MATH529 – Fundamentals of Optimization
Unconstrained Optimization VIII

Marco A. Montes de Oca

Mathematical Sciences, University of Delaware, USA

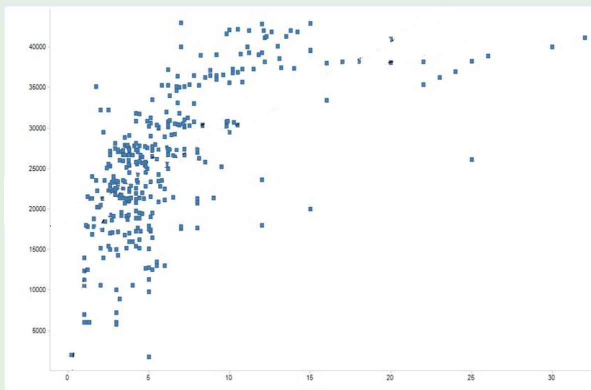
Linear Least Squares

Linear Least Squares

Example

Suppose you use Mechanical Turk (`https://www.mturk.com`) to crowdsource a small research project whose aim is to discover how the salary of actuaries evolves over time after graduation. To do this, you ask volunteers to give you information about their careers.

Your collected data look like this:



Linear Least Squares

Suppose you want to fit a polynomial function

$$f(t) = x_0 + x_1 t + x_2 t^2 + \ldots + x_n t^n$$

where n is selected in advance.

How to choose the coefficients x_i, $i = \{0, 2, \ldots, n\}$?

Linear Least Squares

A common approach is to cast the problem as an optimization problem with an objective function with the following form:

$$\phi(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{m}(r_i(\mathbf{x}))^2$$

where each $r_i : \mathbb{R}^n \rightarrow \mathbb{R}$ is a smooth function referred to as residual and $m \geq n$.

Linear Least Squares

If

$$r_i(\mathbf{x}) = \left(s_i - \sum_{j=0}^{n} x_j t_i^j\right)^2$$

where s_i is the i-th data sample, then

$$\phi(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{m}\left(s_i - \sum_{j=0}^{n} x_j t_i^j\right)^2$$

which can be written in matrix form as follows:

Linear Least Squares

Let

$$A = \begin{pmatrix} 1 & t_1 & t_1^2 & \dots & t_1^n \\ 1 & t_2 & t_2^2 & \dots & t_2^n \\ & & \vdots & & \\ 1 & t_m & t_m^2 & \dots & t_m^n \end{pmatrix}$$

and

$$\mathbf{s} = (s_1, s_2, s_3, \dots, s_m)^T$$

then

$$\phi(\mathbf{x}) = \frac{1}{2}||\mathbf{s} - A\mathbf{x}||^2$$

Linear Least Squares

$$\phi(\mathbf{x}) = \frac{1}{2}||\mathbf{s} - A\mathbf{x}||^2 = \frac{1}{2}(\mathbf{s} - A\mathbf{x})^T(\mathbf{s} - A\mathbf{x}) =$$

$$\frac{1}{2}\left(\mathbf{s}^T\mathbf{s} - 2\mathbf{s}^T A\mathbf{x} + (A\mathbf{x})^T A\mathbf{x}\right) =$$

$$\frac{1}{2}\mathbf{s}^T\mathbf{s} - (A^T\mathbf{s})^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T A^T A\mathbf{x}$$

The gradient and Hessian of ϕ are therefore:

$$\nabla\phi(\mathbf{x}) = -A^T\mathbf{s} + A^T A\mathbf{x}\,; \quad H\phi(\mathbf{x}) = A^T A$$

Remember homework exercise? $H\phi(\mathbf{x})$ is positive semidefinite. Therefore $\phi(\mathbf{x})$ is convex!

The global minimizer is therefore $\nabla\phi(\mathbf{x}^\star) = 0$, so $A^T A\mathbf{x}^\star = A^T\mathbf{s}$ or $\mathbf{x}^\star = (A^T A)^{-1}A^T\mathbf{s}$.

Demonstration

Problem: $\min \frac{1}{2}||\mathbf{r}(\mathbf{x})||^2$, where $\mathbf{r}(\mathbf{x}) = (r_1(\mathbf{x}), r_2(\mathbf{x}), \ldots, r_m(\mathbf{x}))^T$.

If $\mathbf{r}(\mathbf{x}) = \mathbf{s} - A\mathbf{x}$, then the solution of the system must satisfy

$$A^T A\mathbf{x}^\star = A^T\mathbf{s}$$

These equations are called **Normal equations**.

Linear Least Squares

Three solutions for these equations:

- Direct computation (cheapest, numerically unstable).
- QR factorization.
- Singular value decomposition. (expensive, more stable).

Linear Least Squares

QR factorization:

Let Q be an orthogonal matrix (i.e., its columns are orthonormal). Then $Q^T Q = I$.

Orthogonal transformations do not change a vector's norm: $||Q\mathbf{x}||^2 = (Q\mathbf{x})^T (Q\mathbf{x}) = \mathbf{x}^T Q^T Q\mathbf{x} = \mathbf{x}^T\mathbf{x} = ||\mathbf{x}||^2$.

Thanks to this property, the solution to the least squares problem remains unchanged after an orthogonal transformation.

Linear Least Squares

Let $A = QR$, where Q is an orthogonal matrix, and R is an upper triangular matrix.

This factorization transforms the normal equations as follows:

$$A^T A\mathbf{x}^\star = A^T\mathbf{s} = (QR)^T(QR)\mathbf{x}^\star = (QR)^T\mathbf{s}$$

$$R^T Q^T QR\mathbf{x}^\star = (QR)^T\mathbf{s}$$

$$R^T R\mathbf{x}^\star = (QR)^T\mathbf{s}$$

$$\mathbf{x}^\star = (R^T R)^{-1}(QR)^T\mathbf{s}$$

$$\mathbf{x}^\star = R^{-1}(R^T)^{-1}R^T Q^T\mathbf{s}$$

$$\mathbf{x}^\star = R^{-1}Q^T\mathbf{s}$$

Linear Least Squares

Singular Value Decomposition (SVD):

The SVD of A is given by $A = USV^T$, where U, and V are orthogonal matrices, and S is a diagonal matrix with diagonal elements $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$.

Substituting $A = USV^T$ in the normal equations, we obtain:

$$\mathbf{x}^\star = VS^{-1}U^T\mathbf{s} = \sum_{i=1}^{n} \frac{\mathbf{u}_i^T\mathbf{s}}{\sigma_i}\mathbf{v_i}$$

Least Squares in General

In least squares problems we want to minimize the objective function

$$f(\mathbf{x}) = \frac{1}{2}||\mathbf{r}(\mathbf{x})||^2 \, ,$$

where $\mathbf{r}(\mathbf{x}) = (r_1(\mathbf{x}), r_2(\mathbf{x}), \ldots, r_m(\mathbf{x}))^T$.
The gradient of f is given by

$$\nabla f(\mathbf{x}) = \sum_{i=1}^{m} r_i(\mathbf{x}) \nabla r_i(\mathbf{x}) = J(\mathbf{x})^T \mathbf{r}(\mathbf{x}) \, ,$$

where $J(\mathbf{x})$ is the Jacobian of $\mathbf{r}(\mathbf{x})$

$$\begin{pmatrix} \nabla r_1(\mathbf{x})^T \\ \nabla r_2(\mathbf{x})^T \\ \vdots \\ \nabla r_m(\mathbf{x})^T \end{pmatrix}$$

Least Squares in General

The Hessian of f is given by

$$Hf(\mathbf{x}) = \sum_{i=1}^{m} \nabla r_i(\mathbf{x}) \nabla r_i(\mathbf{x})^T + \sum_{i=1}^{m} r_i(\mathbf{x}) Hr_i(\mathbf{x})$$

or

$$Hf(\mathbf{x}) = J(\mathbf{x})^T J(\mathbf{x}) + \sum_{i=1}^{m} r_i(\mathbf{x}) Hr_i(\mathbf{x})$$