Session 10: Linear regression

Optimization and Computational Linear Algebra for Data Science

Léo Miolane

Contents

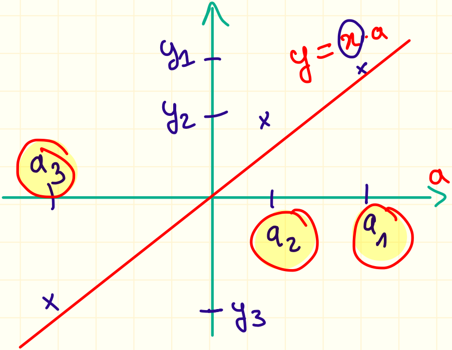
1. Ordinary least squares
2. Penalized linear regression
3. Matrix norms

Introduction

- We have n « feature vectors » $a_1, \dots, a_n \in \mathbb{R}^d$. ← d features
- Each point a_i comes with a « target variable » $y_i \in \mathbb{R}$.

Goal: find a linear relation between the a_i's and the y_i's

→ find $x \in \mathbb{R}^d$ such that $y_i \simeq \langle x, a_i \rangle$ for all i.



Can we have some intercept, that is $y_i \simeq \langle x, a_i \rangle + c$?

Yes We can add a '1' coordinate to the a_i → $\tilde{a}_i = \begin{pmatrix} a_i \\ 1 \end{pmatrix}$ ← d first coords

$\langle x, \tilde{a}_i \rangle = \langle x, a_i \rangle + x_{d+1}$ — c

Solving $Ax = y$ is a bad idea

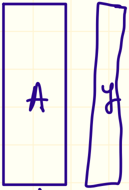
The system $Ax = y$ may have:

$$A = \begin{pmatrix} - a_1^T - \\ - a_n^T - \end{pmatrix} \in \mathbb{R}^{n \times d}$$

- No solution.

Example: if A is a "tall matrix" $(n > d)$

$\rightarrow \dim \operatorname{Im}(A) \leq d < n$ $\qquad \operatorname{Im}(A) \subset \mathbb{R}^n$



$\rightarrow$ y is not very likely to belong to $\operatorname{Im}(A)$ in practice.

$\rightarrow$ no solution

- Infinitely many solutions.

Example: if A is a "fat matrix" $d > n$

then $\dim \operatorname{Ker}(A) \geq d - n > 0$ $\qquad y \in \mathbb{R}^n$

$\rightarrow$ infinitely many solutions.

Ordinary least squares

Least squares problem

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

f is convex (HW 9) therefore

x minimizes $f \iff \nabla f(x) = 0$.

$\iff 2A^TAx - 2A^Ty = 0$

$\iff A^TAx = A^Ty$.

Conclusion: the minimisers of f are exactly the solutions of the linear system $A^TAx = A^Ty$

If A^TA is invertible $\rightarrow x = (A^TA)^{-1}A^Ty$.

What if A^TA is not invertible?

The Moore-Penrose pseudo-inverse

Definition

Let $A = U\Sigma V^\top$ be the SVD of A. The matrix $A^\dagger \stackrel{\text{def}}{=} V\Sigma' U^\top$ is called the (Moore-Penrose) pseudo-inverse of A, where Σ' is the $d \times n$ matrix given by

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

and $\Sigma'_{i,j} = 0$ for $i \neq j$.

$$A = U \begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} V^T \in \mathbb{R}^{n \times d}$$

Exercise: check that if A invertible, then $A^{-1} = A^+$

$$A^+ = V \begin{pmatrix} 1/\sigma_1 & & & & & \\ & \ddots & & & & \\ & & 1/\sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} U^T \in \mathbb{R}^{d \times n}$$

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

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

$$A^T A x^{LS} = V \Sigma^T \cancel{U^T U} \Sigma \cancel{V^T V} \Sigma^\dagger U^T y$$

$$= V \underbrace{\Sigma^T \Sigma \Sigma^\dagger}_{= \Sigma^T} U^T y = V \Sigma^T U^T y = A^T y.$$

Theorem

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

$$A^\dagger y + \underbrace{\mathrm{Ker}(A)}_{\mathrm{Ker}(A^T A)} = \left\{ x^{\mathrm{LS}} + v \,\middle|\, v \in \mathrm{Ker}(A) \right\}.$$

Penalized least squares

Ridge regression

Ridge regression consists in adding a « ℓ_2 penalty » :

(Ridge) Minimize $f(x) = \frac{1}{2}\|Ax-y\|^2 + \frac{\lambda}{2}\|x\|^2$ w.r.t. $x \in \mathbb{R}^d$

for some fixed $\lambda > 0$.

$x_1^2 + \dots + x_d^2$

- f is strongly convex, it admits a unique minimizer

$$\boxed{x^{Ridge} = (A^T A + \lambda\, Id)^{-1} A^T y.}$$

- Why adding the ℓ_2-penalty?

Trade-off :
- this promotes vectors of small norm $\|x^{Ridge}\| \le \|x^{LS}\|$ (exercise!)
- issue : $\|Ax^{Ridge} - y\|^2 \geq \|Ax^{LS} - y\|^2$

$\langle x, a_{new} \rangle = x_1 \cdot a_{new,1} + \dots + x_d \cdot a_{new,d}$

Lasso

The Lasso adds a « ℓ_1 penalty » :

$g(t) = \frac{t^2}{2}$ $g'(t) = t$

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

for some fixed $\lambda > 0$.

- f is not strictly convex in general : there is not a unique minimizer a priori.
- In practice, the minimizer x^{Lasso} is unique.

Why do we add this ℓ_1-penalty ?

→ it promotes sparse vectors x^{Lasso} (lots of coefficients of x^{Lasso} are likely to be zero).

→ Feature selection !

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:

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

Proof: By contradiction, assume that there exists x such that

$$\|Ax - y\|^2 < \|Ax^{\text{Lasso}} - y\|^2$$

$$\|x\|_1 \leq r = \|x^{\text{Lasso}}\|_1$$

$$\rightarrow \|Ax - y\|^2 + \lambda \|x\|_1 < \|Ax^{\text{Lasso}} - y\|^2 + \lambda \|x^{\text{Lasso}}\|_1$$

$\rightarrow$ Contradiction

Application: compressed sensing

- In homework 4 we have seen that we can compress images very well.
- Most of the data can be thrown away !

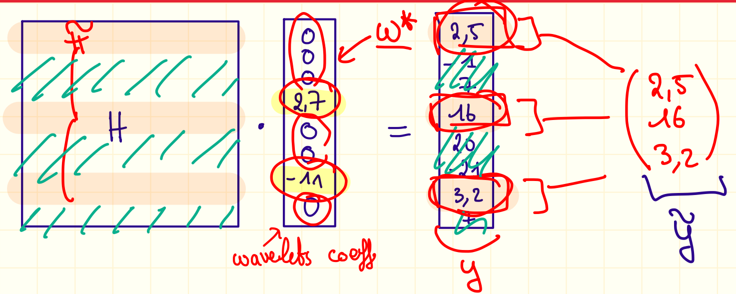
$$H \cdot \begin{pmatrix} 0 \\ 0 \\ 0 \\ 2{,}7 \\ 0 \\ 0 \\ -11 \\ 0 \end{pmatrix} = \begin{pmatrix} 2{,}5 \\ -1 \\ 7 \\ 16 \\ 20 \\ -21 \\ 3{,}2 \\ 7 \end{pmatrix}$$

wavelets coeff

pixels of the image.

Can we directly measure only the useful wavelet coefficients ?

Application: compressed sensing



① Measure only a small fraction of the pixels.

② We have $\boxed{\tilde{H} w^* = \tilde{y}}$ $\boxed{\tilde{H}}$

③ Minimize $f(w) = \|\tilde{H} w - \tilde{y}\|^2 + \lambda \|w\|_1$

to get w^{Lasso} which should be a good estimate of w^*

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

$$A = U \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \ddots \end{pmatrix} V^T$$

$$\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$$

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

$$\|A\|_F^2 = \mathrm{Tr}(AA^T) = \mathrm{Tr}(U\Sigma \cancel{V^T V}\Sigma^T U^T)$$

$$= \mathrm{Tr}(\cancel{U^T U}\Sigma\Sigma^T) = \sigma_1^2 + \dots + \sigma_r^2$$

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

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

largest singular value of A.

$A = U \Sigma V^T$

$A^T A = V \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \end{pmatrix} V^T$

Proof: $\|A\|_{Sp}^2 = \max_{\|x\|=1} \|Ax\|^2$

$= \max_{\|x\|=1} x^T A^T A x$

$= \lambda_1(A^T A) = \sigma_1(A)^2$

The nuclear norm

Definition

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

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

$\longrightarrow$ "ℓ_1 - norm of the singular values"

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

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

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

NP-HARD

→ minimize $\operatorname{rank}(X)$ with respect to $X \in \mathbb{R}^{n \times m}$

verifying $X_{i,j} = M_{i,j}$ for all $(i,j) \in \Omega$

it has been proposed to solve instead:

$$\boxed{\text{minimize} \quad \|X\|_* \quad \text{with respect to } X \in \mathbb{R}^{n \times m} \text{ verifying } X_{i,j} = M_{i,j} \text{ for all } (i,j) \in \Omega}$$

Application to matrix completion

Questions?

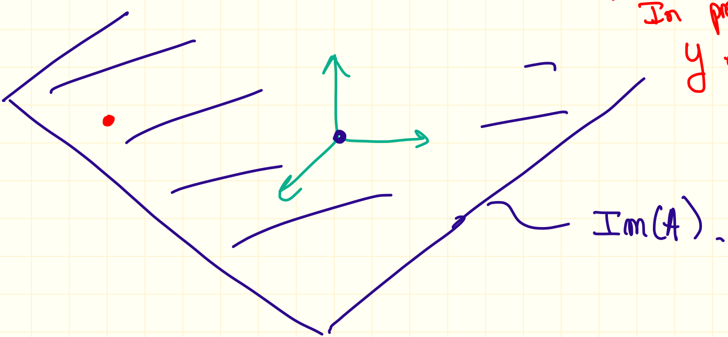
$$A = \begin{array}{cc} \overleftrightarrow{\;2\;} & \\ \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} & \updownarrow 3 \end{array}$$

$\mathrm{Im}(A)$ is a subspace of $\mathbb{R}^3$ of dimension at most $\underline{\underline{2}}$

Assume $\underline{\dim \mathrm{Im}(A) = 2}$

"In practice" $y \notin \mathrm{Im}(A)$

$\mathbb{R}^3$



Questions?