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

Prof. John Paisley

Department of Electrical Engineering & Data Science Institute
Columbia University
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 $y = Xw$.

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 $\ell_2$ regression
One possible solution to the underdetermined problem is

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

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 \( X \):

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

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

In fact, there are an infinite number of possible \( \delta \), because \( d > n \).

We can show that \( w_{ln} \) is the solution with smallest \( \ell_2 \) norm. We will use the proof of this fact as an excuse to introduce two general concepts.
We can use *analysis* to prove that \( w_{ln} \) satisfies the optimization problem

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

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

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

\[
(w - w_{ln})^T w_{ln} = (w - w_{ln})^T X^T (XX^T)^{-1} y
\]

\[
= (X(w - w_{ln}))^T (XX^T)^{-1} y = 0
\]

As a result, \( w - w_{ln} \) is *orthogonal* to \( w_{ln} \). It follows that

\[
\|w\|^2 = \|w - w_{ln} + w_{ln}\|^2 = \|w - w_{ln}\|^2 + \|w_{ln}\|^2 + 2 (w - w_{ln})^T w_{ln} > \|w_{ln}\|^2
\]

\[
= 0
\]
Instead of starting from the solution, start from the problem,

\[ w_{ln} = \arg \min_w w^T w \quad \text{subject to} \quad Xw = y. \]

- Introduce Lagrange multipliers: \( \mathcal{L}(w, \eta) = w^T w + \eta^T (Xw - y). \)
- Minimize \( \mathcal{L} \) over \( w \) maximize over \( \eta \). If \( Xw \neq y \), we can get \( \mathcal{L} = +\infty \).
- The optimal conditions are

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

We have everything necessary to find the solution:

1. From first condition: \( w = -X^T \eta / 2 \)
2. Plug into second condition: \( \eta = -2(XX^T)^{-1}y \)
3. Plug this back into \#1: \( w_{ln} = X^T (XX^T)^{-1}y \)
Sparse $\ell_1$ regression
**LS and RR in high dimensions**

Usually 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”

- Least squares and ridge regression:
  - Treat all dimensions equally without favoring subsets of dimensions
  - The relevant dimensions are averaged with irrelevant ones
  - Problems: Poor generalization to new data, interpretability of results
Penalty terms
Recall: General ridge 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?
**Intuitions**

- Quadratic penalty: Reduction in cost depends on $|w_j|$.

- Suppose we reduce $w_j$ by $\Delta w$. The effect on $\mathcal{L}$ depends on the starting point of $w_j$.

- Consequence: We should favor vectors $w$ whose entries are of similar size, preferably small.
Sparsity

Setting

- Regression problem with \( n \) data points \( x \in \mathbb{R}^d, d \gg n \).
- 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 is

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

so the prediction does not depend on the \( k \)th dimension.
- Feature selection: Find a \( w \) 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 \textit{sparse} solution.
Penalty goal
Find a penalty term which encourages sparse solutions.

Quadratic penalty vs sparsity

- Suppose $w_k$ is large, all other $w_j$ are very small but non-zero
- Sparsity: Penalty should keep $w_k$, 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 small value.

Overall, a quadratic penalty favors many small, but non-zero values.

Solution
Sparsity can be achieved using linear penalty terms.
Sparse regression

**LASSO**: Least Absolute Shrinkage and Selection Operator

With the LASSO, we replace the $\ell_2$ penalty with an $\ell_1$ 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 $\ell_1$-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 $w_j$. 
This figure applies to $d < n$, but gives intuition for $d \gg n$.

- Red: Contours of $(w - w_{LS})^T (X^T X) (w - w_{LS})$ (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
\( \ell_p \) Regression

\( \ell_p \)-norms

These norm-penalties can be extended to all norms:

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

\( \ell_p \)-regression

The \( \ell_p \)-regularized linear regression problem is

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

We have seen:

- \( \ell_1 \)-regression = LASSO
- \( \ell_2 \)-regression = ridge regression
**${\ell}_p$ Penalization Terms**

<table>
<thead>
<tr>
<th>$p$</th>
<th>Behavior of $| . |_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = \infty$</td>
<td>Norm measures largest absolute entry, $|w|_\infty = \max_j</td>
</tr>
<tr>
<td>$p &gt; 2$</td>
<td>Norm focuses on large entries</td>
</tr>
<tr>
<td>$p = 2$</td>
<td>Large entries are expensive; encourages similar-size entries</td>
</tr>
<tr>
<td>$p = 1$</td>
<td>Encourages sparsity</td>
</tr>
<tr>
<td>$p &lt; 1$</td>
<td>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)</td>
</tr>
<tr>
<td>$p \to 0$</td>
<td>Simply records whether an entry is non-zero, i.e. $|w|_0 = \sum_j \mathbb{I}{w_j \neq 0}$</td>
</tr>
</tbody>
</table>
Solution of $\ell_p$ problem

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

$\ell_p$ ($p \geq 1, p \neq 2$) — 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 $L$)
▶ The true solution can be found exactly using iterative algorithms

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

Three techniques formulated as optimization problems

<table>
<thead>
<tr>
<th>Method</th>
<th>Good-o-fit</th>
<th>penalty</th>
<th>Solution method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least squares</td>
<td>$|y - Xw|^2_2$</td>
<td>none</td>
<td>Analytic solution exists if $X^TX$ invertible</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>$|y - Xw|^2_2$</td>
<td>$|w|^2_2$</td>
<td>Analytic solution exists always</td>
</tr>
<tr>
<td>LASSO</td>
<td>$|y - Xw|^2_2$</td>
<td>$|w|^1_1$</td>
<td>Numerical optimization to find solution</td>
</tr>
</tbody>
</table>