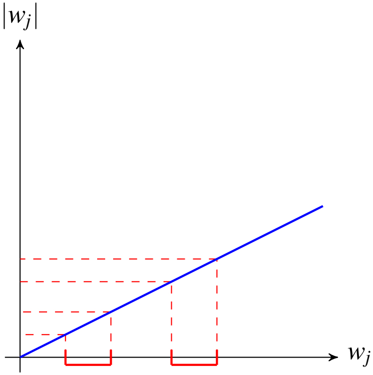
Quadratic Penalties

Quadratic penalty



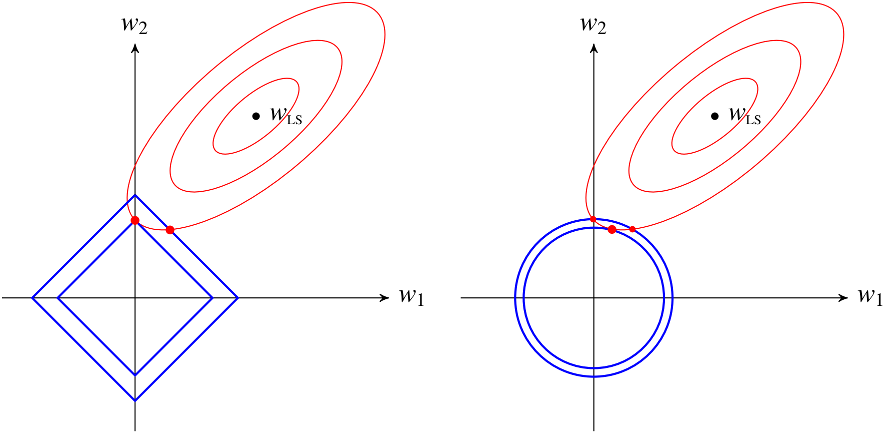
Reducing a large value w_j achieves a larger cost reduction.

Linear penalty



Cost reduction does not depend on the magnitude of w_j.

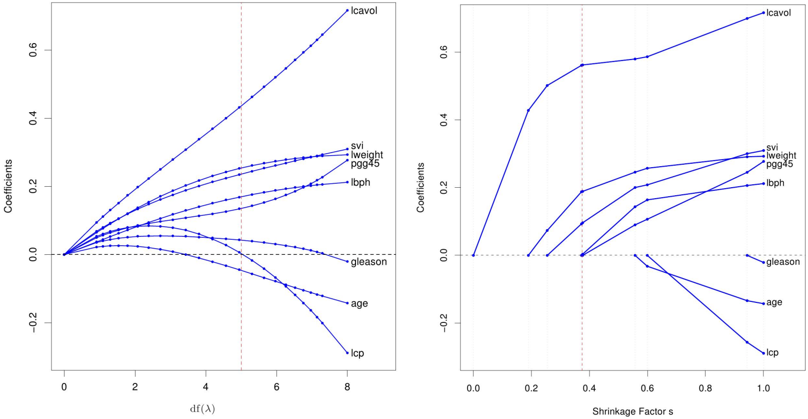
Ridge Regression vs LASSO



This figure applies to $d < n$, but gives intuition for $d \gg n$.

- Red: Contours of $(w - w_{\text{LS}})^T(X^TX)(w - w_{\text{LS}})$ (see Lecture 3)
- Blue: (left) Contours of $\|w\|_1$, and (right) contours of $\|w\|_2^2$

COEFFICIENT PROFILES: RR VS LASSO



(a) $||w||_2$ penalty

(b) $||w||_1$ penalty

ℓ_p Regression

ℓ_p-norms

These norm-penalties can be extended to all norms:

$$\|w\|_p = \left(\sum_{j=1}^{d} |w_j|^p \right)^{\frac{1}{p}} \qquad \text{for } 0 < p \leq \infty$$

ℓ_p-regression

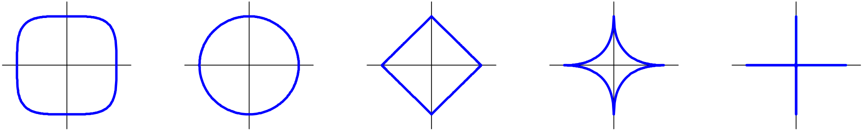
The ℓ_p-regularized linear regression problem is

$$w_{\ell_p} := \arg\min_{w} \|y - Xw\|_2^2 + \lambda \|w\|_p^p$$

We have seen:

- ℓ_1-regression = LASSO
- ℓ_2-regression = ridge regression

ℓ_p Penalization Terms



$p = 4$ $p = 2$ $p = 1$ $p = 0.5$ $p = 0.1$

p	Behavior of $\| \, . \, \|_p$
$p = \infty$	Norm measures largest absolute entry, $\|w\|_\infty = \max_j \|w_j\|$
$p > 2$	Norm focuses on large entries
$p = 2$	Large entries are expensive; encourages similar-size entries
$p = 1$	Encourages sparsity
$p < 1$	Encourages sparsity as for $p = 1$, but contour set is not convex (i.e., no "line of sight" between every two points inside the shape)
$p \to 0$	Simply records whether an entry is non-zero, i.e. $\|w\|_0 = \sum_j \mathbb{I}\{w_j \neq 0\}$

Computing the solution for ℓ_p

Solution of ℓ_p problem

$\boldsymbol{\ell_2}$ aka ridge regression. Has a closed form solution

$\boldsymbol{\ell_p}$ $(\boldsymbol{p \geq 1, p \neq 2})$ — By “convex optimization”. We won’t discuss convex analysis in detail in this class, but two facts are important

- There are no “local optimal solutions” (i.e., local minimum of $\mathcal{L}$)
- The true solution can be found *exactly* using iterative algorithms

$(\boldsymbol{p < 1})$ — We can only find an approximate solution (i.e., the best in its “neighborhood”) using iterative algorithms.

Three techniques formulated as optimization problems

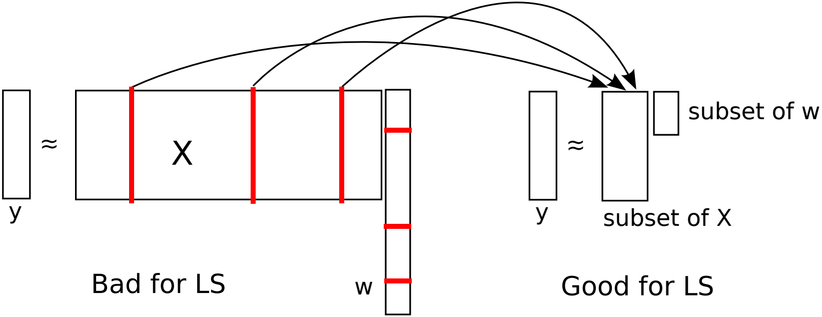
Method	Good-o-fit	penalty	Solution method
Least squares	$\|\|y - Xw\|\|_2^2$	none	Analytic solution exists if $X^T X$ invertible
Ridge regression	$\|\|y - Xw\|\|_2^2$	$\|\|w\|\|_2^2$	Analytic solution exists always
LASSO	$\|\|y - Xw\|\|_2^2$	$\|\|w\|\|_1$	Numerical optimization to find solution

GREEDY SPARSE REGRESSION

Greedy subset selection

There are very many algorithms for finding sparse solutions. Some of them use greedy methods. Many approaches build on least squares. For example,

- Imagine we knew a good subset of $k < n \ll d$ columns of $X \in \mathbb{R}^{n \times d}$ for which the corresponding dimensions of w are $\neq 0$.
- We could pick out that subset of w and learn it using least squares.
- The question is how to pick that subset.



ORTHOGONAL MATCHING PURSUITS

OMP (also called *forward stepwise regression*) sequentially picks columns of X and allows the corresponding dimensions in w to $\neq 0$.

It has two-steps. Given the indexes of k columns selected from X (call it $\mathcal{I}_k$):

1. Find the least squares solution and the approximation error (residual),

$$w_{\text{LS}}^{(k)} = (X_{\mathcal{I}_k}^T X_{\mathcal{I}_k})^{-1} X_{\mathcal{I}_k}^T y, \qquad r^{(k)} = y - X_{\mathcal{I}_k} w_{\text{LS}}^{(k)}.$$

2. "Activate" the column of X that correlates the most with the error,

$$\text{Pick } j\text{th column of } X \text{ (call it } X_j\text{)}, \;\; \text{where } \; j = \arg\max_{j'} \frac{|X_{j'}^T r^{(k)}|}{\|X_{j'}\|_2 \|r^{(k)}\|_2}.$$

Comments:

#1. $X_{\mathcal{I}_k} w_{\text{LS}}^{(k)}$ gets us as close to y as possible using only the active columns.

#2. The angle θ between vectors a and b is found by $a^T b = \|a\|_2 \|b\|_2 \cos\theta$. Thus adding X_j gets us closest to y if it's the last column we can pick.