机器学习引论

彭玺
pengxi@scu.edu.cn
www.pengxi.me

四川大学-机器学习引论

Test Questions

- Q1: What is the vector and matrix?
- Q2: What is the derivative of $A^T x$, w.r.t. x, where A is a matrix and x is a vector?
- Q3 : Two p x p matrices A and B, rank (A) = m, rank (B) = n, m, n <= p. what could you say about rank (A+B)? How about rank (AB)?
- Q4: What is singular value decomposition (SVD)? What is eigenvalue decomposition (EVD)? What's their relationship?
- Q5: What is the essence of matrix and matrix operation including multiplication?

Don't Worry if you fail to recall some of the answers! We will spend the next few classes to review what is needed.

四川大学-机器学习引论

提纲

四川大学-机器学习引论

一．Linear Algebra and Matrix Analysis

二．Vector Space and Optimization

提纲

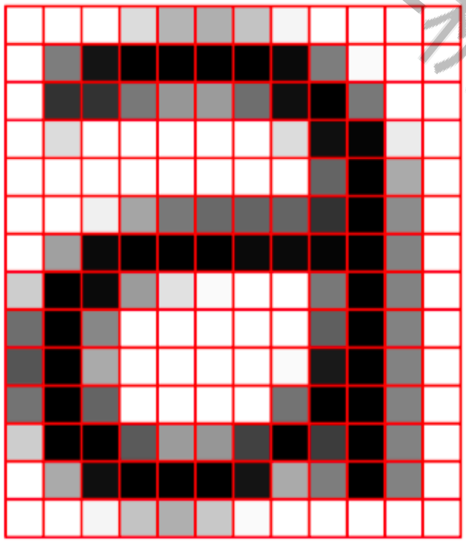
一．Linear Algebra and Matrix Analysis

二．Vector Space and Optimization

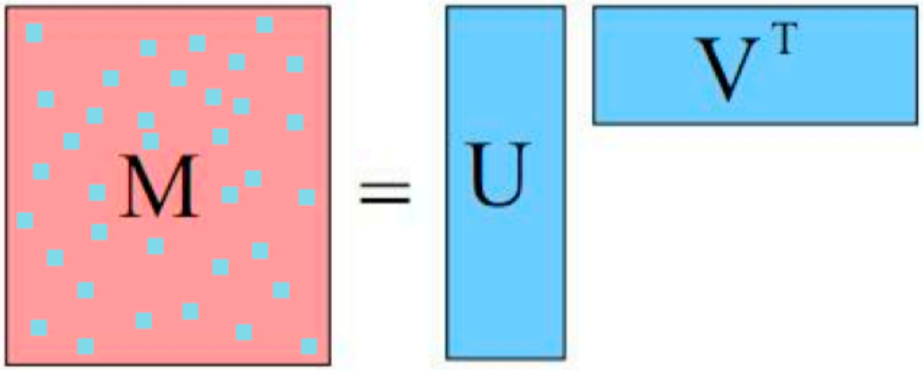
四川大学-机器学习引论

一、Linear Algebra and Matrix Analysis. Why?

- 可计算性
- 每一个图像、音频数据、文本等均以矩阵或向量形式进行表示。

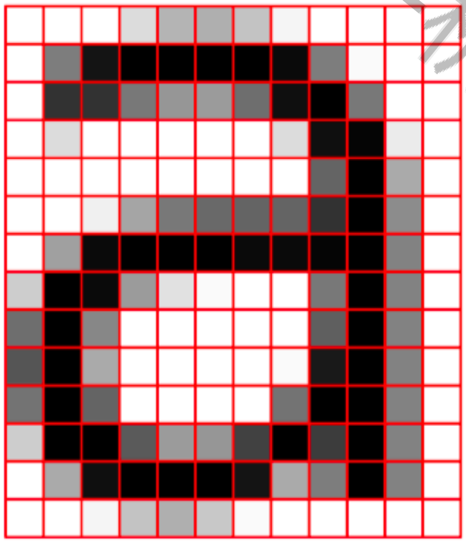


1.0	1.0	1.0	0.9	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0
1.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	1.0	1.0
1.0	0.2	0.2	0.5	0.6	0.6	0.5	0.0	0.0	0.5	1.0	1.0
1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.0	0.0	0.9	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
1.0	1.0	1.0	0.5	0.5	0.5	0.5	0.5	0.4	0.0	0.5	1.0
1.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0
0.9	0.0	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.6	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.7	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.5	1.0
0.6	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.0	0.5	1.0
0.9	0.1	0.0	0.6	0.7	0.7	0.5	0.0	0.5	0.0	0.5	1.0
1.0	0.7	0.1	0.0	0.0	0.0	0.1	0.9	0.8	0.0	0.5	1.0
1.0	1.0	1.0	0.8	0.8	0.9	1.0	1.0	1.0	1.0	1.0	1.0

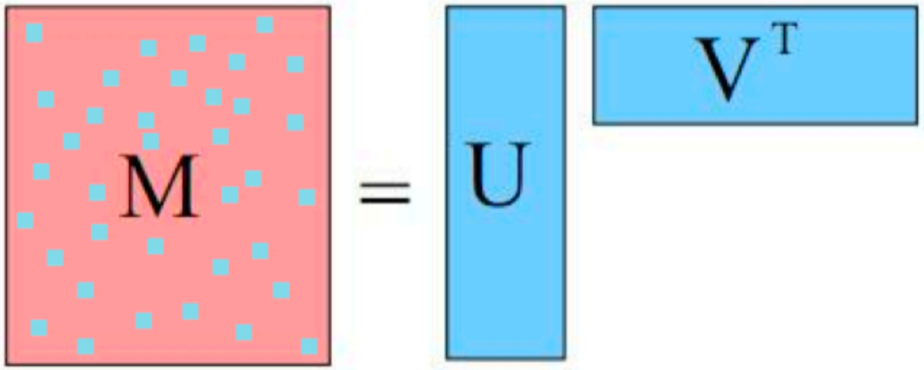


一、Linear Algebra and Matrix Analysis. Why?

- 可计算性
- 每一个图像、音频数据、文本等均以矩阵或向量形式进行表示。
- 数据处理和分析=矩阵操作



1.0	1.0	1.0	0.9	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0
1.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	1.0	1.0
1.0	0.2	0.2	0.5	0.6	0.6	0.5	0.0	0.0	0.5	1.0	1.0
1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.0	0.0	0.9	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
1.0	1.0	1.0	0.5	0.5	0.5	0.5	0.5	0.4	0.0	0.5	1.0
1.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0
0.9	0.0	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.6	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.7	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.5	1.0
0.6	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.0	0.5	1.0
0.9	0.1	0.0	0.6	0.7	0.7	0.5	0.0	0.5	0.0	0.5	1.0
1.0	0.7	0.1	0.0	0.0	0.0	0.1	0.9	0.8	0.0	0.5	1.0
1.0	1.0	1.0	0.8	0.8	0.9	1.0	1.0	1.0	1.0	1.0	1.0

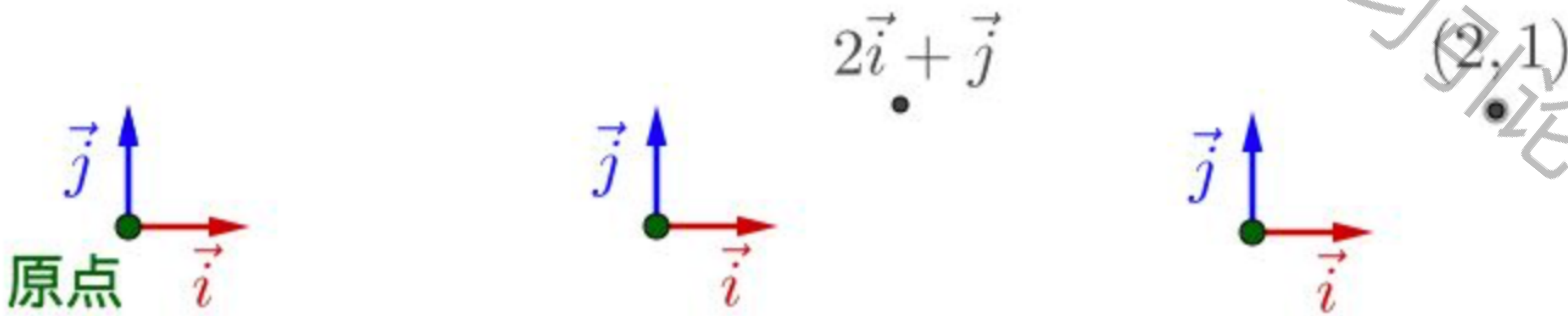


一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

线性代数：研究**线性空间**（vector space）和其上的**线性变换**的学科。

- 线性空间：



整个二维平面上的点，都可以通过 $a\vec{i}+b\vec{j}$ 的方式来表示。

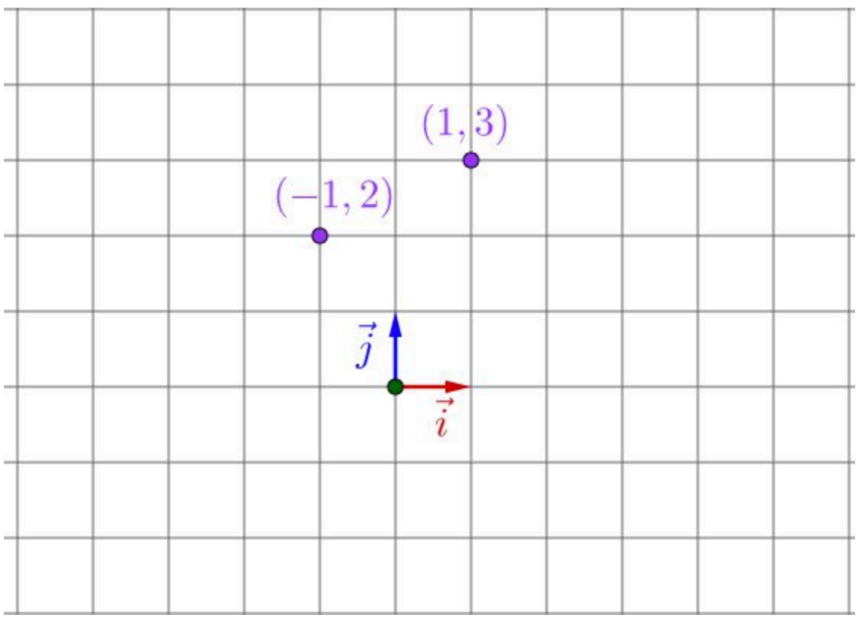
The space spanned by $\vec{i}$ and $\vec{j}$ is the two-diemsnioanl vector space.

一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

线性代数：研究**线性空间**（vector space）和其上的线性变换的学科。

- 线性空间： $a\vec{x}+b\vec{y}, a \in R, b \in R$



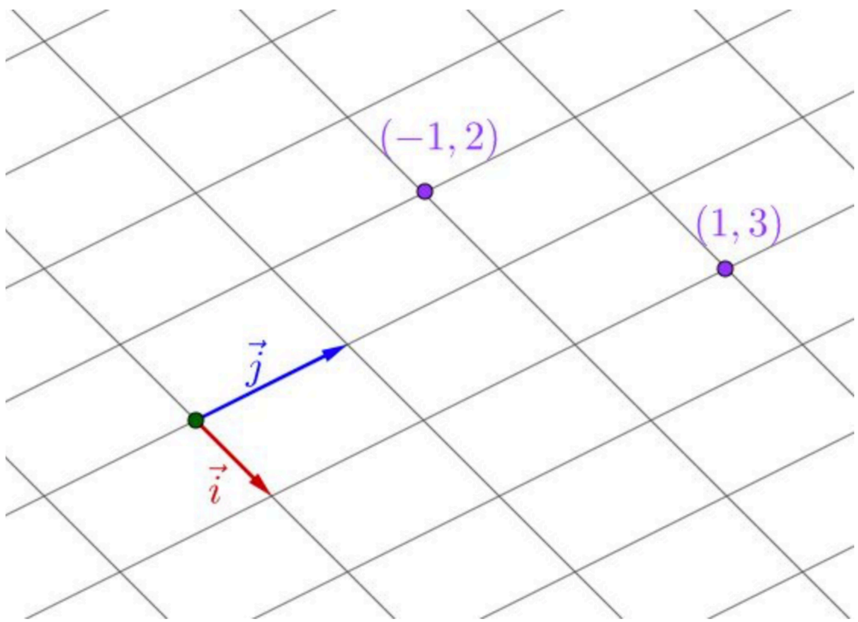
单位正交

一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

线性代数：研究**线性空间（vector space）**和其上的线性变换的学科。

- 线性空间： $a\vec{x}+b\vec{y}, a\in R, b\in R$



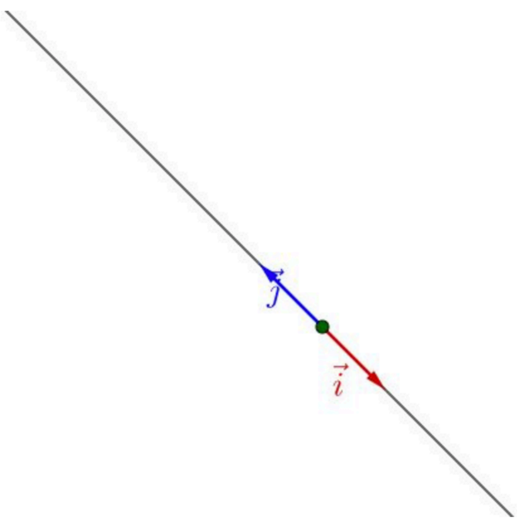
不正交，长度不等

一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

线性代数：研究**线性空间（vector space）**和其上的线性变换的学科。

- 线性空间：$a\vec{x} + b\vec{y}, a \in R, b \in R$



Linear dependent

Null space

一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

线性代数：研究**线性空间（vector space）**和其上的线性变换的学科。

- 线性空间：A vector space (also called a linear space) is a **collection of objects** called **vectors**, which may be **added** together and **multiplied** ("scaled") by numbers, called scalars. Scalars are often taken to be real numbers, but there are also vector spaces with scalar multiplication by complex numbers, rational numbers, or generally any field.

一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

线性空间（vector space）：

- A vector space over a field F is a set V together with **two operations** that satisfy the **eight axioms** listed below.
- The first operation, called **vector addition** or simply addition: $\mathbf{V} + \mathbf{V} \rightarrow \mathbf{V}$, takes any two vectors v and w and assigns to them a third vector which is commonly written as v + w, and called the sum of these two vectors. (Note that the resultant vector is also an element of the set V).
- The second operation, called **scalar multiplication**: $\mathbf{F} \times \mathbf{V} \rightarrow \mathbf{V}$ ， takes any scalar a and any vector v and gives another vector av. (Similarly, the vector av is an element of the set V).
- Consists of null space (0).

一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

线性代数：研究**线性空间**（vector space）和其上的**线性变换**的学科。

公理	说明
向量加法的结合律	$\mathbf{u}+(\mathbf{v}+\mathbf{w})=(\mathbf{u}+\mathbf{v})+\mathbf{w}$
向量加法的交换律	$\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}$
向量加法的单位元	存在一个叫做零向量的元素$\mathbf{0}\in V$，使得对任意$\mathbf{u}\in V$都满足$\mathbf{u}+\mathbf{0}=\mathbf{u}$
向量加法的逆元素	对任意$\mathbf{v}\in V$都存在其逆元素$-\mathbf{v}\in V$使得$\mathbf{v}+(-\mathbf{v})=\mathbf{0}$
标量乘法与标量的域乘法相容	$a(b\mathbf{v})=(ab)\mathbf{v}$
标量乘法的单位元	域F存在乘法单位元1满足$1\mathbf{v}=\mathbf{v}$
标量乘法对向量加法的分配律	$a(\mathbf{u}+\mathbf{v})=a\mathbf{u}+a\mathbf{v}$
标量乘法对域加法的分配律	$(a+b)\mathbf{v}=a\mathbf{v}+b\mathbf{v}$

线性空间

一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

线性代数：研究线性空间（vector space）和其上的**线性变换**的学科。

- 设V和W是在相同域K上的向量空间。$f: V \to W$被称为是线性映射，如果对于V中任何两个向量x和y与K中任何标量a，满足下列两个条件：

可加性： $f(x+y)=f(x)+f(y)$

齐次性： $f(ax)=af(x)$

- 如果A是m x n实矩阵，则A定义了一个从 R^n到 R^m的线性映射，这个映射将列向量$x \in R^n$映射到列向量$Ax \in R^m$。反过来说，在有限维向量空间之间的任何线性映射都可以用这种方式表示。
- 积分生成从在某个区间上所有可积分实函数的空间到R的线性映射。这只是把积分的基本性质（“积分的可加性”和“可从积分号内提出常数倍数”）用另一种说法表述出来。
- 微分是从所有可微分函数的空间到所有函数的空间的线性映射。

一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

Vector norm: measure the size of vector

- 1-norm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
- 2-norm: $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$
- Max-norm: $\|x\|_\infty = \max_{1 \le i \le n} |x_i|$

All the above is a special case of p-norm

$$\|x\|_p = (\sum |x_i|^p)^{1/p}$$

一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

Generally, a vector norm is a mapping $R^n \to R$, with the properties

- $||x|| \geq 0$, for all x
- $||x|| = 0$, if and only if $x = 0$
- $||\alpha x|| = |\alpha|||x||$, $\alpha \in \mathbb{R}$
- $||x + y|| \leq ||x|| + ||y||$, for all x and y

Quick Test: Are They Norms?

$$\|\mathbf{x}\|_\infty := \max\left(|x_1|, \ldots, |x_n|\right)$$

$$|x|_0 = \sum_{i=0}^{d} 1_{x_i \neq 0}$$

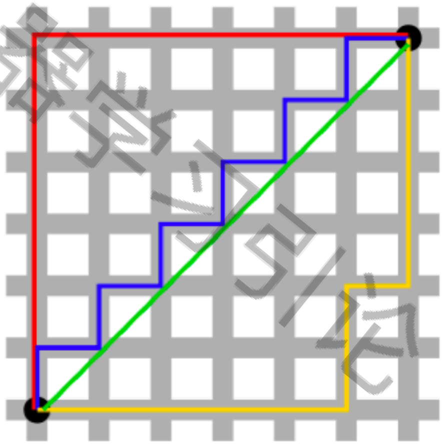
一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

Vector Norm

Norm provides a fundamental definition of "distance" in a vector space

- 1-norm: Manhattan distance
- 2-norm: Euclidean distance (most popular, e.g., MSE, Least Squares…)
- Any other alternative?



How to measure distance between vectors?

- Obvious answer: the distance between two vectors x and y is $||x - y||$, where $|| \cdot ||$ is some vector norm.
- Alternative: use the angle between two vectors x and y to measure the distance between them.
- How to calculate the angle between two vectors?

一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

Vector Norm

Angle between vectors

- The inner product between two vectors is defined by $(x, y) = x^T y$.
- This is associated with the Euclidean norm: $||x||_2 = (x, x)^{1/2}$.

- The angle θ between two vectors x and y is

$$\cos(\theta) = \frac{(x, y)}{||x||_2||y||_2}.$$

- The cosine of the angle between two vectors x and y can be used to measure the similarity between the two vectors: if x and y are close, the angle between them is small, and $\cos(\theta) \approx 1$; if x and y are orthogonal, i.e., (x, y) $= 0$, $\cos(\theta) = 0$ $(\theta = \pi/2)$.

四川大学-机器学习引论

Vector Norm

Why not just use the Euclidean distance?

- Example: term-document matrix
- Each entry tells how many times a term appears in the document:

	Doc1	Doc2	Doc3
Term1	10	1	0
Term2	10	1	0
Term3	0	0	0

- Using the Euclidean distance Documents 1 and 2 look dissimilar, and Documents 2 and 3 look similar. This is just due to the length of the documents.
- Using the cosine of the angle between document vectors Documents 1 and 2 are similar to each other and dissimilar to Document 3.

The Importance of Choosing the Right “Distance” to Measure (a.k.a. Metric)

一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

Why not Euclidean distance, formally

数据的高维（所谓的维灾），本质上是说数据的关键特性主要分布在少量维度（属性）上，其分布于所有维度张成空间的概率接近于0。然而，Euclidean距离平等的对待每一个属性，而不加区分，从而不能真实的描述出数据之间的关系和分布特性。

一、Linear Algebra and Matrix Analysis

Why not Euclidean distance, formally

数据的高维（所谓的维灾），本质上是说数据的关键特性主要分布在少量维度（属性）上，其分布于所有维度张成空间的概率接近于0。然而，Euclidean距离平等的对待每一个属性，而不加区分，从而不能真实的描述出数据之间的关系和分布特性。

如何准确（学习）度量数据之间的关系，是机器学习（人工智能）的核心问题！

四川大学-机器学习引论

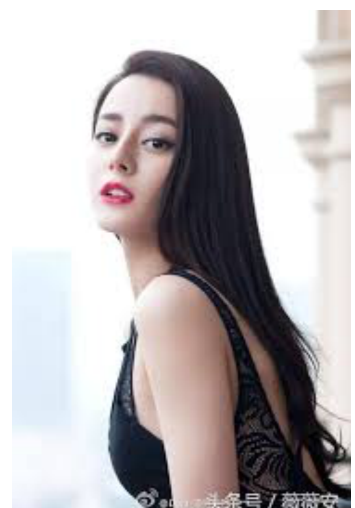
一、Linear Algebra and Matrix Analysis

四川大学-机器学习导论

Why not Euclidean distance, formally

数据的高维（所谓的维灾），本质上是说数据的关键特性主要分布在少量维度（属性）上，其分布于所有维度张成空间的概率接近于0。然而，Euclidean距离平等的对待每一个属性，而不加区分，同时是一种pairwise的距离，故不能真实的描述出数据之间的关系和分布特性。

如何准确（学习）度量数据之间的关系，是机器学习（人工智能）的核心问题！



一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

Linear Independence

- Given a set of vectors $\{v_1, v_2, \cdots, n_n\} \in \mathbb{R}^m$, with $m \geq n$, consider the set of linear combinations $y = \sum_{j=1}^{n} \alpha_j v_j$ for arbitrary coefficients α_j's.
- The vectors $\{v_1, v_2, \cdots, n_n\}$ are linearly independent, if $\sum_{j=1}^{n} \alpha_j v_j = 0$, if and only if $\alpha_j = 0$ for all $j = 1, \cdots, n$.
- A set of m linearly independent vectors of $\mathbb{R}^m$ is called a basis in $\mathbb{R}^m$: any vector in $\mathbb{R}^m$ can be expressed as a linear combination of the basis vectors.

Linear independence could be an effective metric to measure the similarity/distance between two data points lying on different/same **subspaces**.

一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

Matrix Rank

- The rank of a matrix is the maximum number of linearly independent column vectors.
- A square matrix $A \in \mathbb{R}^{n\times n}$ with rank n is called nonsingular.
- A nonsingular matrix A has an inverse A^{-1} satisfying

$$AA^{-1} = A^{-1}A = I_n.$$

- What is the rank of an out-product matrix $xy^T \in \mathbb{R}^{m\times n}$ with $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$?
- Let $A \in \mathbb{R}^{n\times n}$ be nonsingular, and let $B = A + uv^T$ with $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^n$. Then, $B^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}$.

Q: what can be used by the matrix rank?

一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

Range and Null Space

- V is a subspace of $\mathbb{R}^m$, if and only if $\alpha v_1 + \beta v_2 \in V$, for any $v_1, v_2 \in V$ and any scalars α and β.
  - Let W be the set of all points, (x, y), from $\mathbb{R}^2$ in which $x \geq 0$. Is this a subspace of $\mathbb{R}^2$?
  - Let W be the set of all points from $\mathbb{R}^3$ of the form $(0, x_2, x_3)$. Is this a subspace of $\mathbb{R}^3$?
  - Let W be the set of all points from $\mathbb{R}^3$ of the form $(1, x_2, x_3)$. Is this a subspace of $\mathbb{R}^3$?
- The range of a matrix $A \in \mathbb{R}^{m \times n}$ is defined by

$$\operatorname{ran}(A) = \{y \in \mathbb{R}^m : y = Ax \text{ for some } x \in \mathbb{R}^n\}.$$

- The null space of A is defined by

$$\operatorname{null}(A) = \{x \in \mathbb{R}^n : Ax = 0\}.$$

四川大学-机器学习引论

Range and Null Space

- It follows from the definition that $\mathrm{rank}(A) = \dim(\mathrm{ran}(A))$.
  - The dimension of a space S, denoted as $\dim(S)$ denotes the maximum number of linearly independent vectors in S.
- Show that

$$\dim(\mathrm{null}(A)) + \mathrm{rank}(A) = \text{?}$$

Rank–Nullity Theorem (plain language):

The rank and the nullity of a matrix add up to the number of its columns.

Proof: By properties of linear transformations https://en.wikipedia.org/wiki/Rank–nullity_theorem

四川大学-机器学习引论

Eigenvalues and eigenvectors

- Let A be a $n \times n$ matrix. The vector $v \neq 0$ that satisfies

$$Av = \lambda v$$

for some scalar λ is called the eigenvector of A and λ is the eigenvalue corresponding to the eigenvector v.

- An example: $A = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$

$$Av = \lambda v \rightarrow (A - \lambda I_n)v = 0 \rightarrow |A - \lambda I_n| = 0 \rightarrow \left| \begin{pmatrix} 2-\lambda & 1 \\ 1 & 3-\lambda \end{pmatrix} \right| = 0$$

Two eigenvalues $\lambda_1 = 3.62$ and $\lambda_2 = 1.38$. and two eigenvectors:

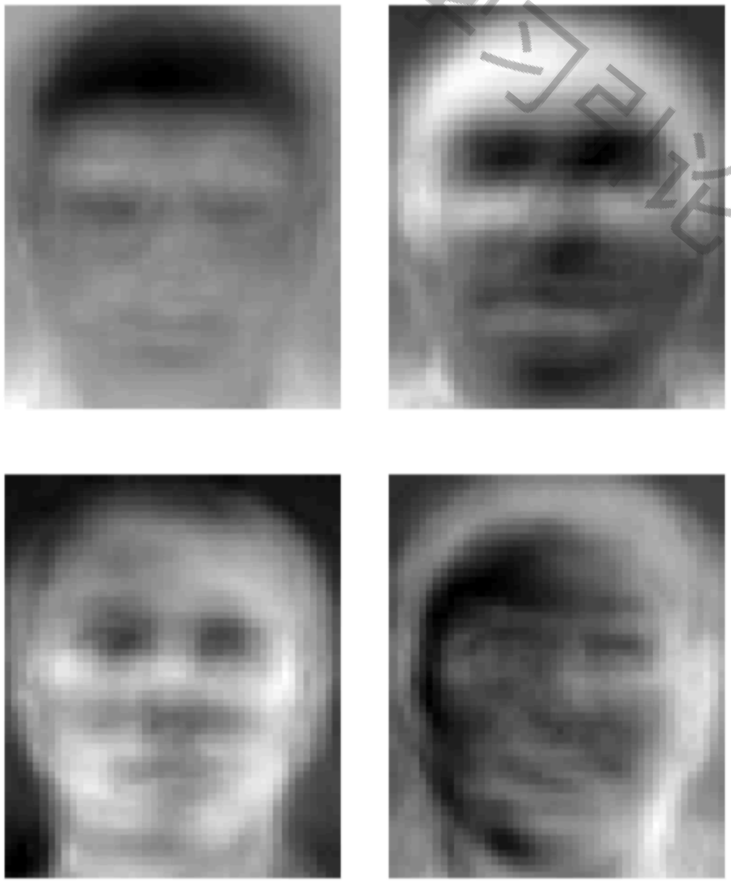
$$v_1 = \begin{pmatrix} 0.52 \\ 0.85 \end{pmatrix}, \quad v_2 = \begin{pmatrix} 0.85 \\ -0.52 \end{pmatrix}$$

一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

Eigenface

- The eigenvectors of the covariance matrix associated with a large set of normalized pictures of faces are called Eigenfaces.
- Eigenfaces reflect a set of “representative appearances” in the given collection of faces
- This is an example of principal component analysis (PCA) too.



一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

Matrix norms

- Let $||\cdot||$ be a vector norm and $A \in \mathbb{R}^{m\times n}$. The corresponding matrix norm is

$$||A|| = \sup_{x\in\mathbb{R}^n : x\neq 0} \frac{||Ax||}{||x||} = \sup_{x\in\mathbb{R}^n : ||x||=1} ||Ax||.$$

- Show that

$$\begin{aligned} ||A+B|| &\leq ||A|| + ||B|| \\ ||Ax|| &\leq ||A||\ ||x|| \\ ||AB|| &\leq ||A||\ ||B|| \end{aligned}$$

一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

Matrix norms

- $||A||_2 = \left(\max_i \lambda_i(A^TA)\right)^{1/2}$: square root of the largest eigenvalue of A^TA.
- $||A||_1 = \max_j \sum_{i=1}^m |a_{ij}|$: maximum over columns.
- $||A||_\infty = \max_i \sum_{j=1}^n |a_{ij}|$: maximum over rows.
- Frobenius norm: does not correspond to any vector norm.

$$||A||_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}$$

  - Define $\mathrm{trace}(B) = \sum_{i=1}^n b_{ii}$ for any matrix $B = (b_{ij}) \in \mathbb{R}^{n\times n}$.
  - Show that $||A||_F^2 = \mathrm{trace}(AA^T)$.

四川大学-机器学习引论

Orthogonality

- Two vectors x and y are orthogonal, if $x^T y = 0$.
- Given a set of orthogonal vectors $\{v_1, v_2, \cdots, n_n\} \in \mathbb{R}^m$, with $m \geq n$, i.e., $v_i^T v_j = 0$, for $i \neq j$, then they are linearly independent. Why?
- Let the set of orthogonal vectors v_j, $j = 1, \cdots, m$ in $\mathbb{R}^m$ be normalized, i.e., $||q|| = 1$. Then they are orthonormal, and constitute an orthonormal basis in $\mathbb{R}^m$.
- A matrix $V = [v_1, v_2, \cdots, v_m]$ is called an orthogonal matrix, if its columns are orthonormal. Prove the following properties of an orthogonal matrix:
  - An orthogonal matrix $Q \in \mathbb{R}^{m \times m}$ has rank m.
  - $Q^{-1} = Q^T$, that is, $Q^T Q = I_m$, and $QQ^T = I_m$.
  - The Euclidean length of a vector $x \in \mathbb{R}^m$ is invariant under an orthogonal transformation Q, that is,
  $$||Qx||_2^2 = ||x||_2^2.$$
  - The product of two orthogonal matrices Q and P is orthogonal.

一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

Singular Value Decomposition (SVD)

- **SVD is extremely powerful and useful**. Numerous practical applications exploit key SVD properties, i.e., its relation to the rank of a matrix and its optimal low-rank approximation of a given matrix.

Let A be an $m \times n$ matrix, with $m \geq n$. It can be factorized as

$$A = U \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T,$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal, and $\Sigma \in \mathbb{R}^{m \times n}$ is diagonal

$$\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \cdots, \sigma_n), \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0.$$

– The quantities σ_i's are called the **singular values** of A and the columns of U and V are called the **left and right singular vectors** of A respectively.

四川大学-机器学习引论

Singular Value Decomposition (SVD)

The following can be derived from SVD:

$$\begin{aligned} Av_j &= \sigma_j u_j, \\ A^T u_j &= \sigma_j v_j. \end{aligned}$$

The columns of U are the eigenvectors of AA^T. The columns of V are the eigenvectors of $A^T A$.

If A is symmetric and positive semi-definite, then $A = U\Sigma U^T$, i.e., A has the same left and right singular vectors. Show that if A is symmetric and positive semi-definite, then we can express A as $A = BB^T$ for some matrix B.

一、Linear Algebra and Matrix Analysis

四川大学-机器学习引论

Singular Value Decomposition (SVD)

Compute the norm of the matrix A:

$$||A||_2 = \sigma_1, \quad ||A||_F = \sqrt{\sum_{i=1}^{n} \sigma_i^2}.$$

The trace norm (or nuclear norm) of the matrix A is defined as:

$$||A||_* = \sum_{i=1}^{n} \sigma_i.$$

The trace norm has become very popular in recent years for matrix completion.

* E. J. Candés and T. Tao. The power of convex relaxation: Near-optimal matrix completion. IEEE Trans. Inform. Theory, 56(5), 2053-2080.
* E. J. Candés and B. Recht. Exact matrix completion via convex optimization. Found. of Comput. Math., 9 717-772.

四川大学-机器学习引论

Skinny SVD

Compact SVD: Only the r column vectors of U and r row vectors of V^T corresponding to the non-zero singular values are calculated.

$$U = (U_1\ U_2)\,\Sigma\,(V_1\ V_2)^T = U_1\Sigma_1V_1^T$$

where Σ_1 includes all nonzero singular values.

An example:

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{pmatrix} \begin{pmatrix} 9.64 & 0 \\ 0 & 5.29 \end{pmatrix} \begin{pmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{pmatrix}$$

四川大学-机器学习引论

SVD: Eckart–Young–Mirsky Theorem

Theorem 3.1 *Let* $U_k = (u_1, \cdots, u_k)$, $V_k = (v_1, \cdots, v_k)$, *and* $\Sigma_k = diag(\sigma_1, \sigma_2, \cdots, \sigma_k)$. *Define*

$$A_k = U_k \Sigma_k V_k^T.$$

Then

$$\min_{B:rank(B) \leq k} ||A - B||_F = ||A - A_k||_F = \sqrt{\sum_{i=k+1}^{n} \sigma_i^2}.$$

- A_k is the best approximation of rank k for the matrix A.

四川大学-机器学习引论

SVD: Eckart–Young–Mirsky Theorem

- This low rank approximation is useful for
  - Compression
  - Noise reduction
  - finding “concepts” or “topics” (text mining/LSI)
  - data exploration and visualizing data
  - classification (e.g. handwritten digits)
- SVD appears under different names:
  - Principal Component Analysis (PCA)
  - Latent Semantic Indexing (LSI)/Latent Semantic Analysis (LSA)
  - Karhunen-Loeve expansion/Hotelling transform (in image processing)

提纲

一．Linear Algebra and Matrix Analysis

二．Vector Space and Optimization

四川大学-机器学习引论

Take Home

- Vector space, linear transformation
- Vector norm, linear independence, matrix rank
- Eigen decomposition, Singular value decomposition, and their possible applications

四川大学-机器学习引论

二、Vector Space and Optimization

四川大学-机器学习引论

Vector Space

A vector space is any set *V* for which two operations are defined:

- **Vector addition**: any vector x_1 and x_2 in set *V* can be added to another vector $x = x_1 + x_2$, and their sum *x* is also in set *V*.
- **Scalar Multiplication**: Any vector *x* in *V* can be multiplied ("scaled") by a real number *c*, to produce a second vector *cx* which is also in V.

Examples: coordinate space, infinite coordinate space, Cartesian product of vector spaces, polynomial vector spaces, functional space…

二、Vector Space and Optimization

四川大学-机器学习引论

Vector Space with Structure

- Review: How to “measure” vectors?
  - Defining a **norm** that measures the distance
  - Defining an **inner product** that measures angles between vectors

- Topological Vector Space
  - Definition: a vector space that blends a uniform topological structure, capturing a particular notion of **convergence** of sequences of functions.
  - **Banach Space: complete vector spaces with norm defined**
  - **Hilber Space: Banach Spaces with inner product defined.** Only in a Hilbert space, both “angle” and “distance” are well-defined concepts.

二、Vector Space and Optimization

四川大学-机器学习引论

目标函数（Objective/loss function）：对客观世界的抽象，对客体之间关系的描述！一般地：

$$\begin{aligned} \min \quad & f(x,y) \\ s.t. \quad & g(x,y) \end{aligned}$$

- **Constrained Optimization**: Minimize an objective function subject to some constraints

$$\begin{aligned} \min_{x} \quad & f_0(x) \\ \text{s.t.} \quad & f_k(x) \le 0, k = 1, 2, \cdots, K \\ & h_j(x) = 0, j = 1, 2, \cdots, J \end{aligned}$$

- **Unconstrained Optimization**: no constraint is enforced, only minimizing f_0

二、Vector Space and Optimization

四川大学-机器学习引论

Optimization in ML

- Linear Regression $$\min_{w} \; ||Xw - y||^2$$
- Logistic Regression $$\min_{w} \; \sum_{i} \log(1 + \exp(-y_i x_i^\top w))$$
- Support Vector Machine (SVM) $$\min_{w} \; ||w||^2 + C \sum_{i} \xi_i$$ $$\text{s.t. } \xi_i \geq 1 - y_i x_i^\top w$$ $$\xi_i \geq 0$$
- And Many More...

What are shared by these function?

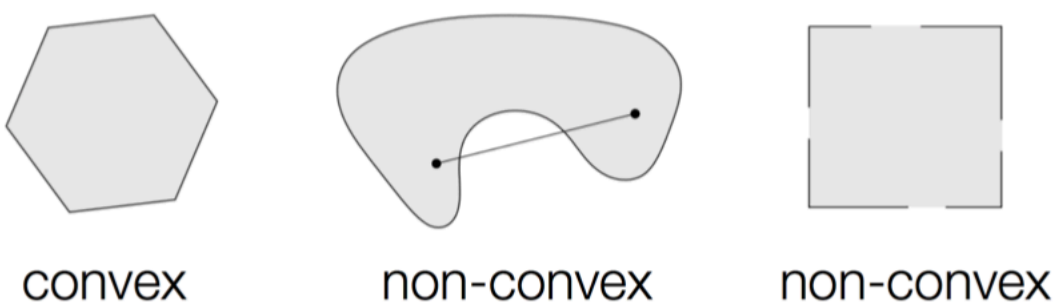
二、Vector Space and Optimization

四川大学-机器学习引论

Convex Set

- **Definition**: Any line segment joining any two elements of a convex set will lie entirely in the same set

$$x_1, x_2 \in C, 0 \leq \theta \leq 1 \implies \theta x_1 + (1-\theta) x_2 \in C$$



四川大学-机器学习引论

Convex Function

- Convexity over all lines

$$f(x) \text{ is convex } \implies f(x_0 + th) \text{ is convex in } t \text{ for all } x_0, h$$

- Positive multiple

$$f(x) \text{ is convex } \implies \alpha f(x) \text{ is convex for all } \alpha \geq 0$$

- Sum of convex functions

$$f_1(x), f_2(x) \text{ convex } \implies f_1(x) + f_2(x) \text{ is convex}$$

- Pointwise maximum

$$f_1(x), f_2(x) \text{ convex } \implies \max\{f_1(x), f_2(x)\} \text{ is convex}$$

- Affine transformation of domain

$$f(x) \text{ is convex } \implies f(Ax + b) \text{ is convex}$$

二、Vector Space and Optimization

四川大学-机器学习引论

Beauty of Convexity

- **Theorem 1:** If x is a local minimizer of a convex optimization problem, it is a **global** minimizer

- **Theorem 2:** If the gradient at *c* is zero, then *c* is the global minimum of *f(x)*

$$\nabla f(c) = 0 \iff c = x^*$$

- "A large portion of convex optimization problems, e.g., linear programming (LP), second order cone programming (SOCP) and semidefinite programming (SDP), can be solved globally optimally in polynomial time." **(By Yurii Nesterov)**

二、Vector Space and Optimization

四川大学-机器学习引论

- Unconstrained optimization Simple but extremely popular

Algorithm 1: Gradient Descent

---

while *Not Converged* **do**

$$x^{(k+1)} = x^{(k)} - \eta^{(k)} \nabla f(x)$$

end

return $x^{(k+1)}$

二、Vector Space and Optimization

四川大学-机器学习引论

- Constrained optimization – unconstrained first

A Constrained Solver: Lagrange Duality

Original optimization problem or primal problem

$$\min_{x} \ f_0(x)$$
$$\text{s.t. } f_k(x) \leq 0, k = 1, 2, \cdots, K$$
$$h_j(x) = 0, j = 1, 2, \cdots, J$$

Lagrangian

$$L(x, \lambda, v) = f_0(x) + \sum_k \lambda_k f_k(x) + \sum_j v_j h_j(x)$$

Lagrange multipliers or dual variables

Take home

- Objective function and constraint
- Convex set
- Unconstrained and constrained optimization, gradient.

四川大学-机器学习引论

Q&A

THANKS!

四川大学-机器学习引论