Least Squares

Stephen Boyd

EE103
Stanford University

October 28, 2017

Outline

Least squares problem

- suppose $m \times n$ matrix A is tall, so $Ax = b$ is over-determined
- for most choices of b, there is no x that satisfies $Ax = b$
- *residual* $r = Ax - b$
- *least squares problem*: choose x to minimize $\|Ax - b\|^2$
- $\|Ax - b\|^2$ is the *objective function*
- $\hat{x}$ is a *solution* of least squares problem if

  $$\|A\hat{x} - b\|^2 \leq \|Ax - b\|^2$$

  for any n-vector x
- idea: $\hat{x}$ makes residual as small as possible, if not 0
- also called *regression* (in data fitting context)

Least squares problem

- $\hat{x}$ called *least squares approximate solution* of $Ax = b$
- $\hat{x}$ is sometimes called 'solution of $Ax = b$ in the least squares sense'
  - this is very confusing
  - never say this
  - do not associate with people who say this

- $\hat{x}$ need not (and usually does not) satisfy $A\hat{x} = b$
- but if $\hat{x}$ does satisfy $A\hat{x} = b$, then it solves least squares problem

Column interpretation

- suppose $a_1, \ldots, a_n$ are columns of A
- then
$$\|Ax - b\|^2 = \|(x_1 a_1 + \cdots + x_n a_n) - b\|^2$$
- so least squares problem is to find a linear combination of columns of A that is closest to b
- if $\hat{x}$ is a solution of least squares problem, the m-vector
$$A\hat{x} = \hat{x}_1 a_1 + \cdots + \hat{x}_n a_n$$
is closest to b among all linear combinations of columns of A

Row interpretation

- suppose $\tilde{a}_1^T, \ldots, \tilde{a}_m^T$ are rows of A
- residual components are $r_i = \tilde{a}_i^T x - b_i$
- least squares objective is

$$\|Ax - b\|^2 = (\tilde{a}_1^T x - b_1)^2 + \cdots + (\tilde{a}_m^T x - b_m)^2$$

  the sum of squares of the residuals
- so least squares minimizes sum of squares of residuals
  - solving $Ax = b$ is making all residuals zero
  - least squares attempts to make them all small

Example

- $A = \begin{bmatrix} 2 & 0 \\ -1 & 1 \\ 0 & 2 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$
- $Ax = b$ has no solution
- $\|Ax - b\|^2 = (2x_1 - 1)^2 + (x_2 - x_1)^2 + (2x_2 + 1)^2$
- least squares approximate solution is $\hat{x} = (1/3, -1/3)$ (say, via calculus)
- $\|A\hat{x} - b\|^2 = 2/3$ is smallest posible value of $\|Ax - b\|^2$
- *e.g.*, with $\tilde{x} = (1/2, -1/2)$, $A\tilde{x} - b = (0, -1, 0)$, and $\|A\tilde{x} - b\|^2 = 1$
- $A\hat{x} = (2/3, -2/3, -2/3)$ is linear combination of columns of A closest to b

Outline

Least squares problem

Solution of least squares problem

Examples

Solution of least squares problem

- we make one assumption: A *has independent columns*
- this implies that Gram matrix $A^T A$ is invertible
- unique solution of least squares problem is

$$\hat{x} = (A^T A)^{-1} A^T b = A^\dagger b$$

- cf. $x = A^{-1} b$, solution of square invertible system $Ax = b$

Derivation via calculus

- define

$$f(x) = \|Ax - b\|^2 = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} A_{ij} x_j - b_i \right)^2$$

- solution $\hat{x}$ satisfies

$$\frac{\partial f}{\partial x_k}(\hat{x}) = \nabla f(\hat{x})_k = 0, \quad k = 1, \ldots, n$$

- taking partial derivatives we get $\nabla f(x)_k = \big(2A^T(Ax - b)\big)_k$
- in matrix-vector notation: $\nabla f(\hat{x}) = 2A^T(A\hat{x} - b) = 0$
- so $\hat{x}$ satisfies *normal equations* $(A^T A)\hat{x} = A^T b$
- and therefore $\hat{x} = (A^T A)^{-1} A^T b$

Direct verification

- let $\hat{x} = (A^T A)^{-1} A^T b$, so $A^T(A\hat{x} - b) = 0$
- for any n-vector x we have

$$\begin{aligned} \|Ax - b\|^2 &= \|(Ax - A\hat{x}) + (A\hat{x} - b)\|^2 \\ &= \|A(x - \hat{x})\|^2 + \|A\hat{x} - b\|^2 + 2(A(x - \hat{x}))^T(A\hat{x} - b) \\ &= \|A(x - \hat{x})\|^2 + \|A\hat{x} - b\|^2 + 2(x - \hat{x})^T A^T(A\hat{x} - b) \\ &= \|A(x - \hat{x})\|^2 + \|A\hat{x} - b\|^2 \end{aligned}$$

- so for any x, $\|Ax - b\|^2 \geq \|A\hat{x} - b\|^2$
- if equality holds, $A(x - \hat{x}) = 0$, which implies $x = \hat{x}$ since columns of A are independent

Computing least squares approximate solutions

- compute QR factorization of A: $A = QR$ $\quad (2mn^2$ flops)
  (exists since columns of A are independent)
- to compute $\hat{x} = A^\dagger b = R^{-1}Q^Tb$
  - form Q^Tb $\quad (2mn$ flops)
  - compute $\hat{x} = R^{-1}(Q^Tb)$ via back substitution $\quad (n^2$ flops)
- total complexity $2mn^2$ flops

- identical to algorithm for solving $Ax = b$ for square invertible A
- but when A is tall, gives least squares approximate solution

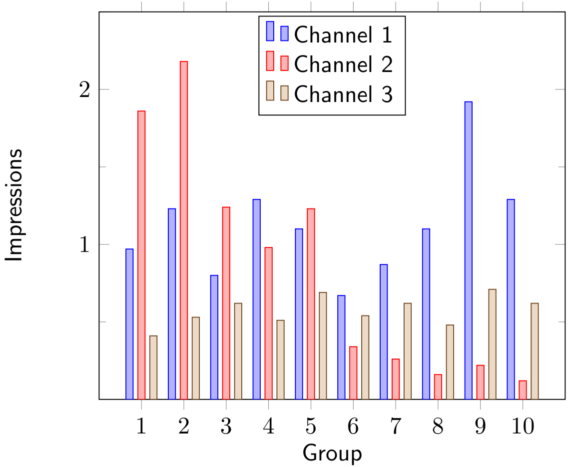
Outline

Advertising purchases

- m demographics groups we want to advertise to
- v^{des} is m-vector of target views or impressions
- n-vector s gives spending on n channels
- $m \times n$ matrix R gives demographic reach of channels
- R_{ij} is number of views per dollar spent (in 1000/$)
- $v = Rs$ is m-vector of views across demographic groups
- $\|v^{\mathrm{des}} - Rs\|/\sqrt{m}$ is RMS deviation from desired views
- we'll use least squares spending $\hat{s} = R^{\dagger} v^{\mathrm{des}}$ (need not be ≥ 0)

Example

$m = 10,\ n = 3$



Least squares advertising purchases

with $v^{\text{des}} = 10^3 \times \mathbf{1}$, $\hat{s} = (62, 100, 1443)$