Sep 30, 2020

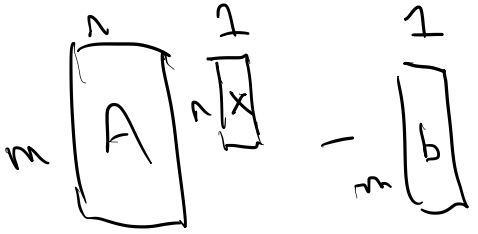
New topic: least squares, QR factorization

Model: $b_i \approx a_i^T x + z \qquad i = 1, \ldots, m$

b_i: outcome; a_i: features; z: intercept

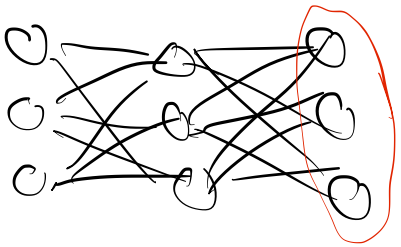
Error: $\sum_{i=1}^{m} (a_i^T x + z - b_i)^2 \Rightarrow \left\| \begin{pmatrix} a_1^T & 1 \\ \vdots & \vdots \\ a_m^T & 1 \end{pmatrix} \begin{pmatrix} x \\ z \end{pmatrix} - b \right\|_2^2$

Linear least squares: $\min_x \|Ax - b\|_2^2$



Examples:

(1) stats/ML a_i = data about page i, b_i = # clicks



final layer $= a_i$

#IJALM

(2) polynomial fitting $a_i^T = (1 \; z_i \; z_i^2 \; z_i^3)$ $b_i = f(z_i)$

$(z_i, f(z_i))$

$$\min_x \|Ax-b\|_2^2$$

Why squared error?

$\|\cdot\|_1$ $\|\cdot\|_{\infty}$ ...

1. not unreasonable
2. computationally nice
3. statistical interpretations
4. building block

Gauss-Markov theorem

$b_i = a_i^T x^* + \varepsilon_i$

- $E(\varepsilon_i) = 0$
- $\mathrm{Var}(\varepsilon_i) < \infty$
- $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$

$\hat{x} = \arg\min_x \|Ax-b\|_2^2$

(1) $E(\hat{x}) = x^*$ (unbiased)

(2) least variance amongst all linear estimators

(BLUE)

$b_i = a_i^T x + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ i.i.d.

$b_i - a_i^T x = \varepsilon_i$ $\quad f(z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} z^2/\sigma^2\right)$

Likelihood$(b; A, x)$

$$= \prod_{i=1}^m f(b_i - a_i^T x) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^m \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^m (b_i - a_i^T x)^2\right)\right)$$

max likelihood $\Rightarrow$ $\min \|Ax-b\|_2^2$

What is the answer?

$$\hat{x} = \arg\min_x \|Ax - b\|_2^2$$

$$(Ax-b)^T(Ax-b)$$

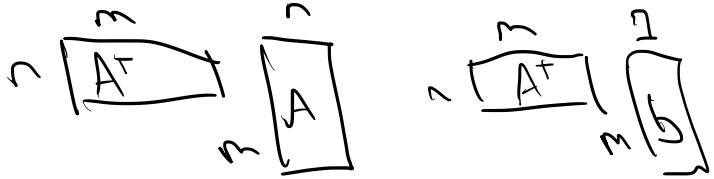
$$= x^T A^T A x - 2x^T A^T b + b^T b$$

Take gradient w/r/t x...

$$2A^TA\hat{x} - 2A^Tb = 0$$

$$\boxed{A^TA\hat{x} = A^Tb}$$

Normal equations



Nonsingular?

$$\hat{x} = (A^TA)^{-1}A^Tb$$ ★

Moore-Penrose pseudoinverse, A^+

Numbers $b_1, \ldots, b_m$. Summary stat x

$$\min_x \sum_{i=1}^m (b_i - x)^2 = \|\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} x - b\|_2^2$$

$A = e$

$$\hat{x} = (e^Te)^{-1}(e^Tb) = \frac{1}{m}\sum b_i = \text{mean}$$

residual $r = A\hat{x} - b$ $\|r\|_2^2$

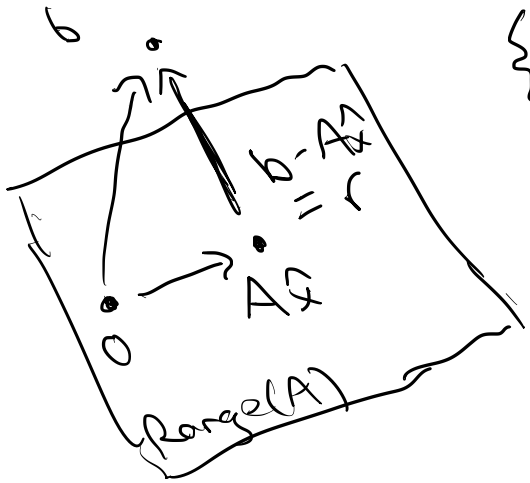
$$r = \frac{1}{m}\sum_{i=1}^m (\hat{x} - b_i)^2 = \text{variance}$$

$$A^TA\hat{x} = A^Tb \iff A^T(\underbrace{A\hat{x} - b}_{r}) = 0$$

$$\min r \iff b - A\hat{x} \in \text{Null}(A^T)$$

$$\iff r \perp \text{Range}(A)$$

$$\text{Range}(A) = \{Ax\}$$



How do we compute $\hat{x}$?

Idea 1: just a linear system!

$(A^TA)\hat{x} = A^Tb$ $\quad A^TA$ SPD $\Rightarrow A^TA = LL^T$ (Cholesky)

Flops: $O(mn^2) + O(mn) + O(n^3) \Rightarrow O(mn^2)$

$\min_x \|r\|_2^2$ $\quad r = b - Ax$ $\quad \|r\|_2^2 \geq 0$ $\quad \hat{x} = A^{-1}b$

$O(n^3)$

Conditioning?

$A = U\Sigma V^T$ $\quad A^TA = V\Sigma^2V^T$ $\quad \kappa_2(A^TA) = \frac{\sigma_1^2}{\sigma_n^2} = \kappa_2(A)^2$

Idea 2: SVD

$A = U \Sigma V^T \qquad \hat{x} = A^+ b = (A^T A)^{-1} A^T b$

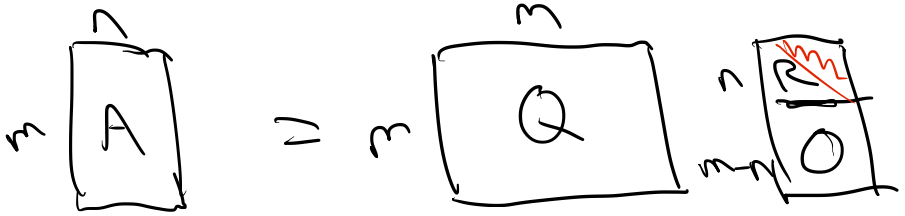
$A^T A = V \Sigma^2 V^T$

$(A^T A)^{-1} A^T b = V \Sigma^{-2} V^T V \Sigma U^T b$

$= V \Sigma^{-1} U^T b$

Can be used but there are faster algorithms (very useful for analysis!)

Idea 3: QR factorization



"full QR"

$Q^TQ = QQ^T = I$

R upper tri;

$$\|Ax-b\|_2^2 = \|Q^T(Ax-b)\|_2^2$$

$$= \left\| \begin{pmatrix} R \\ 0 \end{pmatrix} x - Q^Tb \right\|_2^2$$

~~$Q^TQ \begin{bmatrix} R \\ 0 \end{bmatrix}$~~

$Rx = Q^Tb(1{:}n, 1{:}n)$

"thin/economy/reduced QR"

$$A = \begin{pmatrix} \overset{n}{Q_1} & \overset{m-n}{Q_2} \end{pmatrix} \begin{pmatrix} R \\ 0 \end{pmatrix} = Q_1 R$$

$Q^TQ = I \quad (QQ^T \neq I)$

$$\hat{x} = (A^TA)^{-1}A^Tb = (R^T\cancel{Q^TQ}R)^{-1}R^TQ^Tb$$

$$= R^{-1}\cancel{R^{-T}R^T}Q^Tb$$

$R\hat{x} = Q^Tb$

① Thin QR $O(mn^2)$

② $c = Q^Tb$

③ $R\hat{x} = c$

How to compute QR?

From what we know...

$$A^TA = R^TQ^TQR = R^TR = LL^T \quad L = R^T$$

(1) Cholesky $\Rightarrow L^T = R$

(2) $A = QR \Rightarrow Q = AR^{-1}$

"Cholesky QR"

Graph $\Longleftrightarrow$ A

$d = Ae \qquad d_i = \sum_j A_{ij}$

$v = \operatorname{argmin}_i d_i$

$A(:, v)$ nonzeros $\Rightarrow$ neighbors N

$d[N] -= 1$

$\operatorname{argmin}_{i \neq v} d_i$