Math 484: Nonlinear Programming[1] Mikhail Lavrov

Chapter 4, Lecture 2: Least-squares fit

February 22, 2019 *University of Illinois at Urbana-Champaign*

1 The general least-squares problem

We can generalize the problems we looked at in the previous lecture in at least two ways:

1. We are given a set of points $(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)$ and a degree d. We want to find a polynomial $f(x)$ of degree at most d whose values $f(x_1), f(x_2), \ldots, f(x_k)$ are as close as possible to $y_1, y_2, \ldots, y_k$.
2. We are given k inputs $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(k)} \in \mathbb{R}^n$ and k outputs $y_1, y_2, \ldots, y_k \in \mathbb{R}$. We want to find a linear function from $\mathbb{R}^n$ to $\mathbb{R}$ (that is, a function of the form $f(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x} + b$) whose values $f(\mathbf{x}^{(1)}), f(\mathbf{x}^{(2)}), \ldots, f(\mathbf{x}^{(k)})$ are as close as possible to $y_1, y_2, \ldots, y_k$.

In both cases, "as close as possible" means we take the sum of squares of the errors, as we did before.

These are really the same problem in disguise. In the first problem, if we write $f(x) = a_d x^d + a_{d-1}x^{d-1} + \cdots + a_0$, then we want to minimize the distance between the vector

$$\begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_k) \end{bmatrix} = \begin{bmatrix} a_d x_1^d + a_{d-1}x_1^{d-1} + \cdots + a_0 \\ a_d x_2^d + a_{d-1}x_2^{d-1} + \cdots + a_0 \\ \vdots \\ a_d x_k^d + a_{d-1}x_k^{d-1} + \cdots + a_0 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^d \\ 1 & x_2 & x_2^2 & \cdots & x_2^d \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_k & x_k^2 & \cdots & x_k^d \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_d \end{bmatrix}$$

and a given vector $\mathbf{y} = (y_1, y_2, \ldots, y_k)$. In the second problem we want to minimize the distance between the vector

$$\begin{bmatrix} f(\mathbf{x}^{(1)}) \\ f(\mathbf{x}^{(2)}) \\ \vdots \\ f(\mathbf{x}^{(k)}) \end{bmatrix} = \begin{bmatrix} \mathbf{a} \cdot \mathbf{x}^{(1)} + b \\ \mathbf{a} \cdot \mathbf{x}^{(2)} + b \\ \vdots \\ \mathbf{a} \cdot \mathbf{x}^{(k)} + b \end{bmatrix} = \begin{bmatrix} \text{—}\mathbf{x}^{(1)\mathsf{T}}\text{—} & 1 \\ \text{—}\mathbf{x}^{(2)\mathsf{T}}\text{—} & 1 \\ \vdots & \vdots \\ \text{—}\mathbf{x}^{(k)\mathsf{T}}\text{—} & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \\ b \end{bmatrix}$$

and a given vector $\mathbf{y} = (y_1, y_2, \ldots, y_k)$.

So a common generalization of both problems is to minimize the function $\|A\mathbf{x} - \mathbf{y}\|$ over all $\mathbf{x} \in \mathbb{R}^n$, given a matrix A and a vector $\mathbf{y}$.

This is sometimes called the problem of solving an overconstrained system of equations, because if the system

$$A\mathbf{x} = \mathbf{y}$$

[1] This document comes from the Math 484 course webpage: https://faculty.math.illinois.edu/~mlavrov/courses/484-spring-2019.html

had a solution $\mathbf{x}$, we could use that solution to give $\|A\mathbf{x} - \mathbf{y}\|$ the smallest possible value: 0. So the interesting case is where the system $A\mathbf{x} = \mathbf{y}$ is overconstrained: it has no solutions. Finding a vector $\mathbf{x}$ that minimizes $\|A\mathbf{x} - \mathbf{y}\|$ is like finding the best approximate solution of the system $A\mathbf{x} = \mathbf{y}$.

2 A geometric characterization of the solution

We can rephrase the problem geometrically. If A is an $m \times n$ matrix, then the set $V = \{A\mathbf{x} : \mathbf{x} \in \mathbb{R}^n\}$ is a subspace of $\mathbb{R}^m$: a line or plane or some other higher-dimensional object. Minimizing $\|A\mathbf{x} - \mathbf{y}\|$ means finding the point of V closest to a given vector $\mathbf{y} \in \mathbb{R}^m$.

A geometric intuition says that this point should be obtained by dropping a perpendicular from $\mathbf{y}$ onto V, whatever that looks like. Here is a lemma that makes that precise:

Lemma 2.1. *If V is the subspace $\{A\mathbf{x} : \mathbf{x} \in \mathbb{R}^n\}$, then the point $A\mathbf{x}^* \in V$ is the closest point of V to $\mathbf{y} \in \mathbb{R}^m$ if and only if*

$$A\mathbf{x}^* - \mathbf{y} \perp \mathbf{a}$$

for all $\mathbf{a} \in V$.

Proof. Let's remember the technique we used in Chapter 1 of taking the one-dimensional restriction of a function.

Here, we are minimizing $f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{y}\|^2$ for $\mathbf{x} \in \mathbb{R}^n$. (We square the distance to make the expression simpler, which doesn't change where the distance is minimized.) So let's fix $\mathbf{x}^* \in \mathbb{R}^n$ and define

$$\phi_{\mathbf{u}}(t) = f(\mathbf{x}^* + t\mathbf{u}) = \|A(\mathbf{x}^* + t\mathbf{u}) - \mathbf{y}\|^2.$$

We know that $\mathbf{x}^*$ is a global minimizer of f if and only if for every direction $\mathbf{u}$, 0 is a global minimizer of $\phi_{\mathbf{u}}$.

Let's take a closer look at $\phi_{\mathbf{u}}(t)$. Define $\mathbf{a} = A\mathbf{u}$. (If $\mathbf{u}$ is an arbitrary element of $\mathbb{R}^n$, then $\mathbf{a}$ is an arbitrary element of V.) Then

$$\begin{aligned}
\phi_{\mathbf{u}}(t) &= \|A\mathbf{x}^* - \mathbf{y} + t\mathbf{a}\|^2 \\
&= (A\mathbf{x}^* - \mathbf{y} + t\mathbf{a}) \cdot (A\mathbf{x}^* - \mathbf{y} + t\mathbf{a}) \\
&= (A\mathbf{x}^* - \mathbf{y}) \cdot (A\mathbf{x}^* - \mathbf{y}) + 2t(A\mathbf{x}^* - \mathbf{y}) \cdot \mathbf{a} + t^2(\mathbf{a} \cdot \mathbf{a}) \\
&= \|A\mathbf{x}^* - \mathbf{y}\|^2 + 2t(A\mathbf{x}^* - \mathbf{y}) \cdot \mathbf{a} + t^2\|\mathbf{a}\|^2.
\end{aligned}$$

So now we recognize this as a parabola.

Parabolas $c_2t^2 + c_1t + c_0$ are minimized at $t = 0$ precisely when $c_2 \geq 0$ and $c_1 = 0$: upward-pointing parabolas symmetric about the vertical axis. Here, $c_2 \geq 0$ for a fact: it's a perfect square. So this parabola is minimized at $t = 0$ if and only if, for every $\mathbf{a} \in V$,

$$2(A\mathbf{x}^* - \mathbf{y}) \cdot \mathbf{a} = 0,$$

or $A\mathbf{x}^* - \mathbf{y} \perp \mathbf{a}$, which is the condition we wanted.

Another way to see this is to take the derivative $\phi'_{\mathbf{u}}(0) = 2(A\mathbf{x}^* - \mathbf{y}) \cdot \mathbf{a}$. This is 0 exactly when $A\mathbf{x}^* - \mathbf{y} \perp \mathbf{a}$, and we get the same conclusion. (Here, $\phi_{\mathbf{u}}$ is convex, so 0 is a global minimizer if and only if it is a critical point.) □

We can use this lemma to write down a system of linear equations to solve for $\mathbf{x}^*$, called the *normal equation*:

Theorem 2.1. *A point $\mathbf{x}^* \in \mathbb{R}^n$ minimizes $\|A\mathbf{x} - \mathbf{y}\|$ if and only if*

$$A^\mathsf{T} A\mathbf{x}^* = A^\mathsf{T}\mathbf{y}.$$

Proof. Let $\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(n)}$ be the columns of A. These are elements of the subspace $V = \{A\mathbf{x} : \mathbf{x} \in \mathbb{R}^n\}$: we can write $\mathbf{a}^{(i)}$ as $A\mathbf{e}^{(i)}$, where $\mathbf{e}^{(i)}$ is the i^{th} standard basis vector.

By the lemma, if $A\mathbf{x}^*$ is the closest point of V to $\mathbf{y}$, then $A\mathbf{x}^* - \mathbf{y} \perp \mathbf{a}^{(i)}$, or

$$\mathbf{a}^{(i)} \cdot (A\mathbf{x}^* - \mathbf{y}) = 0,$$

for $i = 1, 2, \ldots, n$. Also, any element of V is a linear combination of $\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(n)}$, so these n dot products capture the full content of the lemma: if all of them are 0, then $A\mathbf{x}^* - \mathbf{y} \perp \mathbf{a}$ for all $\mathbf{a} \in V$, and $A\mathbf{x}^*$ is the closest point of V to $\mathbf{y}$.

To get a prettier-looking statement at the end, we now rephrase the condition "all n of these dot products are 0" to turn it into the equation in the theorem. Stacking these dot products on top of each other, we have

$$\begin{bmatrix} \mathbf{a}^{(1)} \cdot (A\mathbf{x}^* - \mathbf{y}) \\ \mathbf{a}^{(2)} \cdot (A\mathbf{x}^* - \mathbf{y}) \\ \vdots \\ \mathbf{a}^{(n)} \cdot (A\mathbf{x}^* - \mathbf{y}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

and the left-hand side of this equation is a matrix product: we can rewrite the equation as

$$A^\mathsf{T}(A\mathbf{x}^* - \mathbf{y}) = \mathbf{0}.$$

Now expand to $A^\mathsf{T}A\mathbf{x}^* - A^\mathsf{T}\mathbf{y} = \mathbf{0}$, and move $A^\mathsf{T}\mathbf{y}$ to the other side to get the equation we wanted. □

Does there always exist such a minimizer $\mathbf{x}^*$, and is it unique?

A result from linear algebra tells us that the rank of A is equal to the rank of $A^\mathsf{T}A$; in particular, if A has full column rank, $A^\mathsf{T}A$ is invertible. This is the nice case. If this happens, then the normal equation has a unique solution $\mathbf{x}^*$, and so there is a unique minimizer.

In this nice case, $\mathbf{x}^*$ is given by

$$\mathbf{x}^* = (A^\mathsf{T}A)^{-1}A^\mathsf{T}\mathbf{y}.$$

We write $A^\dagger$ for $(A^\mathsf{T}A)^{-1}A^\mathsf{T}$, and call it the *pseudoinverse* of A, because $A^\dagger\mathbf{y}$ is a "pseudo-solution" of the overconstrainted system $A\mathbf{x} = \mathbf{y}$.

The matrix $P = AA^\dagger$ is called a projection matrix: it maps $\mathbf{y} \in \mathbb{R}^m$ to $P\mathbf{y}$, the closest point in V to $\mathbf{y}$. We will have more to say about projection matrices in the next lecture. For now, a quick exercise in matrix algebra is to check the following two properties of P:

- $P^2 = P$ (P is *idempotent*).
- $P^\mathsf{T} = P$ (P is symmetric).

What about the not-so-nice case where A does not have full column rank—when the vectors $\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(n)}$ are linearly dependent?

If A does not have full column rank, then $\mathbf{x}^*$ is not unique. However, dropping the redundant columns of A does not change the vector space $V = \{A\mathbf{x} : \mathbf{x} \in \mathbb{R}^n\}$, so the closest point $A\mathbf{x}^*$ *is* still unique. It just might have multiple representations with different $\mathbf{x}^*$.