

Session 10: Linear regression

Optimization and Computational Linear Algebra for Data Science

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 » $y_i \in \mathbb{R}$.

Solving $Ax = y$ is a bad idea

The system $Ax = y$ may have:

- ❑ No solution.
- ❑ Infinitely many solutions.

Ordinary least squares

Least squares problem

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

The Moore-Penrose pseudo-inverse

Definition

Let $A = U\Sigma V^T$ be the SVD of A . The matrix $A^\dagger \stackrel{\text{def}}{=} V\Sigma'U^T$ 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 $i \neq j$.

Solving $A^\top Ax = A^\top y$

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

Theorem

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

$$A^\dagger y + \text{Ker}(A) = \left\{ x^{\text{LS}} + v \mid v \in \text{Ker}(A) \right\}.$$

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. $x \in \mathbb{R}^n$

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. $x \in \mathbb{R}^n$

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:

$$\text{minimize } \frac{1}{2} \|Ax - y\|^2 \quad \text{subject to } \|x\|_1 \leq r.$$

Intuition behind feature selection

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

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

Proposition

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

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

Application to matrix completion

Questions?

Questions?