Linear algebra for AI and ML

(September-24)

Chapter - 15, 16, 17

[Applied Linear algebra
vectors, matrice & Least squares
S. Boyd]

Multi-objective LS.

$$\min_{x \in \mathbb{R}^n} \boxed{\|Ax - b\|_2^2}$$: LS problem

Objective function

$b \in \mathbb{R}^m$; $A \in \mathbb{R}^{m \times n}$

real values — labels

Data fitting — classification

A is full column rank matrix

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

A^+ : pseudo-inverse.

$$\begin{bmatrix} A_i \end{bmatrix} \begin{bmatrix} x \end{bmatrix} = \begin{bmatrix} b_i \end{bmatrix}$$

$$\min_{x \in \mathbb{R}^n} \|A_i x - b_i\|_2^2$$

multiple objective functions:

$$\underbrace{J_1 = \|A_1 x - b_1\|_2^2}, \; ; \; \underbrace{J_2 = \|A_2 x - b_2\|_2^2}, \cdots,$$

$$J_k = \|A_k x - b_k\|_2^2$$

$A_i \in \mathbb{R}^{m_i \times n}$ and $b_i \in \mathbb{R}^{m_i}$ for $i = 1, 2, \ldots, k$

We seek a common solution $\hat{x} \in \mathbb{R}^n$

Multi-objective LS via weighted-sum

$$J = \lambda_1 J_1 + \lambda_2 J_2 + \cdots + \lambda_k J_k$$

$$\lambda_1, \lambda_2, \ldots, \lambda_k > 0$$

$$J = \lambda_1 \|A_1 x - b\|_2^2 + \lambda_2 \underbrace{\|A_2 x - b_2\|_2^2} + \cdots + \lambda_k \|A_k x - b_k\|_2^2$$

$$J = \lambda_1 \|A_1 x - b_1\|_2^2 + \lambda_2 \|A_2 x - b_2\|_2^2 + \cdots + \lambda_k \|A_k x - b_k\|_2^2$$

$$\min_{x \in \mathbb{R}^n} J$$

$$J = \left\| \begin{bmatrix} \sqrt{\lambda_1}\,(A_1 x - b_1) \\ \sqrt{\lambda_2}\,(A_2 x - b_2) \\ \vdots \\ \sqrt{\lambda_k}\,(A_k x - b_k) \end{bmatrix} \right\|_2^2$$

$\mathbb{R}^{m \times 1}$

where $m = m_1 + m_2 + \cdots + m_k$

$$J = \left\| \underbrace{\begin{bmatrix} \sqrt{\lambda_1} A_1 \\ \sqrt{\lambda_2} A_2 \\ \vdots \\ \sqrt{\lambda_k} A_k \end{bmatrix}} x - \underbrace{\begin{bmatrix} \sqrt{\lambda_1} b_1 \\ \sqrt{\lambda_2} b_2 \\ \vdots \\ \sqrt{\lambda_k} b_k \end{bmatrix}} \right\|_2^2 = \| \underbrace{\tilde{A}} x - \underbrace{\tilde{b}} \|_2^2$$

$\tilde{A} \in \mathbb{R}^{m \times n}$; $\tilde{b} \in \mathbb{R}^{m \times 1}$

If the columns of $\tilde{A}$ are linearly independent, then LS solution is given by:

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

$$\tilde{A} = \begin{bmatrix} \sqrt{\lambda_1} A_1 \\ \sqrt{\lambda_2} A_2 \\ \vdots \\ \sqrt{\lambda_k} A_k \end{bmatrix} \quad ; \quad \tilde{A}^T \tilde{A} = (\lambda_1 A_1^T A_1 + \lambda_2 A_2^T A_2 + \cdots + \lambda_k A_k^T A_k)$$

$$\boxed{\hat{x} = (\lambda_1 A_1^T A_1 + \lambda_2 A_2^T A_2 + \cdots + \lambda_k A_k^T A_k)^{-1} (\lambda_1 A_1^T b_1 + \cdots + \lambda_k A_k^T b_k)}$$

$$G_i = A_i^T A_i$$

case of $k = 2$: (bicriterion / biobjective)

J_1, J_2

$$J = \lambda_1 J_1 + \lambda_2 J_2 = J_1 + \lambda J_2$$

primary objective

secondary objective

Given A & y.

Applications of multi-objective LS

→ Estimation / inversion.

$$\boxed{y} = \boxed{A} x + \boxed{\varepsilon}$$

← measurement error in y

no-constraints on A

$$\begin{bmatrix} & A & \end{bmatrix} \begin{bmatrix} x \end{bmatrix} = \begin{bmatrix} y \end{bmatrix}$$

$y = f(x) + \varepsilon$

$\varepsilon = 0$, A has full column rank

$\underbrace{\hat{x} = A^{+} y}$: exact inversion:

[ε is random measurement noise ("small")
No restrictions/ conditions on A.]

$\rightarrow$ $\boxed{\text{Approximate inversion}}$

(SVD)

$J_1 = \underbrace{\|Ax - y\|_2^2}$

Regularized inversion:

(Prior knowledge "x" has smaller

- $\underbrace{\|x\|_2^2}$: x is "small".
- $\underbrace{\|x - x^{prior}\|_2^2}$: x is close to x^{prior}. $\leftarrow$ (FEMUP)

- $\underbrace{\|\tilde{D}x\|_2^2}$ where D is the differce matrix.

$$\begin{bmatrix} \underbrace{x_{i+1} - x_i} \\ \\ \\ \end{bmatrix} \equiv x \text{ is "smooth".}$$

time-series modeling

Estimate $\hat{x}$ by minimizing

$$\underbrace{\|Ax - b\|_2^2} + \lambda_2 J_2(x) + \cdots + \lambda_k J_k(x)$$

regularized inversion/ regularized estimation.

Tikhonov regularized inversion

$$\hat{x} = \arg\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 + \lambda \|x\|_2^2$$

$$\tilde{A} = \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} x$$

$\Rightarrow \tilde{A}$ has linearly independent columns without any condition on A.

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

$$\boxed{\hat{x} = (A^T A + \lambda I)^{-1} A^T b}$$

Image deblurring:

x : image

A: blurring matrix

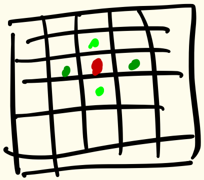
$$y = Ax + \varepsilon$$

observed

blurring image

noise

$$\min_{x \in \mathbb{R}^n} \underset{J_1}{\|Ax - y\|_2^2} + \lambda \underbrace{\left(\|D_h x\|_2^2 + \|D_v x\|_2^2 \right)}_{J_2}$$



Regularized data fitting: (avoids overfitting)

$$\hat{f}(x) = \alpha_1 f_1(x) + \cdots + \alpha_p f_p(x)$$

We can interprete α_i as the amount by which the prediction depends on $f_i(x)$.

If α_i is large, then the prediction will be sensitive to "small" changes in $f_i(x)$.

test data

$$\|y - A\alpha\|_2^2 + \lambda \|\alpha\|_2^2$$

$$\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix}$$

Regression Notation:

$$\min_{\beta} \ \| y - x^T \beta_1 - \beta_0 \|_2^2 + \lambda \| \beta \|_2^2$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

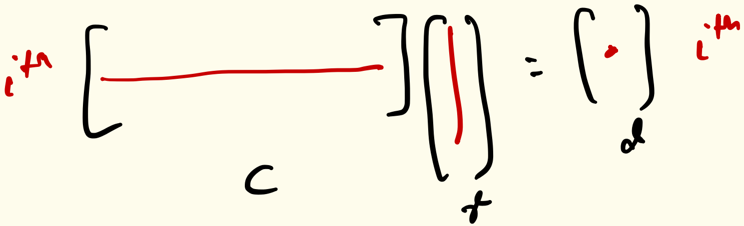
Constrained Least squares:

$$\min_{x \in \mathbb{R}^n} \|Ax-b\|_2^2 \quad \leftarrow \text{objective fn}$$

$$\text{s.t.} \quad Cx = d \quad \leftarrow \text{constraint}$$

$\Big\}$ (LL)

Linearly constrained Least squares problem.



$$c_i^T x = d_i$$

<u>feasible solution</u> : $x \in \mathbb{R}^n$ which satisfies constraints.

optimal solution $\hat{x} \in \mathbb{R}^n$

$\rightarrow \hat{x}$ is feasible

$\rightarrow \|A\hat{x} - b\|_2^2 \leq \|Ax - b\|_2^2 \quad \forall x \in$ feasible set.

---

consider the bi-objective problem.

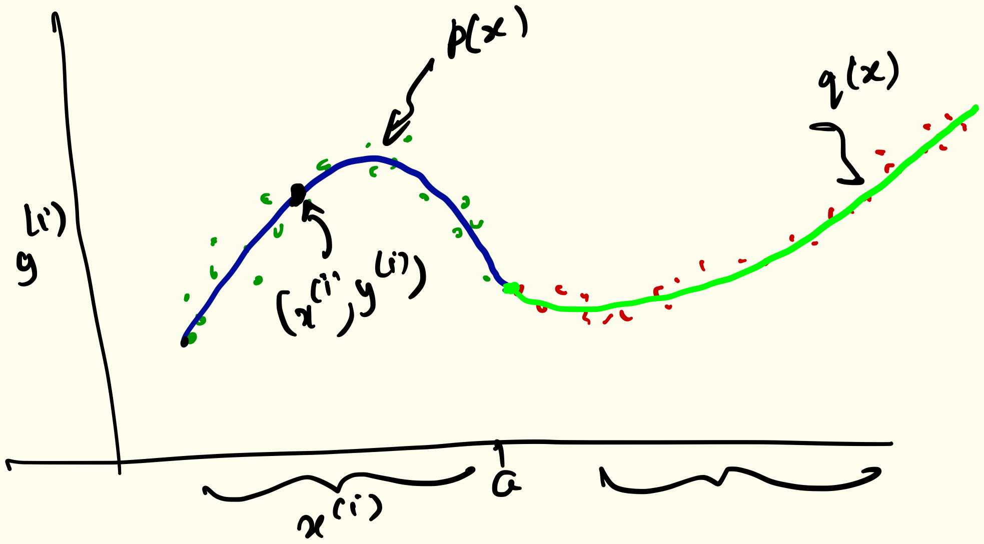
$$\|Ax - b\|_2^2 + \lambda \|Cx - d\|_2^2 \quad \text{—} (*)$$

let $\lambda \to \infty$ the solution $\hat{x}_{MOP}$ of $(*)$

converges / tends to the $\hat{x}_{CLS}$ of $(**)$

Practically let λ to be very large number.

Example: Piecewise polynomial fitting problem.



$$\hat{f}(x) = \begin{cases} p(x) & x \le a \\ q(x) & x > a \end{cases}$$

$p(x)$ & $q(x)$ are polynomials of degree less than or equal to 3.

$$p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3$$
$$q(x) = \alpha_4 + \alpha_5 x + \alpha_6 x^2 + \alpha_7 x^3$$

} compute "best" choice of $\alpha_0, \dots, \alpha_7$

Data: $x^{(1)}, \dots, x^{(N)} \in \mathbb{R}$

$y^{(1)}, \dots, y^{(N)} \in \mathbb{R}$

Let $x^{(1)}, x^{(2)}, \dots, x^{(M)} \leq a$

$x^{(M+1)}, \dots, x^{(N)} > a$

Prediction error:

$$\sum_{i=1}^{M} \left(\alpha_0 + \alpha_1 x^{(i)} + \alpha_2 [x^{(i)}]^2 + \alpha_3 [x^{(i)}]^3 - y^{(i)}\right)^2$$

$$+ \sum_{i=M+1}^{N} \left(\alpha_4 + \alpha_5 x^{(i)} + \alpha_6 [x^{(i)}]^2 + \alpha_7 [x^{(i)}]^3 - y^{(i)}\right)^2$$

Another condition:

continuity at 'a' : $p(a) = q(a) \Rightarrow p(a) - q(a) = 0$

smoothness at 'a' : $p'(a) = q'(a) \Rightarrow p'(a) - q'(a) = 0$

Continuity at 'a' :

$$\rightarrow \alpha_0 + \alpha_1 a + \alpha_2 a^2 + \alpha_3 a^3 - \alpha_4 - \alpha_5 a - \alpha_6 a^2 - \alpha_7 a^3 = 0 \checkmark$$

smoothness at 'a' :

$$\rightarrow \alpha_1 + 2\alpha_2 a + 3\alpha_3 a^2 - \alpha_5 - 2\alpha_6 a - 3\alpha_7 a^2 = 0 \checkmark$$

Define:

$$c = \begin{bmatrix} 1 & a & a^2 & a^3 & -1 & -a & -a^2 & -a^3 \\ 0 & 1 & 2a & 3a^2 & 0 & -1 & -2a & -3a^2 \end{bmatrix} \in \mathbb{R}^{2\times 8}$$

$$d = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \mathbb{R}^{2\times 1}$$

Define:

$$A = \begin{bmatrix} 1 & x^{(1)} & (x^{(1)})^2 & (x^{(1)})^3 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & & \\ 1 & x^{(M)} & (x^{(M)})^2 & (x^{(M)})^3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & x^{(M+1)} & (x^{M+1})^2 & (x^{(M+1)})^3 \\ \vdots & & & & & & & \\ 0 & 0 & 0 & 0 & 1 & x^{(N)} & (x^{(N)})^2 & (x^{(N)})^3 \end{bmatrix}$$

$$\in \mathbb{R}^{N \times 8}$$

$$b = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(M)} \\ y^{(M+1)} \\ \vdots \\ y^{(N)} \end{bmatrix} \in \mathbb{R}^N$$

$$\min \quad \|A\alpha - b\|_2^2$$

$$\text{s.t.} \quad C\alpha = d$$

where $\alpha = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \\ \alpha_7 \end{bmatrix} \begin{array}{l} \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} \leftarrow p(x) \\ \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} \leftarrow q(x) \end{array}$

A particular case:

i) Least norm problem:

$$\min \quad \|x\|_2^2$$

$$\text{s.t.} \quad \boxed{Cx = d}$$

Variation problem: (NOT LS) (orthogonal basis pursuit)

$$\min \|x\|_1 \quad \text{s.t.} \quad Cx = d$$

dictionary learning / sparse representation.

Solution to the linearly constrained least squares problem:

$$\min \|Ax - b\|_2^2$$
$$\text{s.t.} \quad c_i^T x = d_i \qquad i = 1, 2, \cdots, p$$

where c_i^T : i^{th} row of $C \in \mathbb{R}^{p \times n}$

Lagrangian function:

$$L(x, z) = \|Ax - b\|_2^2 + z_1 (c_1^T x - d_1) + z_2 (c_2^T x - d_2)$$

$$+ \cdots + z_p (c_p^T x - d_p)$$

where $z = \begin{pmatrix} z_1 \\ \vdots \\ z_p \end{pmatrix}$: vector of Lagrange multipliers.

Let $\hat{x}$ be the solution to the constrained LS problem. Let $\hat{z}$ be the set of Lagrange multipliers that satisfy:

$\checkmark$ $\dfrac{\partial L}{\partial x_i}(\hat{x}, \hat{z}) = 0$ for $i = 1, 2, \ldots, n$

$\checkmark$ $\dfrac{\partial L}{\partial z_i}(\hat{x}, \hat{z}) = 0$ for $i = 1, 2, \ldots, p$

$$\frac{\partial L}{\partial z_i} = c_i^T \hat{x} - d_i = 0 \qquad \text{for } i = 1, 2, \dots, p$$

$$\frac{\partial L}{\partial x_i}(\hat{x}, \hat{z}) = 2 \sum_{j=1}^{n} (A^T A)_{ij} \hat{x}_j - 2 (A^T b)_i$$

$$+ \; 2 \sum_{j=1}^{p} \hat{z}_j (c_j)_i = 0$$

In matrix form :

$$\underbrace{2 (A^T A) \hat{x} - 2 A^T b} + \underbrace{C^T \hat{z}} = 0$$

$$\underbrace{\begin{bmatrix} 2A^TA & C^T \\ C & 0 \end{bmatrix}}_{\text{KKT matrix}} \begin{bmatrix} \hat{x} \\ \hat{z} \end{bmatrix} = \begin{bmatrix} 2A^Tb \\ d \end{bmatrix}$$

These eq^n: are called as KKT eq^{ns}:

(Invertibility of KKT matrix) $\Longleftrightarrow$ ($\begin{bmatrix} A \\ C \end{bmatrix}$ has linearly indep. columns) ⊛

$$\begin{bmatrix} \hat{x} \\ \hat{z} \end{bmatrix} = \begin{bmatrix} 2A^TA & C^T \\ C & 0 \end{bmatrix}^{-1} \begin{bmatrix} 2A^Tb \\ d \end{bmatrix}$$

■