Multivariate Statistical Analysis

Lecture 01

Fudan University

luoluo@fudan.edu.cn

Outline

1. Course Overview

2. Linear Algebra

3. Convex Optimization

Outline

1. Course Overview
2. Linear Algebra
3. Convex Optimization

Course Overview

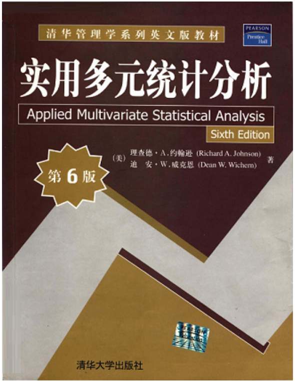
Homepage:

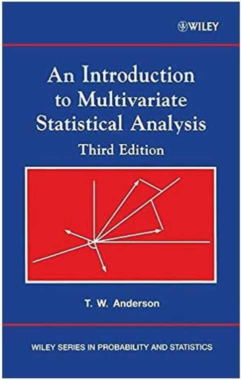
- https://luoluo-sds.github.io/

Prerequisite courses:

- Calculus
- Linear algebra
- Probability and statistics
- Optimization
- Machine learning

Textbook (recommended reading):





Grading Policy

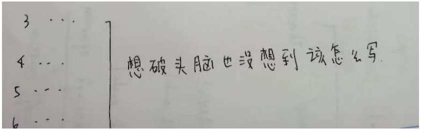
Option I:

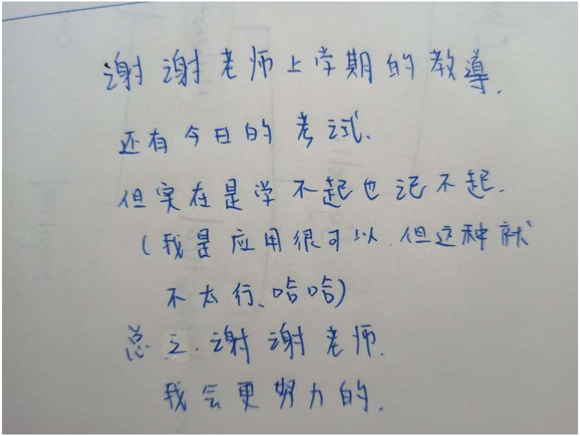
- Homework, 40%
- Final Exam, 60%

Option II:

- Quiz, 20%
- Homework, 30%
- Final Exam, 50%

Grading Policy





What is Multivariate Statistics?

2021–2022 NBA season

1. Points leaders:

Rank	Player	PTS
1	Joel Embiid	30.6
2	LeBron James	30.3
3	Giannis Antetokounmpo	29.9
4	Kevin Durant	29.9
5	Luka Dončić	28.4
6	Trae Young	28.4
7	DeMar DeRozan	27.9
8	Kyrie Irving	27.4
9	Ja Morant	27.4
10	Nikola Jokić	27.1
11	Jayson Tatum	26.9
12	Devin Booker	26.8
13	Donovan Mitchell	25.9
14	Stephen Curry	25.5
15	Karl-Anthony Towns	24.6

Rank	Player	PTS
16	Shai Gilgeous-Alexander	24.5
17	Zach LaVine	24.4
18	CJ McCollum	24.3
19	Paul George	24.3
20	Damian Lillard	24.0
21	Jaylen Brown	23.6
22	De'Aaron Fox	23.2
23	Bradley Beal	23.2
24	Anthony Davis	23.2
25	Pascal Siakam	22.8
26	Brandon Ingram	22.7
27	James Harden	22.5
28	CJ McCollum	22.1
29	Kristaps Porziņģis	22.1
30	James Harden	22.0

2. MVP ranking:

Rank	Player	PTS	TRB	AST	STL	BLK	WIN%
1	Nikola Jokić	27.1	13.8	7.9	1.5	0.9	0.585
2	Joel Embiid	30.6	11.7	4.2	1.1	1.5	0.622
3	Giannis Antetokounmpo	29.9	11.6	5.8	1.1	1.4	0.622

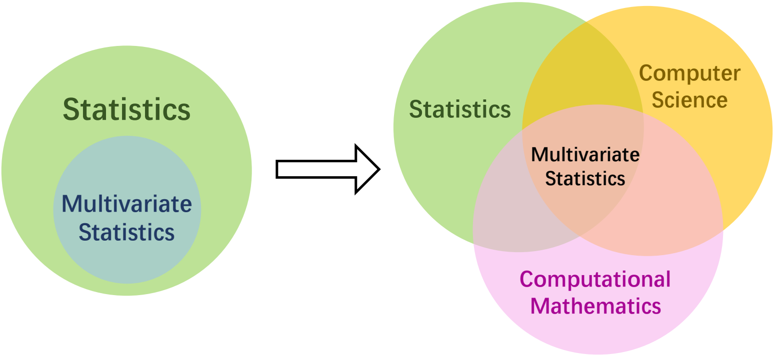
Applications of Multivariate Statistics

1. Investigating of the dependency among variables
2. Hypotheses testing
3. Dimensionality reduction
4. Prediction
5. Clustering

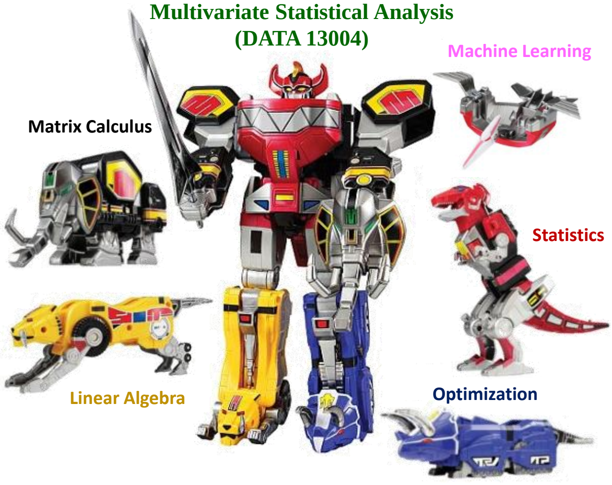
Applications of Multivariate Statistics

课程	学生1	学生2	学生3	学生4	学生5	学生6
习近平新时代中国特色社会主义思想概论	B+	A-	B	A-	C	A
马克思主义原理	A	A	B	B+	B	B+
形势与政策	A-	A-	A	A-	B+	B+
数学分析	A	A	C+	A-	B-	B+
高等代数	A-	A	C	B+	C+	A-
最优化方法	A	A-	C	A-	C+	A-
多元统计分析	A	?	D	?	?	A-
程序设计	B+	A	A	A-	B+	B-
数据库及实现	B+	?	A	B+	B	?
神经网络与深度学习	B+	A-	A-	A-	?	B
计算机视觉	B+	A	A	?	B-	B-
自然语言处理	B+	?	A	A-	B+	B+

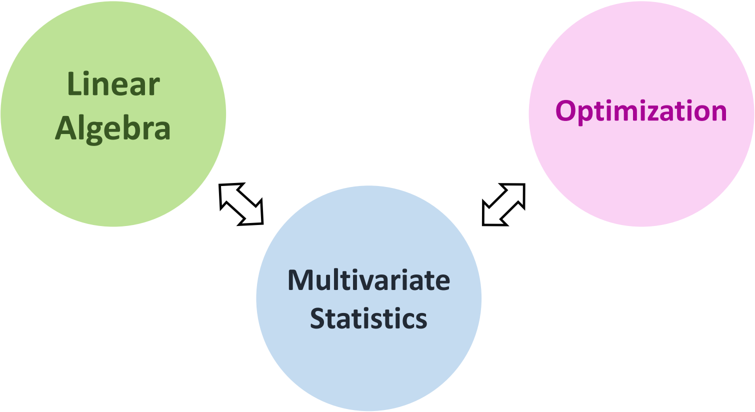
Where is Multivariate Statistics?



Where is Multivariate Statistics?



Where is Multivariate Statistics?



We start from the review of linear algebra and convex optimization.

Outline

1. Course Overview

2. Linear Algebra

3. Convex Optimization

Notations

We use x_i to denote the entry of the n-dimensional vector $\mathbf{x}$ such that

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n.$$

We use a_{ij} or $(\mathbf{A})_{ij}$ to denote the entry of matrix $\mathbf{A}$ with dimension $m \times n$ such that

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

Notations

We can also present the matrix as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1q} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{p1} & \mathbf{A}_{p2} & \cdots & \mathbf{A}_{pq} \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

if the sub-matrices are compatible with the partition.

We define

$$\mathbf{0} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{m\times n} \quad \text{and} \quad \mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

Transpose

The transpose of a matrix results from flipping the rows and columns. Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ such that

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n},$$

then its transpose, written $\mathbf{A}^\top \in \mathbb{R}^{n \times m}$, is an $n \times m$ matrix such that

$$\mathbf{A}^\top = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{n \times m}.$$

Sometimes, we also use $\mathbf{A}'$ the present the transpose of $\mathbf{A}$.

If $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{m\times n}$ are two matrices of the same order, then

$$\mathbf{A}+\mathbf{B} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2n}+b_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \cdots & a_{mn}+b_{mn} \end{bmatrix} \in \mathbb{R}^{m\times n}$$

and

$$\mathbf{A}-\mathbf{B} = \begin{bmatrix} a_{11}-b_{11} & a_{12}-b_{12} & \cdots & a_{1n}-b_{1n} \\ a_{21}-b_{21} & a_{22}-b_{22} & \cdots & a_{2n}-b_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}-b_{m1} & a_{m2}-b_{m2} & \cdots & a_{mn}-b_{mn} \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

Multiplication

The product of $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$ is the matrix

$$\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m\times p},$$

where

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1q} \\ c_{21} & c_{22} & \cdots & c_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pq} \end{bmatrix} \in \mathbb{R}^{m\times p}.$$

and $c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$.

Trace

The trace of a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, denoted $\mathrm{tr}(\mathbf{A})$, is the sum of diagonal elements in the matrix:

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

The trace has the following properties

1. For $\mathbf{A} \in \mathbb{R}^{n\times n}$, we have $\mathrm{tr}(\mathbf{A}) = \mathrm{tr}\left(\mathbf{A}^\top\right)$.
2. For $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times n}$, $c_1 \in \mathbb{R}$ and $c_2 \in \mathbb{R}$, we have
$$\mathrm{tr}(c_1\mathbf{A} + c_2\mathbf{B}) = c_1\mathrm{tr}(\mathbf{A}) + c_2\mathrm{tr}(\mathbf{B}).$$
3. For $\mathbf{A}$ and $\mathbf{B}$ such that $\mathbf{AB}$ is square, $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$.
4. For $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ such that $\mathbf{ABC}$ is square, we have
$$\mathrm{tr}(\mathbf{ABC}) = \mathrm{tr}(\mathbf{BCA}) = \mathrm{tr}(\mathbf{CAB}).$$

Inverse

The inverse of a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is denoted by $\mathbf{A}^{-1}$ and is the unique matrix such that

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I} = \mathbf{A}^{-1}\mathbf{A}.$$

We say that $\mathbf{A}$ is invertible or non-singular if $\mathbf{A}^{-1}$ exists and non-invertible or singular otherwise.

Inverse

If all the necessary inverse exist, we have

1. $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
2. $(c\mathbf{A})^{-1} = c^{-1}\mathbf{A}^{-1}$
3. $(\mathbf{A}^{-1})^\top = (\mathbf{A}^\top)^{-1}$
4. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
5. $\mathbf{A}^{-1} = \mathbf{A}^\top$ if $\mathbf{A}^\top\mathbf{A} = \mathbf{I}$

For $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$, $\mathbf{C} \in \mathbb{R}^{p\times p}$ and $\mathbf{D} \in \mathbb{R}^{p\times n}$, we have

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}$$

if $\mathbf{A}$ and $\mathbf{A} + \mathbf{BCD}$ are non-singular.

A norm of a vector $\mathbf{x} \in \mathbb{R}^n$ written by $\|\mathbf{x}\|$, is informally a measure of the length of the vector.

Formally, a norm is any function $\mathbb{R}^n \to \mathbb{R}$ that satisfies four properties:

1. For all $\mathbf{x} \in \mathbb{R}^n$, we have $\|\mathbf{x}\| \geq 0$ (non-negativity).
2. $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
3. For all $\mathbf{x} \in \mathbb{R}^n$ and $t \in \mathbb{R}$, we have $\|t\mathbf{x}\| = |t| \, \|\mathbf{x}\|$.
4. For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

Vector Norms

There are some examples for $\mathbf{x} \in \mathbb{R}^n$:

1. The ℓ_2 norm is $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$
2. The ℓ_1 norm is $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$
3. The ℓ_p norm is $\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ for $p > 1$.
4. The ℓ_∞ norm is $\|\mathbf{x}\|_\infty = \max_i |x_i|$

Orthogonality

1. Two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are orthogonal if $\mathbf{x}^\top \mathbf{y} = 0$.
2. A vector $\mathbf{x} \in \mathbb{R}^n$ is normalized if $\|\mathbf{x}\|_2 = 1$.
3. A square matrix $\mathbf{U} \in \mathbb{R}^{n \times n}$ is orthogonal if all its columns are orthogonal to each other and are normalized (the columns are then referred to as being orthonormal). In other word, we have

$$\mathbf{U}^\top \mathbf{U} = \mathbf{I} = \mathbf{U}\mathbf{U}^\top.$$

4. Note that if $\mathbf{U}$ is not square, i.e., $\mathbf{U} \in \mathbb{R}^{m \times n}$, $n < m$, but its columns are still orthonormal, then $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$, but $\mathbf{U}\mathbf{U}^\top \neq \mathbf{I}$, we call that $\mathbf{U}$ is column orthonormal.

Quiz

What is the volume of the tetrahedral?

Determinant

Given square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ as

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_{(1)}^\top \\ \mathbf{a}_{(2)}^\top \\ \vdots \\ \mathbf{a}_{(n)}^\top \end{bmatrix},$$

the determinant of $\mathbf{A}$ is the "volume" of the set

$$\mathcal{S} = \left\{ \mathbf{v} \in \mathbb{R}^n : \mathbf{v} = \sum_{i=1}^{n} \beta_i \mathbf{a}_{(i)}, \text{where } 0 \le \beta_i \le 1, i = 1, \dots, n \right\}.$$

The set $\mathcal{S}$ formed by taking all possible linear combinations of the row vectors, where the coefficients are all between 0 and 1.

Determinant

The determinant of a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, is denoted by $\det(\mathbf{A})$ or $|\mathbf{A}|$, which is defined as

$$\det(\mathbf{A}) = \sum_{\tau=(\tau_1,\dots,\tau_n)} \left(\operatorname{sgn}(\tau) \prod_{i=1}^{n} \mathbf{a}_{i,\tau_i} \right)$$

where $\tau = (\tau_1, \dots, \tau_n)$ is permutation of $(1, 2, \dots, n)$. The signature $\operatorname{sgn}(\tau)$ is defined to be $+1$ whenever the reordering given by τ can be achieved by successively interchanging two entries an even number of times, and -1 whenever it can be achieved by an odd number of such interchanges.

Determinant

We can also define determinant recursively

$$\det(\mathbf{A}) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det(\mathbf{A}_{\backslash i, \backslash j}) \quad \text{for any } j \in \{1, \ldots, n\}$$

with the initial condition $\det(a_{ij}) = a_{ij}$, where $\mathbf{A}_{\backslash i, \backslash j}$ is the $(n-1) \times (n-1)$ matrix obtained by deleting the i-th row and j-th column from $\mathbf{A}$.

Determinant

1. $\det(\mathbf{I}) = 1$
2. If we multiply a single row in $\mathbf{A}$ by a scalar $t \in \mathbb{R}^n$, then the determinant of the new matrix is $t \det(\mathbf{A})$.
3. If we exchange any two rows of the square matrix $\mathbf{A}$, then the determinant of the new matrix is $-\det(\mathbf{A})$.
4. For $\mathbf{A} \in \mathbb{R}^{n \times n}$, we have $\det(\mathbf{A}) = 0$ if and only if $\mathbf{A}$ is singular.

Determinant

1. For $\mathbf{A} \in \mathbb{R}^{n\times n}$ is triangular, then $\det(\mathbf{A}) = \prod_{i=1}^{n} a_{ii}$.
2. For $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{p\times p}$ and $\mathbf{C} \in \mathbb{R}^{n\times p}$, we have
$$\det\left(\begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}\right) = \det(\mathbf{A})\det(\mathbf{B})$$
3. For $\mathbf{A} \in \mathbb{R}^{n\times n}$, we have $\det(\mathbf{A}) = \det(\mathbf{A}^\top)$.
4. For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$, we have $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$.
5. For $\mathbf{A} \in \mathbb{R}^{n\times n}$ is orthogonal, we have $\det(\mathbf{A}) = 1$.

Singular Value Decomposition

The singular value decomposition (SVD) of $\mathbf{A} \in \mathbb{R}^{m \times n}$ matrix is

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top,$$

where $\mathbf{U} \in \mathbb{R}^{m \times m}$ is orthogonal, $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ is rectangular diagonal matrix with non-negative real numbers on the diagonal and $\mathbf{V} \in \mathbb{R}^{n \times n}$ is orthogonal.

1. The diagonal entries of $\mathbf{\Sigma}$ are uniquely determined by $\mathbf{A}$ and are known as the singular values of $\mathbf{A}$.
2. The number of non-zero singular values is equal to the rank of $\mathbf{A}$.
3. The columns of $\mathbf{U}$ and the columns of $\mathbf{V}$ are called left-singular vectors and right-singular vectors of $\mathbf{A}$, respectively.

Singular Value Decomposition

The term SVD sometimes refers to the compact SVD, that is

$$\mathbf{A} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^\top$$

in which $\mathbf{\Sigma}_r$ is square diagonal of size $r \times r$, where $r \leq \min\{m, n\}$ is the rank of $\mathbf{A}$, and has only the non-zero singular values.

In this variant, $\mathbf{U}_r$ is an $m \times r$ column orthogonal matrix and $\mathbf{V}_r$ is an $n \times r$ column orthogonal matrix such that

$$\mathbf{U}_r^\top \mathbf{U}_r = \mathbf{V}_r^\top \mathbf{V}_r = \mathbf{I}.$$

Matrix Norms

Matrix norm is any function $\mathbb{R}^{m\times n} \to \mathbb{R}$ that satisfies

1. For all $\mathbf{A} \in \mathbb{R}^{m\times n}$, we have $\|\mathbf{A}\| \geq 0$.
2. $\|\mathbf{A}\| = 0$ if and only if $\mathbf{A} = \mathbf{0}$.
3. For all $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $t \in \mathbb{R}$, we have $\|t\mathbf{A}\| = |t|\,\|\mathbf{A}\|$.
4. For all $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$, we have $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$.

Matrix Norms

Given any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, its spectral norm is defined as

$$\|\mathbf{A}\|_2 = \sup_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sup_{\mathbf{x}\in\mathbb{R}^n, \|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2\,;$$

and its Frobenius norm is defined as

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2} = \sqrt{\mathrm{tr}(\mathbf{A}^\top\mathbf{A})}.$$

We can show that

$$\|\mathbf{A}\|_2 = \sigma_1 \quad \text{and} \quad \|\mathbf{A}\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2},$$

where $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_r \geq 0$ are the non-zero singular values of $\mathbf{A}$.

Let $\mathbf{A} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^\top$ be condense SVD of rank-r matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and partition

$$\mathbf{U}_r = [\mathbf{u}_1, \dots, \mathbf{u}_r] \in \mathbb{R}^{m \times r}, \ \mathbf{\Sigma}_r = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \in \mathbb{R}^{r \times r}, \ \mathbf{V}_r = [\mathbf{v}_1, \dots, \mathbf{v}_r] \in \mathbb{R}^{n \times r}.$$

The matrix $\mathbf{A}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^\top$ is the best rank-k approximation of $\mathbf{A}$ ($k \leq r$), where

$$\mathbf{U}_k = [\mathbf{u}_1, \dots, \mathbf{u}_k] \in \mathbb{R}^{m \times k}, \ \mathbf{\Sigma}_k = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix} \in \mathbb{R}^{k \times k}, \ \mathbf{V}_k = [\mathbf{v}_1, \dots, \mathbf{v}_k] \in \mathbb{R}^{n \times k}.$$

We have

$$\mathbf{A}_k = \operatorname*{arg\,min}_{\mathrm{rank}(\mathbf{X}) \leq k} \|\mathbf{A} - \mathbf{X}\|_2 = \operatorname*{arg\,min}_{\mathrm{rank}(\mathbf{X}) \leq k} \|\mathbf{A} - \mathbf{X}\|_F \,.$$

Quadratic Forms

Given a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and a vector $\mathbf{x} \in \mathbb{R}^n$, the scalar $\mathbf{x}^\top \mathbf{A}\mathbf{x}$ is called a quadratic form and we have

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j.$$

We often implicitly assume that the matrices appearing in a quadratic form are symmetric.

Definiteness

We introduce the definiteness as follows.

1. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is positive definite if for all non-zero vectors $\mathbf{x} \in \mathbb{R}^n$ holds that $\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$. This is usually denoted by $\mathbf{A} \succ \mathbf{0}$.
2. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is positive semi-definite if for all vectors $\mathbf{x} \in \mathbb{R}^n$ holds that $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$. This is usually denoted by $\mathbf{A} \succeq \mathbf{0}$.

Similarly, we can define negative definite and negative semi-definite matrices.

Schur Complement

Given matrices $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\mathbf{B} \in \mathbb{R}^{p\times q}$, $\mathbf{C} \in \mathbb{R}^{q\times p}$ and $\mathbf{D} \in \mathbb{R}^{q\times q}$, we suppose $\mathbf{D}$ is non-singular and let

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \in \mathbb{R}^{(p+q)\times(p+q)}.$$

Then the Schur complement of the block $\mathbf{D}$ for $\mathbf{M}$ is

$$\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} \in \mathbb{R}^{p\times p}.$$

Then we can decompose the matrix $\mathbf{M}$ as

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{B}\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix}$$

and the inverse of $\mathbf{M}$ can be written as

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix} \begin{bmatrix} (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{B}\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

Schur Complement

The decomposition

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{B}\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix}$$

means we have $\det(\mathbf{M}) = \det(\mathbf{D})\det(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})$.

We consider the symmetric matrix

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{D} \end{bmatrix}$$

with non-singular $\mathbf{D}$ and let $\mathbf{S} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}^\top$, then

1. $\mathbf{M} \succ \mathbf{0} \iff \mathbf{D} \succ \mathbf{0}$ and $\mathbf{S} \succ \mathbf{0}$.
2. If $\mathbf{D} \succ \mathbf{0}$, then $\mathbf{M} \succeq \mathbf{0} \iff \mathbf{S} \succeq \mathbf{0}$.

Low-Rank Approximation and Beyond

For symmetric positive-definite $\mathbf{A} \in \mathbb{R}^{n\times n}$, its best rank-$k$ approximation is

$$\mathbf{A}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{U}_k^\top = \operatorname*{arg\,min}_{\mathrm{rank}(\mathbf{X})\leq k} \|\mathbf{A}-\mathbf{X}\|_2 = \operatorname*{arg\,min}_{\mathrm{rank}(\mathbf{X})\leq k} \|\mathbf{A}-\mathbf{X}\|_F .$$

Inspired by *probabilistic PCA*, we find the better estimator

$$\widehat{\mathbf{A}}_k = \mathbf{U}_k(\mathbf{\Sigma}_k - \hat{\delta}\mathbf{I}_k)\mathbf{U}_k^\top + \hat{\delta}\mathbf{I}_d, \quad \text{where} \quad \hat{\delta} = \frac{1}{n-k}\sum_{i=k+1}^{n} \sigma_i.$$

We can verify

$$\left(\mathbf{U}_k(\mathbf{\Sigma}_k - \hat{\delta}\mathbf{I}_k)^{1/2}, \hat{\delta}\right) = \operatorname*{arg\,min}_{\mathrm{rank}(\mathbf{B})\leq k, \delta\in\mathbb{R}} \left\|\mathbf{A} - (\mathbf{B}\mathbf{B}^\top + \delta\mathbf{I}_d)\right\|_F$$

and

$$\left\|\mathbf{A} - \widehat{\mathbf{A}}_k\right\|_F \leq \|\mathbf{A}-\mathbf{A}_k\|_F .$$

The Gradient

Suppose that $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a differentiable function that takes as input a matrix $\mathbf{X}$ of size $m \times n$ and returns a real value. Then the gradient of f with respect to $\mathbf{X}$ is

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \nabla f(\mathbf{X}) = \begin{bmatrix} \frac{\partial f(\mathbf{X})}{\partial x_{11}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{X})}{\partial x_{m1}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{mn}} \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

Some Basic Results

1. For $\mathbf{X} \in \mathbb{R}^{m\times n}$, we have $\dfrac{\partial(f(\mathbf{X}) + g(\mathbf{X}))}{\partial \mathbf{X}} = \dfrac{\partial f(\mathbf{X})}{\partial \mathbf{X}} + \dfrac{\partial g(\mathbf{X})}{\partial \mathbf{X}}$.
2. For $\mathbf{X} \in \mathbb{R}^{m\times n}$ and $t \in \mathbb{R}$, we have $\dfrac{\partial t f(\mathbf{X})}{\partial \mathbf{X}} = t\dfrac{\partial f(\mathbf{X})}{\partial \mathbf{X}}$.
3. For $\mathbf{A}, \mathbf{X} \in \mathbb{R}^{m\times n}$, we have $\dfrac{\partial \mathrm{tr}(\mathbf{A}^\top \mathbf{X})}{\partial \mathbf{X}} = \mathbf{A}$.
4. For $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{x} \in \mathbb{R}^{n}$, we have $\dfrac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$.

   If $\mathbf{A}$ is symmetric, we have $\dfrac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$.

We can find more results in the matrix cookbook:
https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Hessian

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a twice differentiable function. Then its Hessian with respect to $\mathbf{x}$, written as $\nabla^2 f(\mathbf{x})$, which is defined as

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_1} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_n} \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Taylor's expansion:

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \nabla f(\mathbf{a})^\top (\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^\top \nabla^2 f(\mathbf{a})(\mathbf{x} - \mathbf{a}).$$

Outline

1 Course Overview

2 Linear Algebra

3 Convex Optimization

Convex Function

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if it holds

$$f(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\alpha \in [0, 1]$.

Theorem (first-order condition)

If a function $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, then it is convex if and only if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

holds for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

If a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable, then $\mathbf{x}^*$ is the global minimizer of $f(\cdot)$ if and only if $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

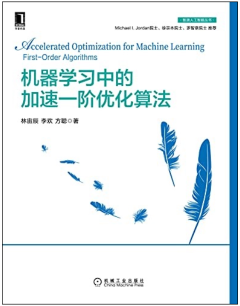
Convex Function

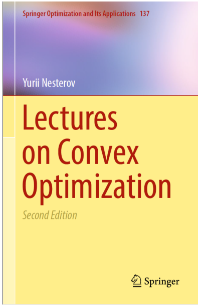
Theorem (second-order condition)

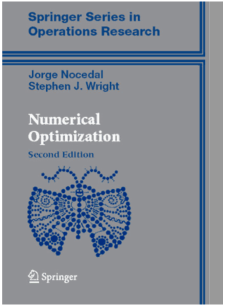
If a function $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable, then it is convex if and only if

$$\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$$

holds for any $\mathbf{x} \in \mathbb{R}^d$.







Example: Least Squares

Consider the least square problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{A}\mathbf{x} - \mathbf{b}\right\|_2^2.$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ is full rank, $\mathbf{b} \in \mathbb{R}^m$ and $m \geq n$.

The solution is

$$\mathbf{x}^* = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{b}.$$

Pseudo Inverse

Let $\mathbf{A} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^\top \in \mathbb{R}^{m \times n}$ be the condense SVD, where r is the rank of $\mathbf{A}$. We define the pseudo inverse of $\mathbf{A}$ as

$$\mathbf{A}^\dagger = \mathbf{V}_r \mathbf{\Sigma}_r^{-1} \mathbf{U}_r^\top \in \mathbb{R}^{n \times m}.$$

In special case, we have

1. If $\mathrm{rank}(\mathbf{A}) = n$, we have $\mathbf{A}^\dagger = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top$.
2. If $\mathrm{rank}(\mathbf{A}) = m$, we have $\mathbf{A}^\dagger = \mathbf{A}^\top (\mathbf{A}\mathbf{A}^\top)^{-1}$.
3. If $\mathbf{A}$ is square and non-singular, we have $\mathbf{A}^\dagger = \mathbf{A}^{-1}$.

The solution of the general least square problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) = \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$$

is $\{\mathbf{x} : \mathbf{x} = \mathbf{A}^\dagger \mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger \mathbf{A})\mathbf{b},\ \mathbf{y} \in \mathbb{R}^n\}$.

Gradient Descent Method

We consider the optimization problem

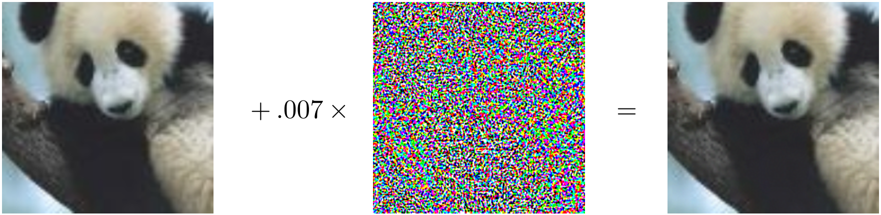
$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}),$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable.

The most popular method is gradient descent, which follows

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t),$$

where $\eta_t > 0$.



"panda"
57.7% confidence

noise

"gibbon"
99.3 % confidence

We can only access the output of a big model.

Zeroth-Order Optimization

We consider the optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}),$$

where the gradient of $f:\mathbb{R}^d\to\mathbb{R}$ is difficult to access.

We can solve the problem by iteration

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \cdot \frac{f(\mathbf{x}_t + \delta\mathbf{u}_t) - f(\mathbf{x}_t)}{\delta} \cdot \mathbf{u}_t$$

for some $\eta_t > 0$ and $\delta > 0$, where $\mathbf{u}_t \in \mathbb{R}^d$ is *a random vector*.

It also works for nonsmooth nonconvex optimization.