Optimization and Computational Linear Algebra for Data Science
Lecture 11: Linear regression, matrix completion

Léo Miolane · leo.miolane@gmail.com

August 20, 2019

Warning: *This material is not meant to be lecture notes. It only gathers the main concepts and results from the lecture, without any additional explanation, motivation, examples, figures...*

1 Least squares

Assume that we are given point $a_i = (a_{i,1}, \dots, a_{i,d}) \in \mathbb{R}^d$ with labels $y_i \in \mathbb{R}$ for $i = 1 \dots n$. We aim at finding a vector $x \in \mathbb{R}^d$ such that

$$y_i \simeq \langle a_i, x \rangle = \sum_{j=1}^{d} a_{i,j} x_j, \qquad \text{for} \quad i = 1 \dots n.$$

If we denote by A the $n \times d$ matrix whose rows are $a_1, \dots, a_n$, i.e. $A_{i,j} = a_{i,j}$, we are looking for some x such that $Ax \simeq y$.

1.1 Solving the system $Ax = y$

As we have seen in Lecture 2, we can distinguish two cases:

- If $y \notin \mathrm{Im}(A)$ then the equation $Ax = y$ does not admit any solution (by definition of $\mathrm{Im}(A)$).
- If $y \in \mathrm{Im}(A)$ then the equation $Ax = y$ admits at least a solution x_0 (by definition of $\mathrm{Im}(A)$). Moreover, the set of (all) solutions is

  $$x_0 + \mathrm{Ker}(A) = \{x_0 + v \,|\, v \in \mathrm{Ker}(A)\}.$$

  In particular, if $\mathrm{Ker}(A) = \{0\}$ then the equation admits a unique solution.

In the second case, one can obtain an expression for a particular solution x_0 using the SVD of A. Let $r = \mathrm{rank}(A)$, $\sigma_1, \sigma_2, \dots, \sigma_r > 0$ be the non-zero singular values of A and $\Sigma = \mathrm{Diag}(\sigma_1, \dots, \sigma_r)$. Finally, let $A = U \Sigma V^{\mathsf{T}}$ be the SVD of A, where $V \in \mathbb{R}^{n \times r}$ and $U \in \mathbb{R}^{d \times r}$ are matrices that have orthonormal columns.

Notice that $V^{\mathsf{T}} V = \mathrm{Id}$ and that UU^{T} is the orthogonal projection on $\mathrm{Im}(A)$. Hence, if we let $x_0 = V \Sigma^{-1} U^{\mathsf{T}} y$, we have

$$Ax_0 = U \Sigma V^{\mathsf{T}} V \Sigma^{-1} U^{\mathsf{T}} y = UU^{\mathsf{T}} y = y$$

because we assumed that $y \in \mathrm{Im}(A)$. This motivates the following definition:

Definition 1.1 (*Moore-Penrose pseudo-inverse*)

The matrix $A^{\dagger} \overset{\text{def}}{=} V \Sigma^{-1} U^{\mathsf{T}}$ is called the (Moore-Penrose) pseudo-inverse of A.

Notice that in the case where A is invertible, $A^{\dagger} = A^{-1}$. From the analysis above, we deduce:

Proposition 1.1

The set of solution of the linear system $Ax = y$ is

- $\emptyset$ *if* $y \notin \mathrm{Im}(A)$.
- $A^{\dagger} y + \mathrm{Ker}(A)$ *otherwise.*

1.2 Least squares

In general, there is no reason for y to belong to $\mathrm{Im}(A)$, especially when $n > d$. (Exercise: why?) Therefore one is rather interested by solving

$$\min_{x \in \mathbb{R}^d} \|Ax - y\|^2. \tag{1}$$

The function $f : x \mapsto \|Ax - y\|^2$ is convex (Exercise: why?) and differentiable. Hence x is solution of (1) if and only if $\nabla f(x) = 0$. Compute

$$f(x) = (Ax - y)^{\top}(Ax - y) = x^{\top} A^{\top} A x - 2 y^{\top} A x + \|y\|^2.$$

Hence $\nabla f(x) = 2A^{\top}Ax - 2A^{\top}y$. We conclude

$$x \ \text{ is solution of } \ (1) \qquad \Longleftrightarrow \qquad A^{\top}Ax = A^{\top}y.$$

If $A^{\top}A$ is invertible there is a unique minimizer $x^* = (A^{\top}A)^{-1}A^{\top}y$. In the general case, we see that the solutions of (1) are the solutions of the linear system $A^{\top}Ax = A^{\top}y$. From Proposition 1.1 we get that the solutions of (1) are

$$(A^{\top}A)^{\dagger}A^{\top}y + \mathrm{Ker}(A^{\top}A).$$

This expression simplifies a lot. First (exercise!) we have $\mathrm{Ker}(A^{\top}A) = \mathrm{Ker}(A)$. Then if we let $A = U\Sigma V^{\top}$ be the SVD of A, we have

$$A^{\top}A = V\Sigma^2 V^{\top}.$$

$V\Sigma^2V^{\top}$ is therefore the SVD of $A^{\top}A$. Hence $(A^{\top}A)^{\dagger} = V\Sigma^{-2}V^{\top}$. This gives $(A^{\top}A)^{\dagger}A^{\top} = V\Sigma^{-2}V^{\top}V\Sigma U^{\top} = A^{\dagger}$. We conclude:

Proposition 1.2

The set of solution of the minimization problem $\min_{x \in \mathbb{R}^n} \|Ax - y\|^2$ *is*

$$A^{\dagger}y + \mathrm{Ker}(A).$$

2 Penalized least squares: Ridge regression and Lasso

3 Norms for matrices

4 Low-rank matrix estimation and matrix completion

Further reading

References