Session 10: Linear regression

Optimization and Computational Linear Algebra for Data Science

Marylou Gabrié (based on material by Léo Miolane)

Contents

1. Ordinary least squares
2. Penalized linear regression
3. Matrix norms

Introduction

- We have n « feature vectors » $a_1, \dots, a_n \in \mathbb{R}^d$.
- Each point a_i comes with a « target variable » $y_i \in \mathbb{R}$.
- **Goal.** Find a linear relation between the a_is and the y_is:

- **Prediction:**

Can we have an intercept?

Find $x \in \mathbb{R}^d$ such that $y_i = \langle x, a_i \rangle + \mathbf{c}$.

Solving $Ax = y$ is a bad idea

The system $Ax = y$ may have:

- No solution.

- Infinitely many solutions.

1. Ordinary least squares

Least squares problem

(LS) Minimize $f(x) = \|Ax - y\|^2$ with respect to $x \in \mathbb{R}^d$.

The Moore-Penrose pseudo-inverse

What if $A^\top A$ is not invertible?

Definition

Let $A = U\Sigma V^\mathsf{T}$ be the SVD of A. The matrix $A^\dagger \stackrel{\text{def}}{=} V\Sigma' U^\mathsf{T}$ is called the **(Moore-Penrose) pseudo-inverse** of A, where $\Sigma' \in \mathbb{R}^{d\times n}$ is

$$\Sigma'_{i,i} = \begin{cases} 1/\Sigma_{i,i} & \text{if } \Sigma_{i,i} \neq 0 \\ 0 & \text{otherwise} \end{cases}, \quad \text{and} \quad \Sigma'_{i,j} = 0 \ \text{ for } \ i \neq j$$

Exercise: Check that if A is invertible then $A^{-1} = A^\dagger$.

Solving $A^\mathsf{T} Ax = A^\mathsf{T} y$

Claim: The vector $x^{\mathrm{LS}} \stackrel{\mathrm{def}}{=} A^\dagger y$ is a solution of $A^\mathsf{T} Ax = A^\mathsf{T} y$

Theorem

The set of the minimizers of $f(x) = \|Ax - y\|^2$ is

$$\left\{ x^{\mathrm{LS}} + v \,\middle|\, v \in \mathrm{Ker}(A) \right\}.$$

2. Penalized least squares

Ridge regression

Ridge regression consists in adding a « ℓ_2 penalty » :

(Ridge) Minimize $f(x) = \|Ax - y\|^2 + \lambda\|x\|^2$ w.r.t. $x \in \mathbb{R}^d$.

for some fixed $\lambda > 0$.

Lasso

The Lasso adds a « ℓ_1 penalty » :

(Lasso) Minimize $f(x) = \|Ax - y\|^2 + \lambda\|x\|_1$ w.r.t. $x \in \mathbb{R}^d$.

for some fixed $\lambda > 0$.

Intuition behind feature selection

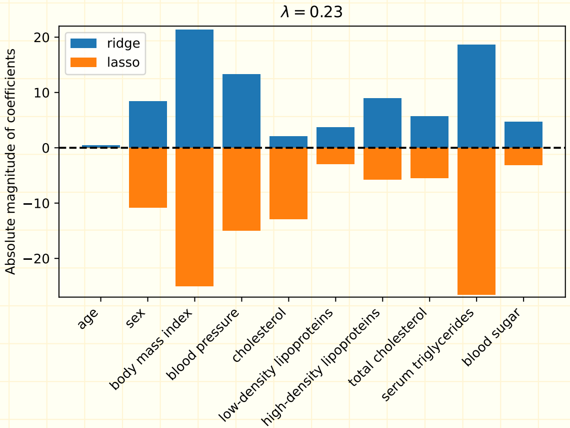
Lemma

Let x^{Lasso} be a minimizer of the Lasso cost function and let $r = \|x^{\text{Lasso}}\|_1$. Then x^{Lasso} is a solution to the constrained optimization problem:

$$\text{minimize} \quad \|Ax - y\|^2 \quad \text{subject to} \quad \|x\|_1 \leq r.$$

Effect of Regularization

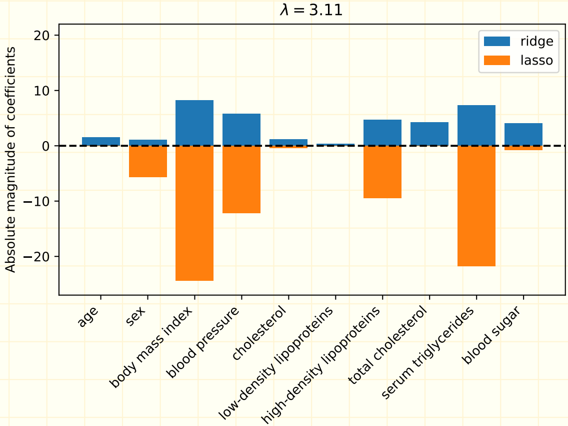
Consider a data set with $n = 442$ patients, with $d = 10$ dimensions feature vectors and the prediction of diabetese disease progression



https://scikit-learn.org/stable/datasets/toy_dataset.html

Effect of Regularization

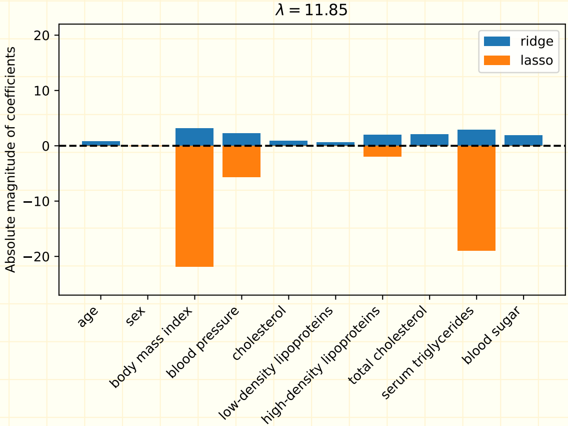
Consider a data set with $n = 442$ patients, with $d = 10$ dimensions feature vectors and the prediction of diabetese disease progression



https://scikit-learn.org/stable/datasets/toy_dataset.html

Effect of Regularization

Consider a data set with $n = 442$ patients, with $d = 10$ dimensions feature vectors and the prediction of diabetese disease progression



https://scikit-learn.org/stable/datasets/toy_dataset.html

3. Matrix norms

Frobenius norm

Definition

The Frobenius norm of a matrix $A \in \mathbb{R}^{n \times m}$ is defined as

$$\|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} A_{i,j}^2}$$

Proposition

$$\|A\|_F = \sqrt{\sum_{i=1}^{\min(n,m)} \sigma_i(A)^2}$$

The spectral norm

Definition

The spectral norm of a matrix $A \in \mathbb{R}^{n\times m}$ is defined as

$$\|A\|_{\mathrm{Sp}} = \max_{\|x\|=1} \|Ax\|.$$

Proposition

$$\|A\|_{\mathrm{Sp}} = \sigma_1(A).$$

The nuclear norm

Definition

The nuclear norm of a matrix $A \in \mathbb{R}^{n \times m}$ is defined as

$$\|A\|_\star = \sum_{i=1}^{\min(n,m)} \sigma_i(A).$$

Application to matrix completion

We have a data matrix $M \in \mathbb{R}^{n\times m}$ that we only observe partially. That is we only have access to

$$M_{i,j} \quad \text{for } (i,j) \in \Omega,$$

where $\Omega \subset \{1,\dots,n\} \times \{1,\dots m\}$ is a subset of the complete set of the entries.

Application to matrix completion

Questions?

Questions?