Mathematics for Machine Learning I

Ngày 5 tháng 6 năm 2021

Overview

1. Linear algebra
   - About this chapter
   - Related topics
2. Probability theory
   - About this chapter
   - Related topics

Linear Algebra

Linear algebra: About this chapter

- Not cover all of linear algebra
- Focus on some relevant topics to ML and DS

Text books:

- Strang, Gilbert. *Introduction to Linear Algebra*
- Nguyễn Hữu Việt Hưng. *Đại số tuyến tính*
- Goodfellow, Ian, et al. *Deep Learning*. Vol. 1. No. 2. Cambridge: MIT press, 2016
- Vũ Hữu Tiệp. *Machine Learning cơ bản*

Online course:

- MIT OpenCourseWare, *Introduction to Linear Algebra*

Linear algebra: Related topics

- Scalars, Vectors, Matrices, and Tensors
- Multiplying Matrices and Vectors
- Identity and Inverse Matrices
- Linear Dependence and Span
- Norms
- Special Kinds of Matrices and Vectors
- Eigendecomposition
- Singular Value Decomposition
- The Moore-Penrose Pseudoinverse
- The Trace Operator
- The Determinant

Scalars

- In linear algebra, real numbers or other elements of a field are called scalars
- In our lecture, a scalar is just a single number: integer, real number, rational number, etc
- It is denoted with an italic font:

$$n, a, x$$

Vectors

- In mathematics and physics, a vector is an element of a vector space
- A vector space (also called a linear space) is a set of objects called vectors, which may be added together and multiplied ("scaled") by numbers, called scalars. Example: Euclidean vector, a geometric object that has magnitude (or length) and direction
- In computer science, a vector is a one-dimensional array data structure

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \tag{1}$$

- Example notation for its type and size $\mathbf{x} \in \mathbb{R}^m$

Matrices

- A matrix is a two-dimensional array data structure. In our lecture, a matrix is a two-dimensional array of numbers

$$\mathbf{A} = \begin{pmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,n} \\ A_{3,1} & A_{3,2} & \cdots & A_{3,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1} & A_{m,2} & \cdots & A_{m,n} \end{pmatrix} \tag{2}$$

- Example notation for its type and size $\mathbf{A} \in \mathbb{R}^{m \times n}$

Tensors

- A tensor is an algebraic object that describes a (multilinear) relationship between sets of algebraic objects related to a vector space
- In our lecture, a tensor is an array of numbers, that may have
  - zero dimensions, and be a scalar
  - one dimension, and be a vector
  - two dimensions, and be a matrix
  - or more dimensions

Matrix Transpose

- The transpose of the matrix can be thought of as a mirror image across the main diagonal: $(\mathbf{A}^\top)_{i,j} = \mathbf{A}_{j,i}$

$$\mathbf{A} = \begin{pmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,n} \\ A_{3,1} & A_{3,2} & \cdots & A_{3,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1} & A_{m,2} & \cdots & A_{m,n} \end{pmatrix}$$

, then

$$\mathbf{A}^\top = \begin{pmatrix} A_{1,1} & A_{2,1} & A_{3,1} & \cdots & A_{m,1} \\ A_{1,2} & A_{2,2} & A_{3,2} & \cdots & A_{m,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{1,n} & A_{2,n} & A_{3,n} & \cdots & A_{m,n} \end{pmatrix} \tag{3}$$

- Property:

$$(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top \tag{4}$$

Matrix Sum

- We can add matrices to each other, as long as they have the same shape, just by adding their corresponding elements
- $$\mathbf{C} = \mathbf{A} + \mathbf{B} \tag{5}$$
- $$\mathbf{C}_{i,j} = \mathbf{A}_{i,j} + \mathbf{B}_{i,j} \tag{6}$$

Matrix Product (Dot Product)

- In order for defining the product of matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A}$ must have the same number of columns as $\mathbf{B}$ has rows. Precisely, if $\mathbf{A}$ is of shape $m \times n$ and $\mathbf{B}$ is of shape $n \times p$, then

$$\mathbf{C} = \mathbf{AB} \tag{7}$$

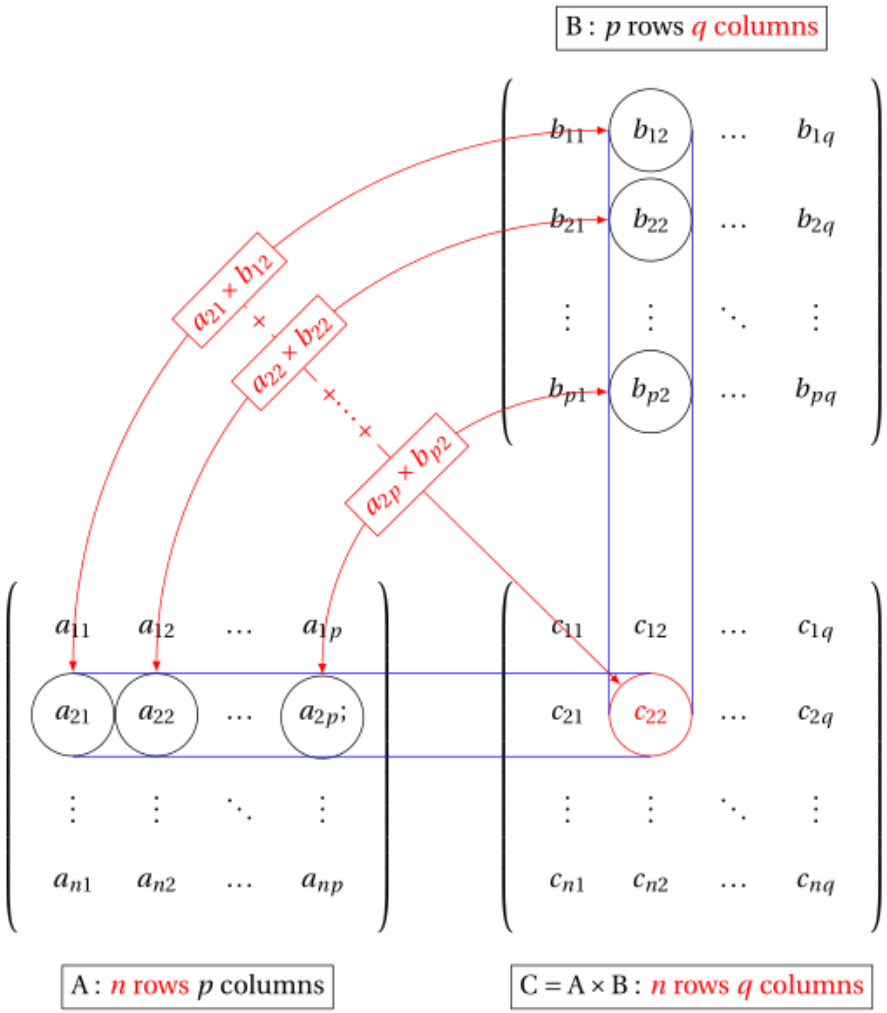
  is of shape $m \times p$
- The product operation is defined by

$$\mathbf{C}_{i,j} = \sum_{k} \mathbf{A}_{i,k} \mathbf{B}_{k,j} \tag{8}$$

- The dot product between two vectors $\mathbf{x}$ and $\mathbf{y}$ of the same dimension is the matrix product $\mathbf{x}^{\mathsf{T}}\mathbf{y}$.

Matrix Product (Dot Product)

Hình: Matrix multiplication. Source: Alain Matthes, TEXample.net



Linear Equations System

A system of linear equations is

$$\mathbf{Ax} = \mathbf{b}, \tag{9}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a known matrix, $\mathbf{b} \in \mathbb{R}^m$ is a known vector, and $\mathbf{x} \in \mathbb{R}^n$ is a vector of unknown variables we would like to solve for. Each element x_i of $\mathbf{x}$ is one of these unknown variables. Each row of $\mathbf{A}$ and each element of $\mathbf{b}$ provide a constraint. All the constraints are

$$A_{1,1}x_1 + A_{1,2}x_2 + \cdots + A_{1,n}x_n = b_1, \tag{10}$$

$$A_{2,1}x_1 + A_{2,2}x_2 + \cdots + A_{2,n}x_n = b_2, \tag{11}$$

$$\cdots \tag{12}$$

$$A_{m,1}x_1 + A_{m,2}x_2 + \cdots + A_{m,n}x_n = b_m. \tag{13}$$

Solving Linear Equations System

A linear system of equations can have:

- No solution
- Many solutions
- Exactly one solution: Multiplication by the inverse matrix of $\mathbf{A}$ to obtain the solution.

Identity and Inverse Matrices

- We denote the **identity matrix** that preserves n-dimensional vectors as $\mathbf{I}_n$. Formally, $\mathbf{I}_n \in \mathbb{R}^{n \times n}$, and $\forall \mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{I}_n \mathbf{x} = \mathbf{x}. \tag{14}$$

- The structure of the identity matrix is simple: all the entries along the main diagonal are 1, while all the other entries are zero.

$$\mathbf{I}_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \tag{15}$$

- The **inverse matrix** of $\mathbf{A}$ is denoted as $\mathbf{A}^{-1}$, and it is defined as the matrix such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n. \tag{16}$$

Solving a linear equations system using an inverse

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{17}$$

$$\mathbf{A}^{-1}\mathbf{A}\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \tag{18}$$

$$\mathbf{I}_n\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \tag{19}$$

- Useful for abstract analysis
- Numerically unstable

Invertible Matrices

A matrix cannot be inverted (called singular or degenerate matrix) if

- More rows than columns or more columns than rows
- Redundant rows/columns (linearly dependent, low rank)

Norms

- Given a vector space $\mathbf{V}$ over a subfield $\mathbf{F}$ of the complex numbers $\mathbb{C}$, a norm on $\mathbf{V}$ is a nonnegative-valued real-valued function $f : \mathbf{V} \to \mathbb{R}$ with the following properties, where $|a|$ denotes the usual absolute value of a,
- For all a in $\mathbf{F}$ and all $\mathbf{u}, \mathbf{v}$ in $\mathbf{V}$,
  - 1. $f(\mathbf{u} + \mathbf{v}) \leq f(\mathbf{u}) + f(\mathbf{v})$ (being subadditive or satisfying the triangle inequality).
  - 2. $f(a\mathbf{v}) = |a| f(\mathbf{v})$ (being absolutely homogeneous or absolutely scalable).
  - 3. If $f(\mathbf{v}) = 0$ then $\mathbf{v} = \mathbf{0}$ is the zero vector (being positive definite or being point-separating).
- A seminorm on $\mathbf{V}$ is a function $f : \mathbf{V} \to \mathbb{R}$ with the properties 1 and 2 above.

Norms

In machine learning, we usually measure the size of vectors using the L^p norm, given by

$$||\mathbf{x}||_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}}, \tag{20}$$

for $p \in \mathbb{R}, p \geq 1$.

- The L^2 norm, with $p = 2$, is known as the Euclidean norm, which is simply the Euclidean distance from the origin to the point identified by $\mathbf{x}$. It is also common to measure the size of a vector using the squared L^2 norm, which can be calculated simply as $\mathbf{x}^\top \mathbf{x}$.
- The squared L^2 norm is more convenient to work with mathematically and computationally than the L^2 norm itself. For example, each derivative of the squared L^2 norm with respect to each element of x depends only on the corresponding element of x, while all the derivatives of the L^2 norm depend on the entire vector.

Norms

- In many contexts, the squared L^2 norm may be undesirable because it increases very slowly near the origin. In several machine learning applications, it is important to discriminate between elements that are exactly zero and elements that are small but nonzero. In these cases, we turn to a function that grows at the same rate in all locations, but that retains mathematical simplicity: the L^1 norm. The L^1 norm may be simplified to

$$||\mathbf{x}||_1 = \sum_i |x_i|. \tag{21}$$

- We sometimes measure the size of the vector by counting its number of nonzero elements. Some persons refer to this function as the 'L^0' norm, but this is incorrect terminology. The number of nonzero entries in a vector is not a norm, because scaling the vector by α does not change the number of nonzero entries. The L^1 norm is often used as a substitute for the number of nonzero entries.

Norms

- One other norm that commonly arises in machine learning is the L^∞ norm, also known as the max norm. This norm simplifies to the absolute value of the element with the largest magnitude in the vector

$$||\mathbf{x}||_\infty = \max_i |x_i|. \tag{22}$$

- Sometimes we may also wish to measure the size of a matrix. In the context of deep learning, the most common way to do this is with the otherwise obscure Frobenius norm

$$||\mathbf{A}||_F = \sqrt{\sum_{i,j} A_{i,j}^2}, \tag{23}$$

  which is analogous to the L^2 norm of a vector.
- The dot product of two vectors can be rewritten in terms of norms. Specifically, $\mathbf{x}^\mathsf{T}\mathbf{y} = ||\mathbf{x}||_2||\mathbf{y}||_2 cos\theta$, where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$.

Special Kinds of Matrices and Vectors

- **Diagonal matrices.** A matrix $\mathbf{D}$ is diagonal if and only if $D_{i,j} = 0$ for all $i \neq j$. Not all diagonal matrices need be square. It is possible to construct a rectangular diagonal matrix. Nonsquare diagonal matrices do not have inverses, but we can still multiply by them cheaply.
- A **symmetric matrix** is any matrix that is equal to its own transpose $\mathbf{A} = \mathbf{A}^\top$.
- A **unit vector** is a vector with unit norm $||\mathbf{x}||_2 = 1$.
- A vector $\mathbf{x}$ and a vector $\mathbf{y}$ are **orthogonal** to each other if $\mathbf{x}^\top \mathbf{y} = 0$. If the vectors not only are orthogonal but also have unit norm, we call them **orthonormal**.
- An **orthogonal matrix** is a square matrix whose rows are mutually orthonormal and whose columns are mutually orthonormal $\mathbf{A}^\top \mathbf{A} = \mathbf{A}\mathbf{A}^\top = I$.

Eigendecomposition

- An eigenvector of a square matrix $\mathbf{A}$ is a nonzero vector $\mathbf{v}$ such that multiplication by $\mathbf{A}$ alters only the scale of $\mathbf{v}$

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}. \tag{24}$$

- The scalar λ is known as the eigenvalue corresponding to this eigenvector.
- If $\mathbf{v}$ is an eigenvector of $\mathbf{A}$, then so is any rescaled vector $s\mathbf{v}$ for $s \in \mathbb{R}$, $s \neq 0$. Moreover, $s\mathbf{v}$ still has the same eigenvalue. For this reason, we usually look only for unit eigenvectors.

Eigendecomposition

- Suppose that a matrix $\mathbf{A}$ has n linearly independent eigenvectors $\{\mathbf{v}^1, \ldots, \mathbf{v}^n\}$ with corresponding eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$. We may concatenate all the eigenvectors to form a matrix $\mathbf{V}$ with one eigenvector per column: $\mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^n]$. Likewise, we can concatenate the eigenvalues to form a vector $\lambda = [\lambda_1, \ldots, \lambda_n]^\intercal$. The eigendecomposition of $\mathbf{A}$ is then given by

$$\mathbf{A} = \mathbf{V} diag(\lambda) \mathbf{V}^{-1}. \tag{25}$$

- Not every matrix can be decomposed into eigenvalues and eigenvectors. In some cases, the decomposition exists but involves complex rather than real numbers. Specifically, every real symmetric matrix can be decomposed into an expression using only real-valued eigenvectors and eigenvalue.

Eigendecomposition

- A matrix whose eigenvalues are all positive is called positive definite. A matrix whose eigenvalues are all positive or zero valued is called positive semidefinite.
- Likewise, if all eigenvalues are negative, the matrix is negative definite, and if all eigenvalues are negative or zero valued, it is negative semidefinite.
- Positive semidefinite matrices are interesting because they guarantee that $\forall \mathbf{x}, \mathbf{x}^\intercal \mathbf{A}\mathbf{x} \geq 0$. Positive definite matrices additionally guarantee that $\mathbf{x}^\intercal \mathbf{A}\mathbf{x} = 0 \implies \mathbf{x} = \mathbf{0}$.

Singular Value Decomposition

- The singular value decomposition (SVD) provides another way to factorize a matrix, into singular vectors and singular values.
- Every real matrix has a singular value decomposition, but the same is not true of the eigenvalue decomposition. For example, if a matrix is not square, the eigendecomposition is not defined, and we must use a singular value decomposition instead.
- The singular value decomposition is similar to the eigendecomposition, except this time we will write $\mathbf{A}$ as a product of three matrices:

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathsf{T}}, \tag{26}$$

where $\mathbf{A}$ is an $m \times n$ matrix, $\mathbf{U}$ is an $m \times m$ matrix, $\mathbf{D}$ is an $m \times n$ matrix, and $\mathbf{V}$ is an $n \times n$ matrix. Each of these matrices is defined to have a special structure. The matrices U and V are both defined to be orthogonal matrices. The matrix $\mathbf{D}$ is defined to be a diagonal matrix. Note that $\mathbf{D}$ is not necessarily square.

Moore-Penrose Pseudoinverse

- Matrix inversion is not defined for matrices that are not square. Suppose we want to make a left-inverse $\mathbf{B}$ of a matrix $\mathbf{A}$ so that we can solve a linear equation $\mathbf{Ax} = \mathbf{y}$ by left-multiplying each side to obtain $\mathbf{x} = \mathbf{By}$. Depending on the structure of the problem, it may not be possible to design a unique mapping from $\mathbf{A}$ to $\mathbf{B}$.
- If $\mathbf{A}$ is taller than it is wide, then it is possible for this equation to have no solution. If $\mathbf{A}$ is wider than it is tall, then there could be multiple possible solutions.
- The Moore-Penrose pseudoinverse enables us to make some headway in these cases. The pseudo inverse of $\mathbf{V}$ is defined as a matrix

$$\mathbf{A}^{+} = \lim_{\alpha \searrow 0} (\mathbf{A}^{\intercal}\mathbf{A} + \alpha \mathbf{I})^{-1}\mathbf{A}^{\intercal}. \tag{27}$$

Moore-Penrose Pseudoinverse

- Practical algorithms for computing the pseudoinverse are based not on this definition, but rather on the formula $\mathbf{A}^+ = \mathbf{V}\mathbf{D}^+\mathbf{U}^\mathsf{T}$, where $\mathbf{U}$, $\mathbf{D}$, and $\mathbf{V}$ are the singular value decomposition of $\mathbf{A}$, and the pseudoinverse $\mathbf{D}^+$ of a diagonal matrix D is obtained by taking the reciprocal of its nonzero elements then taking the transpose of the resulting matrix.
- When $\mathbf{A}$ has more columns than rows, then solving a linear equation using the pseudoinverse provides one of the many possible solutions. Specifically, it provides the solution $\mathbf{x} = \mathbf{A}^+\mathbf{y}$ with minimal Euclidean norm $||\mathbf{x}||_2$ among all possible solutions.
- When $\mathbf{A}$ has more rows than columns, it is possible for there to be no solution. In this case, using the pseudoinverse gives us the $\mathbf{x}$ for which $\mathbf{A}\mathbf{x}$ is as close as possible to $\mathbf{y}$ in terms of Euclidean norm $||\mathbf{A}\mathbf{x} - \mathbf{y}||_2$.

Trace Operator

The trace operator gives the sum of all the diagonal entries of a matrix:

$$Tr(\mathbf{A}) = \sum_i A_{i,i}. \tag{28}$$

Properties

$$Tr(\mathbf{A}) = Tr(\mathbf{A}^\mathsf{T}) \tag{29}$$

$$Tr(\mathbf{ABC}) = Tr(\mathbf{CAB}) = Tr(\mathbf{BCA}). \tag{30}$$

This invariance to cyclic permutation holds even if the resulting product has a different shape. For example, for $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times m}$, we have

$$Tr(\mathbf{AB}) = Tr(\mathbf{BA}), \tag{31}$$

even though $\mathbf{AB} \in \mathbb{R}^{m \times m}$ and $\mathbf{BA} \in \mathbb{R}^{n \times n}$.

Determinant

- The determinant of a square matrix, denoted $det(\mathbf{A})$, is a function that maps matrices to real scalars.
- The determinant is equal to the product of all the eigenvalues of the matrix.
- The absolute value of the determinant can be thought of as a measure of how much multiplication by the matrix expands or contracts space. If the determinant is 0, then space is contracted completely along at least one dimension, causing it to lose all its volume. If the determinant is 1, then the transformation preserves volume.

Probability theory

Probability theory: About this chapter

- Not cover all of Probability theory
- Focus on some relevant topics to ML and DS

Text books:

- Ross, Sheldon. *A first course in probability*
- Đặng Hùng Thắng. *Mở đầu về lí thuyết xác suất và các ứng dụng*
- Goodfellow, Ian, et al. *Deep Learning*. Vol. 1. No. 2. Cambridge: MIT press, 2016
- Vũ Hữu Tiệp. *Machine Learning cơ bản*

Online course:

- MIT OpenCourseWare, *Introduction to Probability*

Probability theory: Related topics

- Random Variables
- Probability Distributions
- Marginal Probability
- Conditional Probability
- The Chain Rule of Conditional Probabilities
- Independence and Conditional Independence
- Expectation, Variance and Covariance
- Common Probability Distributions
- Bayes' Rule