A Quick Math Recap for Regression

Mohammad Sadegh Talebi
m.shahi@di.ku.dk
Department of Computer Science

Vectors

- We use small **boldface** letters to denote vectors.
- A column vector $\boldsymbol{x} \in \mathbb{R}^d$,

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

- Alternatively, $\boldsymbol{x} = \left[x_1, \dots, x_d\right]^\top$, with $\top$ denoting 'Transpose'.
- Inner product of vectors: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^d x_i y_i$
- Euclidean (or L_2) norm for vectors: $\|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}} = \sqrt{\sum_{i=1}^d x_i^2}$

Vectors

- We use small **boldface** letters to denote vectors.
- A column vector $\boldsymbol{x} \in \mathbb{R}^d$,

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

- Alternatively, $\boldsymbol{x} = \begin{bmatrix} x_1, \dots, x_d \end{bmatrix}^\top$, with $\top$ denoting 'Transpose'.
- Inner product of vectors: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^d x_i y_i$
- Euclidean (or L_2) norm for vectors: $\|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}} = \sqrt{\sum_{i=1}^d x_i^2}$
- $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are linearly independent *if and only if*

$$a_1 \boldsymbol{x}_1 + \ldots + a_n \boldsymbol{x}_n = \mathbf{0} \quad \Longleftrightarrow \quad a_1 = \ldots = a_n = 0$$

I.e., none of them is a linear combination of the rest.

Matrices

- We use capital **boldface** letters to denote vectors.
- A matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$,

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1d} \\ A_{21} & A_{22} & \dots & A_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nd} \end{bmatrix}$$

Matrices

- We use capital **boldface** letters to denote vectors.
- A matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$,

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1d} \\ A_{21} & A_{22} & \dots & A_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nd} \end{bmatrix}$$

- Outer product of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$\boldsymbol{x}\boldsymbol{y}^\top = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \begin{bmatrix} y_1, \dots, y_d \end{bmatrix} = \begin{bmatrix} x_1y_1 & x_1y_2 & \dots & x_1y_d \\ x_2y_1 & x_2y_2 & \dots & x_2y_d \\ \vdots & \vdots & \ddots & \vdots \\ x_dy_1 & x_dy_2 & \dots & x_dy_d \end{bmatrix}$$

While $\boldsymbol{x}^\top\boldsymbol{y}$ is a scalar, $\boldsymbol{x}\boldsymbol{y}^\top$ is a matrix.

Matrices

- We use capital **boldface** letters to denote vectors.
- A matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$,

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1d} \\ A_{21} & A_{22} & \dots & A_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nd} \end{bmatrix}$$

- Outer product of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$\boldsymbol{x}\boldsymbol{y}^\top = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \begin{bmatrix} y_1, \dots, y_d \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \dots & x_1 y_d \\ x_2 y_1 & x_2 y_2 & \dots & x_2 y_d \\ \vdots & \vdots & \ddots & \vdots \\ x_d y_1 & x_d y_2 & \dots & x_d y_d \end{bmatrix}$$

  While $\boldsymbol{x}^\top \boldsymbol{y}$ is a scalar, $\boldsymbol{x}\boldsymbol{y}^\top$ is a matrix.
- Column space (or range) of $\mathbf{A}$ is the collection of all *linear combination* of columns of $\mathbf{A}$.

Gradient

Consider a real-valued multivariate function $f : \mathbb{R}^d \to \mathbb{R}$. The gradient of f evaluated at $\boldsymbol{x}$, if exists:

$$\nabla f(\boldsymbol{x}) = \left[\frac{\partial f}{\partial x_1}(\boldsymbol{x}), \ldots, \frac{\partial f}{\partial x_d}(\boldsymbol{x})\right]^\top$$

Gradient

Consider a real-valued multivariate function $f : \mathbb{R}^d \to \mathbb{R}$. The gradient of f evaluated at $\boldsymbol{x}$, if exists:

$$\nabla f(\boldsymbol{x}) = \left[\frac{\partial f}{\partial x_1}(\boldsymbol{x}), \dots, \frac{\partial f}{\partial x_d}(\boldsymbol{x})\right]^\top$$

- Note that $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$.
- $\nabla f(\boldsymbol{x})$ indicates the direction and rate of fastest increase in f at $\boldsymbol{x}$.

Gradient

Consider a real-valued multivariate function $f : \mathbb{R}^d \to \mathbb{R}$. The gradient of f evaluated at $\boldsymbol{x}$, if exists:

$$\nabla f(\boldsymbol{x}) = \Big[\frac{\partial f}{\partial x_1}(\boldsymbol{x}), \dots, \frac{\partial f}{\partial x_d}(\boldsymbol{x})\Big]^\top$$

- Note that $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$.
- $\nabla f(\boldsymbol{x})$ indicates the direction and rate of fastest increase in f at $\boldsymbol{x}$.
- Example: affine function $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$, with $\nabla f(\boldsymbol{x}) = \boldsymbol{w}$
- Example: quadratic function $f(\boldsymbol{x}) = \boldsymbol{x}^\top \mathbf{A} \boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x} + c$, with

$$\nabla f(\boldsymbol{x}) = (\mathbf{A} + \mathbf{A}^\top)\boldsymbol{x} + \boldsymbol{b}$$

Linear System

Consider a system of linear equations

$$\mathbf{A}\boldsymbol{x} = \boldsymbol{b}, \quad \mathbf{A} \in \mathbb{R}^{n \times d}.$$

A solution $\boldsymbol{x}$ exists if and only if $\boldsymbol{b}$ is in the column space of $\mathbf{A}$.

Linear System

Consider a system of linear equations

$$\mathbf{A}\boldsymbol{x} = \boldsymbol{b}, \quad \mathbf{A} \in \mathbb{R}^{n \times d}.$$

A solution $\boldsymbol{x}$ exists if and only if $\boldsymbol{b}$ is in the column space of $\mathbf{A}$.

Linear System

Consider a system of linear equations

$$\mathbf{A}\boldsymbol{x} = \boldsymbol{b}, \quad \mathbf{A} \in \mathbb{R}^{n \times d}.$$

A solution $\boldsymbol{x}$ exists if and only if $\boldsymbol{b}$ is in the column space of $\mathbf{A}$.

Three possibilities 'in general':

- Single solution
- No solution (overdetermined system)
- Infinitely many solutions (underdetermined system)

Linear System

Consider a system of linear equations

$$\mathbf{A}\boldsymbol{x} = \boldsymbol{b}, \quad \mathbf{A} \in \mathbb{R}^{n \times d}.$$

Assume (i) $n > d$ and (ii) $\boldsymbol{b}$ is not in range of $\mathbf{A}$.

Linear System

Consider a system of linear equations

$$\mathbf{A}\boldsymbol{x} = \boldsymbol{b}, \quad \mathbf{A} \in \mathbb{R}^{n\times d}.$$

Assume (i) $n > d$ and (ii) $\boldsymbol{b}$ is not in range of $\mathbf{A}$.

- An overdetermined system, hence no solution exists ...
- ... yet we are interested in $\boldsymbol{x}$ s.t. $\mathbf{A}\boldsymbol{x} \approx \mathbf{b}$.
- More concretely, we wish to find $\boldsymbol{x}$ s.t. $\|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2$ is small:

$$\operatorname*{argmin}_{\boldsymbol{x}} \|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2$$

Overdetermined Linear System

$$\boldsymbol{x}^\star = \operatorname*{argmin}_{\boldsymbol{x}} \|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2 = \operatorname*{argmin}_{\boldsymbol{x}} \|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2^2$$

Overdetermined Linear System

$$\boldsymbol{x}^\star = \operatorname*{argmin}_{\boldsymbol{x}} \|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2 = \operatorname*{argmin}_{\boldsymbol{x}} \|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2^2$$

An unconstrained optimization problem: To find $\boldsymbol{x}^\star$, we solve $\nabla\|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2^2 = \mathbf{0}$

$$\|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2^2 = (\mathbf{A}\boldsymbol{x} - \boldsymbol{b})^\top(\mathbf{A}\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{x}^\top\mathbf{A}^\top\mathbf{A}\boldsymbol{x} - 2\boldsymbol{b}^\top\mathbf{A}\boldsymbol{x} + \boldsymbol{b}^\top\boldsymbol{b}$$
$$\implies \nabla\|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2^2 = 2\mathbf{A}^\top\mathbf{A}\boldsymbol{x} - 2\mathbf{A}^\top\boldsymbol{b}$$

Overdetermined Linear System

$$\boldsymbol{x}^\star = \operatorname*{argmin}_{\boldsymbol{x}} \|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2 = \operatorname*{argmin}_{\boldsymbol{x}} \|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2^2$$

An unconstrained optimization problem: To find $\boldsymbol{x}^\star$, we solve $\nabla\|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2^2 = \mathbf{0}$

$$\|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2^2 = (\mathbf{A}\boldsymbol{x} - \boldsymbol{b})^\top(\mathbf{A}\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{x}^\top\mathbf{A}^\top\mathbf{A}\boldsymbol{x} - 2\boldsymbol{b}^\top\mathbf{A}\boldsymbol{x} + \boldsymbol{b}^\top\boldsymbol{b}$$
$$\Longrightarrow \nabla\|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2^2 = 2\mathbf{A}^\top\mathbf{A}\boldsymbol{x} - 2\mathbf{A}^\top\boldsymbol{b}$$

Hence, $\boldsymbol{x}^\star$ satisfies: $\mathbf{A}^\top\mathbf{A}\boldsymbol{x}^\star = \mathbf{A}^\top\boldsymbol{b}$

- If columns of $\mathbf{A}$ are linearly independent, $\mathbf{A}^\top\mathbf{A}$ is invertible so that

$$\boldsymbol{x}^\star = \left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top\boldsymbol{b} := \mathbf{A}^\dagger\boldsymbol{b}$$

- $\mathbf{A}^\dagger = \left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top$ is called the Moore-Penrose inverse of $\mathbf{A}$.
- In practice, $\mathbf{A}^\dagger$ is found via QR-decomposition to reduce computation.

Overdetermined Linear System

$$\boldsymbol{x}^\star = \underset{\boldsymbol{x}}{\operatorname{argmin}} \|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2 = \underset{\boldsymbol{x}}{\operatorname{argmin}} \|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2^2$$

An unconstrained optimization problem: To find $\boldsymbol{x}^\star$, we solve $\nabla\|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2^2 = \mathbf{0}$

$$\|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2^2 = (\mathbf{A}\boldsymbol{x} - \boldsymbol{b})^\top(\mathbf{A}\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{x}^\top\mathbf{A}^\top\mathbf{A}\boldsymbol{x} - 2\boldsymbol{b}^\top\mathbf{A}\boldsymbol{x} + \boldsymbol{b}^\top\boldsymbol{b}$$
$$\implies \nabla\|\mathbf{A}\boldsymbol{x} - \mathbf{b}\|_2^2 = 2\mathbf{A}^\top\mathbf{A}\boldsymbol{x} - 2\mathbf{A}^\top\boldsymbol{b}$$

Hence, $\boldsymbol{x}^\star$ satisfies: $\mathbf{A}^\top\mathbf{A}\boldsymbol{x}^\star = \mathbf{A}^\top\boldsymbol{b}$

- If columns of $\mathbf{A}$ are linearly independent, $\mathbf{A}^\top\mathbf{A}$ is invertible so that

$$\boldsymbol{x}^\star = \left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top\boldsymbol{b} := \mathbf{A}^\dagger\boldsymbol{b}$$

- $\mathbf{A}^\dagger = \left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top$ is called the Moore-Penrose inverse of $\mathbf{A}$.
- In practice, $\mathbf{A}^\dagger$ is found via QR-decomposition to reduce computation.