Ordinary least squares

DS-GA 1013 / MATH-GA 2824 Mathematical Tools for Data Science

https://cims.nyu.edu/~cfgranda/pages/MTDS_spring20/index.html

Carlos Fernandez-Granda

Prerequisites

Linear algebra (vectors, matrices)

Mean-squared-error estimation

Goal

Derive ordinary-least-squares estimator (in two different ways)

Regression

Goal: Estimate response (or dependent variable)

Data: Several observed variables, known as features (or covariates, or independent variables)

Probabilistic perspective

Response: random variable $\tilde{y}$

Features: random vector $\tilde{x}$

Linear minimum MSE estimator

$$\Sigma_{\tilde{x}}^{-1}\Sigma_{\tilde{x}\tilde{y}} = \arg\min_{\beta} \mathrm{E}\left[(\tilde{y} - \tilde{x}^T \beta)^2\right]$$

We need to compute covariance and cross-covariance from data!

Data

Training data: (y_1, x_1), (y_2, x_2), $\ldots$, (y_n, x_n), where $y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$

We define a response vector $y \in \mathbb{R}^n$ and a feature matrix

$$X := \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$$

Estimation via averaging

$$\Sigma_{\tilde{x}} \approx \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T = \frac{1}{n} X X^T$$

$$\Sigma_{\tilde{y}\tilde{x}} \approx \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i[1] y_i \\ \frac{1}{n}\sum_{i=1}^{n} x_i[2] y_i \\ \cdots \\ \frac{1}{n}\sum_{i=1}^{n} x_i[p] y_i \end{bmatrix} = \frac{1}{n} X y$$

$$\Sigma_{\tilde{x}}^{-1}\Sigma_{\tilde{y}\tilde{x}} \approx (XX^T)^{-1} X y$$

Ordinary least squares cost function

Reasonable cost function beyond probabilistic assumptions

$$\begin{aligned}\beta_{\text{OLS}} &:= \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i^T \beta \right)^2 \\ &= \arg\min_{\beta} \|y - X^T \beta\|_2^2 \\ &= \arg\min_{\beta} \beta^T X X^T \beta - 2 y^T X^T \beta + y^T y\end{aligned}$$

Quadratic form

If XX^T is positive definite, then solution is point where gradient is zero

Ordinary least squares

If X is full rank, for any $v \neq 0$

$$v^T XX^T v = ||Xv||_2^2 > 0$$

so XX^T is positive definite

$$\nabla f(\beta) = 2XX^T\beta - 2Xy$$

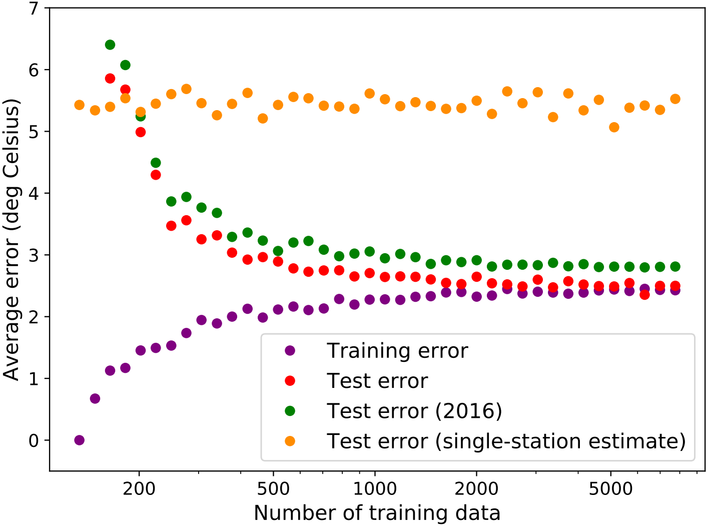
Setting to zero yields

$$\beta_{\text{OLS}} = (XX^T)^{-1}Xy$$

Temperature prediction via linear regression

- Dataset of hourly temperatures measured at weather stations all over the US
- Goal: Predict temperature in Yosemite from other temperatures
- Response: Temperature in Yosemite
- Features: Temperatures in 133 other stations ($p = 133$) in 2015
- Test set: 10^3 measurements
- Additional test set: All measurements from 2016

Results


Average error (deg Celsius)
Number of training data
0
1
2
3
4
5
6
7
200
500
1000
2000
5000
Training error
Test error
Test error (2016)
Test error (single-station estimate)

What have we learned?

OLS estimator can be derived from linear minimum MSE estimator or from least-squares cost function