Least squares and the normal equations

Alex Townsend

March 1, 2015

If A is of size 3×2 or 10×3, then $Ax = b$ usually does not have a solution. For example,

$$\begin{pmatrix} 1 & 2 \\ 1 & 3/2 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$$

does not have a unique solution because there is no line $y = c + dx$ that goes through $(2, 1)$, $(3/2, 2)$, and $(4, 1)$.

Instead, for rectangular matrices we seek the least squares solution. That is, we minimize the sum of squares of the error

$$\|b - Ax\|_2^2 = (b - Ax)_1^2 + \ldots + (b - Ax)_n^2.$$

In the above example the least squares solution finds the global minimum of the sum of squares, i.e.,

$$f(c, d) = (1 - c - 2d)^2 + (2 - c - 3/2d)^2 + (1 - c - 4d)^2. \tag{1}$$

At the global minimium the gradient of f vanishes. That is,

$$\begin{pmatrix} \frac{\partial f}{\partial c} \\ \frac{\partial f}{\partial c} \end{pmatrix} = \begin{pmatrix} -2(1 - c - 2d) - 2(2 - c - \frac{3}{2}d) - 2(1 - c - 4d) \\ -4(1 - c - 2d) - 3(2 - c - \frac{3}{2}d) - 8(1 - c - 4d) \end{pmatrix} = \begin{pmatrix} 6c + 15d - 8 \\ 15c + \frac{89}{2}d - 18 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

These equations can be solved by the following linear system (using elimination, say):

$$\begin{pmatrix} 6 & 15 \\ 15 & \frac{89}{2} \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} 8 \\ 18 \end{pmatrix}.$$

MATLAB calculates the global minimum of (1) as $8/21$ when $(c, d) = (43/21, -2/7)$. This is the least squares solution. The line of best-fit is $y = 43/21 - 2/7x$. This is not remarkable.

But this is:

$$2A^T A = 2 \begin{pmatrix} 1 & 1 & 1 \\ 2 & 3/2 & 4 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 1 & 3/2 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} 6 & 15 \\ 15 & \frac{89}{2} \end{pmatrix}, \qquad 2A^T \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 8 \\ 18 \end{pmatrix}.$$

There is no need to differentiate to solve a minimization problem! Just solve the normal equations!

NORMAL EQUATIONS:

$$A^T Ax = A^T b$$

Why the normal equations? To find out you will need to be slightly crazy and totally comfortable with calculus.

In general, we want to minimize[1]

$$f(x) = \|b - Ax\|_2^2 = (b - Ax)^T(b - Ax) = b^T b - x^T A^T b - b^T Ax + x^T A^T Ax.$$

If x is a global minimum of f, then its gradient $\nabla f(x)$ is the zero vector. Let's take the gradient of f remembering that

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}.$$

We have the following three gradients:

$$\nabla(x^T A^T b) = A^T b, \quad \nabla(b^T Ax) = A^T b, \quad \nabla(x^T A^T Ax) = 2A^T Ax.$$

To calculate these gradients, write out $x^T A^T b$, $b^T Ax$, and $x^T A^T Ax$, in terms of sums and differentiate with respect to $x_1, \ldots, x_n$ (this gets very messy).

Thus, we have

$$\nabla f(x) = 2A^T Ax - 2A^T b,$$

just like we saw in the example. We can solve $\nabla f(x) = 0$ or, equivalently $A^T Ax = A^T b$ to find the least squares solution. Magic.

Is this the global minimum? Could it be a maximum, a local minimum, or a saddle point? To find out we take the "second derivative" (known as the Hessian in this context):

$$Hf = 2A^T A.$$

Next week we will see that $A^T A$ is a positive semi-definite matrix and that this implies that the solution to $A^T Ax = A^T b$ is a global minimum of $f(x)$. Roughly speaking, $f(x)$ is a function that looks like a bowl.

[1]Here, x is a vector not a 1D variable.