15.095: Machine Learning under a Modern Optimization Lens

Lecture 3: Robust Linear Regression

Outline

Robustness View of Regression

- In reality, data is uncertain — $\mathbf{X}$ and $\mathbf{y}$ are not known exactly.

- We will focus on $\mathbf{X}$.

$$\text{e.g.} \qquad \mathbf{X} = \begin{pmatrix} 1.33 & -83.5 \\ -10.1 & 0.7 \\ 2.2 & 12.4 \end{pmatrix} \qquad \leftrightsquigarrow \qquad \widetilde{\mathbf{X}} = \begin{pmatrix} 1.2 & -83.5 \\ 10.1 & 1.7 \\ 2.0 & 12.3 \end{pmatrix}$$

Why Does Robustness Matter?

Let's go back to the diabetes example from previous lecture:

$n = 350$ patients and $p = 55$:

- 10 baseline variables x_i (age, sex, cholesterol levels, etc.)
- Second-order interactions $x_i \cdot x_j$ for $i < j$
- Predicting hemoglobin measure in one year

Using ordinary least squares, the linear model coefficients are

	Age	Sex	LDL	HDL	$\cdots$
Original data	0.05	-0.20	2.91	-2.75	$\cdots$
Perturbed data	0.05	-0.20	-2.62	2.18	$\cdots$

If you randomly perturb the 10 baseline measurements by just 1%, the coefficients can change dramatically.

Robustness View of Linear Regression

Account for uncertainty by considering $\mathbf{X} + \boldsymbol{\Delta}$ for all $\boldsymbol{\Delta} \in \mathcal{U} \subseteq \mathbb{R}^{n \times p}$

The set $\mathcal{U}$ is an **uncertainty set** which captures our belief about the noise in the data $\mathbf{X}$.

Objective:

$$
\begin{tikzcd}
\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_q \arrow[r] \arrow[d, squiggly] & \min\limits_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_q \arrow[d, squiggly] \\
\max\limits_{\boldsymbol{\Delta} \in \mathcal{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_q \arrow[r] & \min\limits_{\boldsymbol{\beta}} \max\limits_{\boldsymbol{\Delta} \in \mathcal{U}} \|(\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_q
\end{tikzcd}
$$

where $\|\beta\|_q := \left(\sum_i |\beta_i|^q\right)^{1/q}$ for $q \in [1, \infty]$.

Regularization View of Linear Regression

For $q, r \in \{1, 2\}$:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_q + \lambda \|\boldsymbol{\beta}\|_r$$

Examples:

- $q = r = 2$: regularized least squares (ridge regression)
- $q = 2,\ r = 1$: Lasso

Uncertainty sets

Some examples of possible uncertainty sets:

1. **Frobenius** norm sets:

$$\mathcal{U}_{F(q)} = \{\boldsymbol{\Delta} \in \mathbb{R}^{p\times n} \mid \|\boldsymbol{\Delta}\|_{q-F} \leq \lambda\},$$

   where $\|\boldsymbol{\Delta}\|_{q-F} := \left(\sum_{ij} |\Delta_{ij}|^q\right)^{1/q}$.

2. **Induced** norm sets:

$$\mathcal{U}_{I(r,q)} = \{\boldsymbol{\Delta} \in \mathbb{R}^{p\times n} \mid \|\boldsymbol{\Delta}\|_{r,q} \leq \lambda\},$$

   where $\|\boldsymbol{\Delta}\|_{r,q} := \max_{\mathbf{x}} \dfrac{\|\boldsymbol{\Delta}\mathbf{x}\|_q}{\|\mathbf{x}\|_r}$.

Example of Robust Problem

An example of robust linear regression problem:

$$\mathcal{U} = \mathcal{U}_{F(2)} = \{\boldsymbol{\Delta} \in \mathbb{R}^{n \times p} \mid \|\boldsymbol{\Delta}\|_{2-F} \leq \lambda\}$$

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\beta\|_q = \min_{\boldsymbol{\beta}} \max_{\substack{\widetilde{\mathbf{X}}: \\ \|\widetilde{\mathbf{X}} - \mathbf{X}\|_{2-F} \leq \lambda}} \|\mathbf{y} - \widetilde{\mathbf{X}}\beta\|_q$$

Perturbations $\boldsymbol{\Delta}$ constrained to have $\sum_{ij} \Delta_{ij}^2 \leq \lambda^2$.

Example of Robust Problem

Another example of robust linear regression problem:

$$\mathcal{U} = \mathcal{U}_{I(1,2)} = \{\mathbf{\Delta} \in \mathbb{R}^{n\times p} \mid \|\mathbf{\Delta}\|_{1,2} \le \lambda\} = \{\mathbf{\Delta} \mid \|\mathbf{\Delta x}\|_2 \le \lambda\|\mathbf{x}\|_1 \text{ for all } \mathbf{x}\}$$

Can show that

$$\mathcal{U} = \{\mathbf{\Delta} \mid \text{every column } \mathbf{\Delta}_i \text{ has } \|\mathbf{\Delta}_i\|_2 \le \lambda\}$$

Therefore,

$$\min_{\boldsymbol{\beta}} \max_{\mathbf{\Delta}\in\mathcal{U}} \|\mathbf{y} - (\mathbf{X} + \mathbf{\Delta})\boldsymbol{\beta}\|_q$$

allows for **feature-wise** perturbations $\mathbf{\Delta}$ (in contrast with $\mathcal{U}_{F(2)}$).

Equivalence of robustness and regularization

Theorem

1. *For* $\mathcal{U}_{F(q)} = \{\mathbf{\Delta} \in \mathbb{R}^{n\times p} \mid \|\mathbf{\Delta}\|_{q-F} \leq \lambda\}$,

$$\min_{\boldsymbol{\beta}} \max_{\mathbf{\Delta}\in\mathcal{U}_{F(q)}} \|\mathbf{y} - (\mathbf{X}+\mathbf{\Delta})\beta\|_q = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\beta\|_q + \lambda\|\boldsymbol{\beta}\|_{q^*},$$

   where $\frac{1}{q} + \frac{1}{q^*} = 1$.

2. *For* $\mathcal{U}_{I(r,q)} = \{\mathbf{\Delta} \in \mathbb{R}^{n\times p} \mid \|\mathbf{\Delta}\|_{r,q} \leq \lambda\}$,

$$\min_{\boldsymbol{\beta}} \max_{\mathbf{\Delta}\in\mathcal{U}_{I(r,q)}} \|\mathbf{y} - (\mathbf{X}+\mathbf{\Delta})\beta\|_q = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\beta\|_q + \lambda\|\boldsymbol{\beta}\|_r$$

Examples of Equivalence

Theorem

For $\mathcal{U}_{F(q)} = \{\boldsymbol{\Delta} \in \mathbb{R}^{n \times p} \mid \|\boldsymbol{\Delta}\|_{q-F} \leq \lambda\}$,

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{F(q)}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_q = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_q + \lambda \|\boldsymbol{\beta}\|_{q^*},$$

where $1/q + 1/q^* = 1$.

An example of equivalence:

$$\mathcal{U} = \mathcal{U}_{F(2)} = \{\boldsymbol{\Delta} \in \mathbb{R}^{n \times p} \mid \|\boldsymbol{\Delta}\|_{2-F} \leq \lambda\}$$

Using the theorem (with $q = 2$, so $q^* = 2$),

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2 = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \lambda \|\boldsymbol{\beta}\|_2.$$

Gives interpretation of ridge regression as protecting against **global** perturbations $\boldsymbol{\Delta}$ with $\left(\sum_{ij} \Delta_{ij}^2\right)^{1/2} \leq \lambda$.

Examples of Equivalence

Theorem

For $\mathcal{U}_{I(r,q)} = \{\boldsymbol{\Delta} \in \mathbb{R}^{n \times p} \mid \|\boldsymbol{\Delta}\|_{r,q} \leq \lambda\}$,

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{I(r,q)}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_q = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_q + \lambda\|\boldsymbol{\beta}\|_r$$

Another example of equivalence:

$\mathcal{U} = \mathcal{U}_{I(1,2)} = \{\boldsymbol{\Delta} \in \mathbb{R}^{n \times p} \mid \|\boldsymbol{\Delta}\|_{1,2} \leq \lambda\}$

Using the theorem (with $q = 2$ and $r = 1$),

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2 = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \lambda\|\boldsymbol{\beta}\|_1.$$

Gives interpretation of Lasso as protecting against **feature-wise** perturbations $\boldsymbol{\Delta}$.

Proof Idea

Focusing on case when $\mathcal{U} = \mathcal{U}_{I(r,q)}$ and loss function is ℓ_q.

- Using the norm properties, we have that
$$\|\mathbf{y} - (\mathbf{X} + \mathbf{\Delta})\beta\|_q \leq \|\mathbf{y} - \mathbf{X}\beta\|_q + \|\mathbf{\Delta}\beta\|_q.$$
- Since $\|\mathbf{\Delta}\beta\|_q \leq \|\mathbf{\Delta}\|_{r,q}\|\beta\|_r$ and for $\|\mathbf{\Delta}\|_{r,q} \leq \lambda$
$$\|\mathbf{\Delta}\beta\|_q \leq \lambda\|\beta\|_r.$$
- Thus,
$$\|\mathbf{y} - (\mathbf{X} + \mathbf{\Delta})\beta\|_q \leq \|\mathbf{y} - \mathbf{X}\beta\|_q + \lambda\|\beta\|_r.$$
- We can select a $\mathbf{\Delta}^0 \in \mathcal{U}$ such that
$$\|\mathbf{y} - (\mathbf{X} + \mathbf{\Delta}^0)\beta\|_q = \|\mathbf{y} - \mathbf{X}\beta\|_q + \lambda\|\beta\|_r.$$

  (Check for yourself!)
- Leads to
$$\max_{\mathbf{\Delta} \in \mathcal{U}} \|\mathbf{y} - (\mathbf{X} + \mathbf{\Delta})\beta\|_q = \|\mathbf{y} - \mathbf{X}\beta\|_q + \lambda\|\beta\|_r.$$

How do we solve the robust problems?

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \lambda\|\boldsymbol{\beta}\|_1$$

Rewrite as

$$\begin{array}{cc} \min & t + \lambda\|\boldsymbol{\beta}\|_1 \\ \text{subject to} & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 \leq t \end{array}$$

$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 \leq t$ is a quadratic constraint

$\|\boldsymbol{\beta}\|_1 = |\beta_1| + \ldots + |\beta_p|$ can be expressed with linear constraints using *auxiliary variables* $\mathbf{a}$:

$$\beta_j \leq a_j \quad \text{and} \quad -\beta_j \leq a_j.$$

Specialized R codes available as well.

A Cutting Plane Approach

We can use the equivalence theorem to instead solve robust problems using cutting planes.

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \lambda\|\boldsymbol{\beta}\|_1 = \min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{I(1,2)}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2$$

1. Pick some $\boldsymbol{\Delta}_1 \in \mathcal{U}$ and set $\mathcal{U}_1 = \{\boldsymbol{\Delta}_1\}$.
2. For $t \geq 1$, solve
$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_t} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2.$$
3. If solution $\boldsymbol{\beta}_t^*$ to Step 2 has
$$\max_{\boldsymbol{\Delta} \in \mathcal{U}_t} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}_t^*\|_2 < \max_{\boldsymbol{\Delta} \in \mathcal{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}_t^*\|_2,$$
then set $\mathcal{U}_{t+1} := \mathcal{U}_t \cup \{\boldsymbol{\Delta}_t^*\}$, where $\boldsymbol{\Delta}_t^* \in \operatorname*{argmax}_{\boldsymbol{\Delta} \in \mathcal{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}_t^*\|_2$ and go back to Step 2.

Real world data sets

- UCI Machine Learning Repository
- Data sizes:

Data set	n	p
Abalone	4177	9
Auto MPG	392	8
Comp Hard	209	7
Concrete	1030	8
Housing	506	13
Space shuttle	23	4
WPBC	46	32

Evaluation procedure

- Testing "Rob q-r" for $q, r \in \{1, 2\}$:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\beta\|_q + \lambda \|\boldsymbol{\beta}\|_r$$

- Training, Validation, testing sets: 50%, 25%, 25%.
- λ was chosen as the value giving the best mean squared prediction error on the validation set.

Effect of Robustness

Average out-of-sample mean squared error:

	Regular OLS	**Rob 1-1**	**Rob 1-2**	**Rob 2-1** Lasso	**Rob 2-2** Ridge
Abalone	5.74	5.67	5.65	5.63	**5.53**
Auto MPG	18.79	18.72	18.70	18.69	**18.58**
Comp Hard	2026.00	2014.32	1978.12	1965.75	**1925.13**
Concrete	132.47	131.46	131.32	131.08	**129.31**
Forest Fires	5526.00	5312.18	5229.14	**4994.81**	5266.40
Housing	39.80	39.54	39.49	39.42	**39.07**
Space shuttle	0.53	0.52	**0.51**	0.52	0.52
WPBC	4723.07	4676.20	4657.98	4630.19	**4489.20**

Robust Classification

Similar modifications can be made to create **robust** classification algorithms—address uncertainties in both features and labels.

- Binary classification problems:
  - Given data $(\mathbf{x}_i, y_i), i = 1, \ldots n$, with features $\mathbf{x}_i \in \mathbb{R}^p$ and labels $y_i \in \{-1, +1\}$;
  - find a function $f : \mathbb{R}^p \to \{-1, +1\}$ to classify new data points.
- Maximum Likelihood Estimator with logistic loss function

$$\max_{\boldsymbol{\beta}, \beta_0} \sum_{i=1}^{n} -\log(1 + e^{-y_i(\beta^T \mathbf{x}_i + \beta_0)}) \tag{1}$$

Summary

- Robustness improves regression.
- Robustness can be accomplished by adding regularization.
- The computational cost of achieving robustness is small.
- Regularized problems easily solvable in Jump.
- Can incorporate **both** sparsity and robustness into regression models (using techniques here and from Lecture 2).