Linear Algebra for AI & ML

(September. 17)

LS problem:

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 \quad (2)$$

Assumption: Columns of A are linearly independent.

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

$$\hat{x} = A^+ b$$

$\hat{x}$ which minimizer should satisfy

$$A^T A \hat{x} = A^T b \quad : \text{normal eqns}$$

Column interpretation of the LS problem:

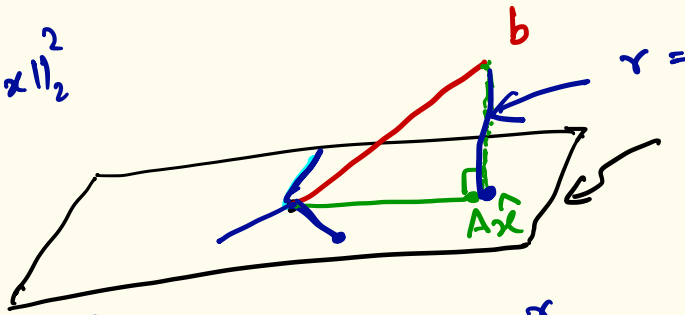
$$Ax = x_1\begin{bmatrix} \\ A_1 \\ \\ \end{bmatrix} + x_2\begin{bmatrix} \\ A_2 \\ \\ \end{bmatrix} + \cdots + x_n\begin{bmatrix} \\ A_n \\ \\ \end{bmatrix}$$

A_i : i^{th} column of A.

Orthogonality principle in the least squares

$$\min_{x \in \mathbb{R}^n} \|Ax-b\|_2^2 \quad (2) \quad = \quad \min_{x \in \mathbb{R}^n} \|x_1A_1 + \cdots + x_nA_n - b\|_2^2$$

$\langle x, x \rangle = \|x\|_2^2$



col space(A)
Range(A)

$\mathcal{L} \oplus \mathcal{L}^{\perp} = \mathbb{R}^m$

$\rightarrow A^T(b - A\hat{x}) = 0 \Rightarrow$ $A^Tb = A^TA\hat{x}$ (with $r = b - A\hat{x}$)

$b \notin \text{colspan}(A)$

$\text{Range}(A)$

Find $\hat{x}$ s.t. $A\hat{x}$ is (closest) to b

$$\|Ax - b\|_2$$

$$\min_{x \in \mathbb{R}^n} \|Ax-b\|_\infty = \min_{x \in \mathbb{R}^n} \{ \max |r_i| \} \leftarrow \text{LASSO}$$

$r_i = (Ax-b)_i$

$$\|Ax-b\|_\infty = \max_i |(Ax-b)_i|$$

$$\min_{x \in \mathbb{R}^n} \|Ax-b\|_1 = \min_{x \in \mathbb{R}^n} \{ |r_1| + \cdots + |r_m| \}$$

$$\min_{x \in \mathbb{R}^n} \|Ax-b\|_2^2 = \min_{x \in \mathbb{R}^n} \{ (r_1)^2 + \cdots + r_m^2 \} \leftarrow \text{LS}$$

How to compute $\hat{x}$??

$$(A^TA)\hat{x} = A^Tb \quad \ldots \text{ normal eqn}$$

$$\hat{x} = (A^TA)^{-1}A^Tb$$

$\hat{x}$ will <u>NOT</u> be computed like this!!

$$[A]_{m\times n} = QR$$

$$\begin{bmatrix} \\ \\ \\ \end{bmatrix}_{m\times n} \begin{bmatrix} \\ \\ \end{bmatrix}_{n\times n}$$

$$(A^TA)^{-1}A^T = R^{-1}Q^T$$

$$\hat{x} = R^{-1}Q^Tb$$

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

back substitution.

Given a matrix A, we should avoid computation of $A^T A$ as well.

(problem case: matrix A has almost linearly dependent columns)

Least squares data fitting / classification

binary multiclass

$f: \mathbb{R}^n \longrightarrow \mathbb{R}$

$y = f(x)$

$x \in \mathbb{R}^n$; $y \in \mathbb{R}$

response variable

independent variables

Data: $\left\{ x^{(1)}, x^{(2)}, \ldots, x^{(N)} \right.$
$\left. y^{(1)}, y^{(2)}, \ldots, y^{(N)} \right\} \in \mathbb{R}^{(n+1)}$

$x^{(i)}, y^{(i)}$: Data pair

$x_j^{(i)}$: j^{th} value of i^{th} observation

Model: Form a model to describe the relationship between x & y

$$y \approx \hat{f}(x)$$

where $\hat{f} : \mathbb{R}^n \longrightarrow \mathbb{R}$ is the **model**.

Linear in parameters model:

$$\hat{f}(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \cdots + \alpha_p f_p(x)$$

where $f_i : \mathbb{R}^n \longrightarrow \mathbb{R}$ are basis functions (feature mappings)

For a fixed $x \in \mathbb{R}^n$; $\hat{f}$ is a linear f^{m} in $(\alpha_1, \ldots \alpha_p)^T$; $\hat{f}(x) = (\alpha_1 \cdots, \alpha_p)^T \begin{pmatrix} f_1(x) \\ \vdots \\ f_p(x) \end{pmatrix}$

Prediction error:

Our objective is to choose the model $\hat{f}$ so that it is consistent with the data i.e. $y^{(i)} \approx \hat{f}(x^{(i)})$ for $i = 1, 2, \dots, N$

Denote $\hat{y}^{(i)} = \hat{f}(x^{(i)})$

Residual: $r^{(i)} = y^{(i)} - \hat{y}^{(i)}$

In vector notation: $y^d = (y^{(1)}, \dots, y^{(N)})^T$

$\hat{y}^d = (\hat{y}^{(1)}, \dots, \hat{y}^{(N)})^T$

$$r^d = (y^{(1)} - \hat{y}^{(1)}, \dots, y^{(N)} - \hat{y}^{(N)})^T$$

Least square model fitting:

minimize $\|r^d\|_2^2$
$\alpha_1, \alpha_2, \cdots, \alpha_p$

Express $\hat{y}^{(i)} = A_{i1}\theta_1 + \cdots + A_{ip}\theta_p$ $\quad i = 1, 2, \cdots, N$

where $A_{ij} = f_j(x^{(i)})$ $\quad$ for $i = 1, 2, \cdots, N$
$j = 1, 2, \cdots, p$

$A = [A_{ij}] \in \mathbb{R}^{N \times p}$

Matrix-vector notation

$$\hat{y}^d = A \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix}$$

where $\alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix}$

$$\|r^d\|_2^2 = \|y^d - \hat{y}^d\|_2^2 = \|y^d - A\alpha\|_2^2$$

$$= \|A\alpha - y^d\|_2^2$$

$$\min_{\alpha_1,\ldots,\alpha_p} \|r^d\|_2^2 = \min_{\alpha_1,\ldots,\alpha_p} \|A\alpha - y^d\|_2^2$$

$$\boxed{\alpha = (A^T A)^{-1} A^T y^d = A^+ y^d}$$

If A has linearly dependent columns, then one of the column (let us assume p^{th} column) can be expressed as a linear combination of the remaining columns.

$\frac{\|y^d - A\alpha\|^2}{N}$: Mean square error.

For $\hat{\alpha}$ mean square error is minimum.

$\sqrt{\frac{\|y^d - A\alpha\|^2}{N}}$: Root mean square (RMS).

In general : $\varphi(r^d)$ may be used to determine how "closely" $\hat{f}$ matches the data $x^{(i)}, y^{(i)}$ $i = 1, 2, \ldots, N$.

For w: $\varphi(r^d) = RMS(r^d) = \sqrt{\frac{\|y^d - A\alpha\|_2^2}{N}}$ → LS estimate α.

Least squares fit on a constant:

(Interpretation of mean & variance)

$p = 1$ and $f_1(x) = 1 \quad \forall\, x \in \mathbb{R}^n$

In this case, the model $\hat{f}$ is constant.

$$\hat{f} = \alpha_1 f_1 = \alpha_1$$

$$A \in \mathbb{R}^{N \times 1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{N \times 1} = 1_N$$

$$y^d : \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{bmatrix}$$

$$\hat{\alpha}_1 = (A^T A)^{-1} A^T y^d$$

$$= \frac{1}{N} 1_N^T y^d$$

$$\hat{\alpha}_1 = \text{avg}(y^d) \quad ; \; \text{RMS}(r^d) = \text{std}(y^d)$$

Univariate fit.

$n = 1$

$x \in \mathbb{R}^n$; $y \in \mathbb{R}$

$x \in \mathbb{R}$; $y \in \mathbb{R}$

Straight line fit:

$f_1(x) = 1$, $f_2(x) = x$

$$\begin{array}{cc} x^{(1)} & y^{(1)} \\ \vdots & \vdots \\ x^{(N)} & y^{(N)} \end{array}$$

$\hat{f}(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x)$

$\hat{f}(x) = \alpha_1 + \alpha_2 x$

$$A = \begin{bmatrix} 1 & x^{(1)} \\ 1 & x^{(2)} \\ \vdots & \vdots \\ 1 & x^{(N)} \end{bmatrix} \in \mathbb{R}^{N \times 2}, \quad y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{bmatrix}$$

Polynomial fit:

$n = 1$; $f_1(x) = 1$, $f_2(x) = x$; $f_3(x) = x^2$, ... $f_p(x) = x^{p-1}$; $f_{p+1}(x) = x^p$

Data: $(x^{(1)}, y^{(1)}), \dots, (x^{(N)}, y^{(N)})$

$$A = \begin{bmatrix} 1 & x^{(1)} & (x^{(1)})^2 & \cdots & (x^{(1)})^p \\ 1 & x^{(2)} & (x^{(2)})^2 & \cdots & (x^{(2)})^p \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x^{(N)} & (x^{(N)})^2 & \cdots & (x^{(N)})^p \end{bmatrix} \in \mathbb{R}^{N \times (p+1)}$$

Vandermonde matrix

(ill-conditioned matrices) = $\frac{\text{maxmag}(A)}{\text{minmag}(A)}$

log transform of dependent variable.

y is positive and varies over a large range.

$$w = \log y$$

use least squares to develop a model for w.

$$\hat{w} = \hat{g}(x)$$

$$\hat{y} = e^{\hat{w}} = e^{\hat{g}(x)}$$

Generalization ability: