Foundations of Data Science

Lecture 4: Vector, Matrix and Operations

MING GAO

DaSE@ECNU
(for course related communications)
mgao@sei.ecnu.edu.cn

Oct. 9, 2016

Outline

1. Vector
2. Matrix
3. Matrix derivatives
4. Linear Model
5. Eigenvalue and Eigenvector

Linear equations and matrix

Linear equations

- The subject of algebra arose from studying equations.
- If $x_1, x_2, \cdots, x_n$ are variables and $a_1, a_2, \cdots, a_n$ and c are constant, then the equation $a_1x_1 + a_2x_2 + \cdots + a_nx_n = c$ is said to be a linear equation, where a_i are the coefficient.
- More generally, a linear system consisting of m equations in n unknowns will look like:

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= c_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= c_2 \\
&\cdots \\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= c_m
\end{aligned}
$$

- The main problem is to find the solution set of a linear system (Gaussian reduction). While that is not the focus of this course

Vector

An n-tuple (pair, triple, quadruple, ...) of scalars can be written as a horizontal row or vertical column. A column is called a vector. For example $x = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}$ and $x^T = [x_1, x_2, \cdots, x_n]$

Operations

- Addition: $x + y = [x_1 + y_1, x_2 + y_2, \cdots, x_n + y_n]$
  - $x + y = y + x$
  - $cx = [cx_1, cx_2, \cdots, cx_n]$ and $0x = \overrightarrow{0}$
  - $x + (y + z) = (x + y) + z$
- Manipulation: $x^T y = \sum_{i=1}^{n} x_i y_i$
  - $x^T y = y^T x$
  - $(x + y)^T z = x^T z + y^T z$

Examples of vector

Examples

- Entity: an entity can be modeled as a vector $x = [x_1, x_2, \cdots, x_n]$, where n denotes the number of features and x_i denotes the value of the $i-$th feature. For example, patients, email, student, and user, etc.
- Set: given a universal set, a subset of the universal set can be model as a binary vector $x = [0, 1, 1, \cdots, 0]^T$, where the dimension of X is the size of universal set and $x_i = 1$ means that the $i-$th item exists in the set. For example, document VS. words, vertex VS. neighbors, string VS. n-gram, user VS. products, etc.
- Distribution: given a discrete sample space Ω, PMF can be modeled as a vector $p = [p_1, p_2, \cdots, p_n]$, where n is the cardinality of Ω, $0 \leq p_i \leq 1$ and $\sum_{i=1}^{n} p_i = 1$. For example, document VS. topics, Markov chain VS. states, tweet VS. polarity, entity VS. classes, etc.
- Latent vector: matrix factorization, topic modeling, and word embedding, etc.

How far from two vectors

Distance

Distance is a numerical description of how far apart vectors are. Given two vectors x and y, there are many ways to measure distance of two vectors.

- Distance in Euclidean space: the Minkowski distance of order p (p-norm distance) is defined as
    - 1-norm distance (Manhattan distance): $\sum_{i=1}^{n} |x_i - y_i|$
    - 2-norm distance (Euclidean distance): $\left(\sum_{i=1}^{n} (x_i - y_i)^2\right)^{1/2}$
    - p-norm distance: $\left(\sum_{i=1}^{n} (x_i - y_i)^p\right)^{1/p}$
    - Infinity norm distance: $\lim_{p\to\infty} \left(\sum_{i=1}^{n} (x_i - y_i)^p\right)^{1/p} = \max\{x_i - y_i | i = 1, 2, \cdots, n\}$
- Mahalanobis distance: it is defined as a dissimilarity between two random vectors X and Y of the same distribution with covariance matrix S: $D(X, Y) = \left((X - Y)^T S^{-1} (X - Y)\right)^{1/2}$. If $S = I$, the Mahalanobis distance reduces to the Euclidean distance

Matrix

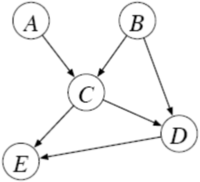
Definition

An $m \times n$ matrix $A = (a_{ij})$ $(1 \le i \le m, 1 \le j \le n)$ is a rectangular array of mn scalars in m rows and n columns, such as

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

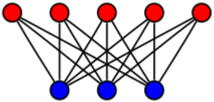
- The set of all $m \times n$ matrices with real entries will be denoted by $\mathbb{R}^{m \times n}$
- A is called the identity matrix if $a_{ii} = 1$ and $a_{ij} = 0$ $(i \neq j)$
- If $m = n$, A is called a square matrix
- $A^T = (a_{ji}) (\in \mathbb{R}^{n \times m})$ is a $n \times m$ matrix
- If A is a $n \times n$ square matrix, $Trace(A) = \sum_i^n a_{ii} = Trace(A^T)$

Examples of matrix



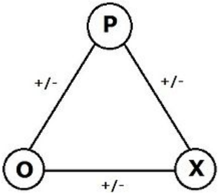
Homogeneous graph

Vertex can be Web page, user, protein, road, route, etc. Edge can be directed, undirected, weighted or labeled.



Heterogeneous graph

Vertex can be user-Web page, user-product, user-service, user-paper, etc.



Signed graph

user friend and enemy networks, like and dislike networks, etc.

Examples of matrix Cont.

Image processing

color, texture and shape, etc

Object	F1	F2	F3	F4
1	1.3	12.5	0.4	234.8
2	2.2	23.1	0.45	255.6
3	1.9	7.4	0.54	301.3
4	?	14.2	0.51	278.3

Feature representation

Many applications can be modeled in this manner, such as search engine, email classification, disease diagnosis, churn predication, anomaly detection, etc.

Operations of matrix

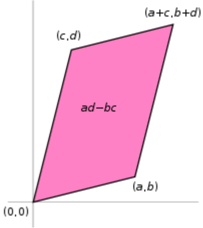
Operations

- Addition: $A + B = (a_{ij} + b_{ij})$
  - $A + cB = (a_{ij} + cb_{ij})$, especially $A - B = (a_{ij} - b_{ij})$, where c is a scalar
  - $A + B = B + A$
  - $A + (B + C) = (A + B) + C$
- Manipulation: If A is an $n \times m$ matrix and B is an $m \times p$ matrix, then $AB = (\sum_{k=1}^{m} a_{ik} b_{kj})$
  - Not commutative: $AB \neq BA$
  - Distributive over matrix addition: $(A + B)C = AC + BC$
  - Scalar multiplication is compatible with matrix multiplication: $\lambda AB = (\lambda A)B = A(\lambda B)$
  - $(AB)^T = B^T A^T$
  - $Trace(AB) = Trace(BA)$ and $Trace(ABCD) = Trace(BCDA) = Trace(CDAB) = Trace(DABC)$. In general, $Trace(ABC) \neq Trace(ACB)$

Determinant of square matrix

Definition

The determinant of a matrix A is denoted $det(A)$ or $|A|$. It can be viewed as the scaling factor of the transformation described by the matrix.



In case of 2×2 matrix

$$det(A) = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

Properties

- $det(I_n) = 1$, where I_n is a $n \times n$ identity matrix
- $det(A) = det(A^T)$ and $det(A^{-1}) = \frac{1}{det(A)}$
- $det(AB) = det(A)det(B)$ and $det(cA) = c^n det(A)$, where c is a constant.

Matrix derivatives

Type	scalar	vector	matrix
scalar	$\frac{\partial y}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial Y}{\partial x}$
vector	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	
matrix	$\frac{\partial y}{\partial X}$		

Derivatives by scalar

Assume that $x, y, a \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^{n\times 1}$, $X, Y, A \in \mathbb{R}^{n\times m}$, and $\mathbf{y}, \mathbf{a} \in \mathbb{R}^{m\times 1}$. Let a, $\mathbf{a}$ and A be constant scalar, vector and matrix.

- $\frac{\partial y}{\partial x}$ and $\frac{\partial a}{\partial x} = 0$
- $\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \cdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}$ and $\frac{\partial \mathbf{a}}{\partial x} = \mathbf{0}$ (vector)
- $\frac{\partial Y}{\partial x} = \begin{bmatrix} \frac{\partial y_{11}}{\partial x} & \cdots & \frac{\partial y_{1m}}{\partial x} \\ \cdots & \cdots & \cdots \\ \frac{\partial y_{n1}}{\partial x} & \cdots & \frac{\partial y_{nm}}{\partial x} \end{bmatrix}$ and $\frac{\partial A}{\partial x} = \mathbf{0}$ (matrix)

Matrix derivatives Cont.

Derivatives by vector

- $\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \cdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix}^T$ and $\frac{\partial a}{\partial \mathbf{x}} = \mathbf{0}^T$ (vector)
- $\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \cdots & \cdots & \cdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$, $\frac{\partial \mathbf{a}}{\partial \mathbf{x}} = \mathbf{0}$ (matrix) and $\frac{\partial \mathbf{x}}{\partial \mathbf{x}} = \mathbf{I}$ (matrix)

Derivatives by matrix

$$\frac{\partial y}{\partial X} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \cdots & \frac{\partial y}{\partial x_{n1}} \\ \cdots & \cdots & \cdots \\ \frac{\partial y}{\partial x_{1m}} & \cdots & \frac{\partial y}{\partial x_{nm}} \end{bmatrix} \text{ and } \frac{\partial a}{\partial X} = \mathbf{0}^T \text{ (matrix)}$$

Common properties of matrix derivatives

Properties

- c1 $\frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^T\mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}^T$
- c2 $\frac{\partial \mathbf{x}^T\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}^T$
- c3 $\frac{\partial (\mathbf{x}^T\mathbf{a})^2}{\partial \mathbf{x}} = 2\mathbf{x}^T\mathbf{a}\mathbf{a}^T$
- c4 $\frac{\partial A\mathbf{x}}{\partial \mathbf{x}} = A$ and $\frac{\partial \mathbf{x}^T A}{\partial \mathbf{x}} = A^T$
- c5 $\frac{\partial \mathbf{x}^T A\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^T(A + A^T)$

Proof c2. Let $s = \mathbf{x}^T\mathbf{x} = \sum_{i=1}^n x_i^2$. Then, $\frac{\partial s}{\partial x_i} = 2x_i$. So, $\frac{\partial \mathbf{x}^T\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}^T$.

Properties of matrix derivatives cont.

Properties for scalar by scaler

ss1 $\frac{\partial(u+v)}{\partial x} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial x}$

ss2 $\frac{\partial uv}{\partial x} = u\frac{\partial v}{\partial x} + v\frac{\partial u}{\partial x}$ (product rule)

ss3 $\frac{\partial g(u)}{\partial x} = \frac{\partial g(u)}{\partial u}\frac{\partial u}{\partial x}$ (chain rule)

ss4 $\frac{\partial f(g(u))}{\partial x} = \frac{\partial f(g)}{\partial g}\frac{\partial g(u)}{\partial u}\frac{\partial u}{\partial x}$ (chain rule)

Properties for vector by scaler

vs1 $\frac{\partial(a\mathbf{u}+\mathbf{v})}{\partial x} = a\frac{\partial \mathbf{u}}{\partial x} + \frac{\partial \mathbf{v}}{\partial x}$, where a is not a function of x.

vs2 $\frac{\partial Au}{\partial x} = A\frac{\partial u}{\partial x}$ where A is not a function of x.

vs3 $\frac{\partial \mathbf{u}^T}{\partial x} = \left(\frac{\partial \mathbf{u}}{\partial x}\right)^T$

vs4 $\frac{\partial g(\mathbf{u})}{\partial x} = \frac{\partial g(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial x}$ (chain rule)

vs5 $\frac{\partial f(g(\mathbf{u}))}{\partial x} = \frac{\partial f(g)}{\partial g}\frac{\partial g(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial x}$ (chain rule)

Properties of matrix derivatives cont.

Properties for matrix by scaler

ms1 $\frac{\partial aU}{\partial x} = a\frac{\partial U}{\partial x}$, where a is not a function of x.

ms2 $\frac{\partial AUB}{\partial x} = A\frac{\partial U}{\partial x}B$ where A and B are not a function of x.

ms3 $\frac{\partial (U+V)}{\partial x} = \frac{\partial U}{\partial x} + \frac{\partial V}{\partial x}$

ms4 $\frac{\partial UV}{\partial x} = U\frac{\partial V}{\partial x} + \frac{\partial U}{\partial x}V$ (product rule)

Properties for scalar by vector

sv1 $\frac{\partial (au+v)}{\partial \mathbf{x}} = a\frac{\partial u}{\partial \mathbf{x}} + \frac{\partial v}{\partial \mathbf{x}}$, where a is not a function of x.

sv2 $\frac{\partial uv}{\partial \mathbf{x}} = u\frac{\partial v}{\partial \mathbf{x}} + v\frac{\partial u}{\partial \mathbf{x}}$ (product rule)

sv3 $\frac{\partial f(g(u))}{\partial \mathbf{x}} = \frac{\partial f(g)}{\partial g}\frac{\partial g(u)}{\partial u}\frac{\partial u}{\partial \mathbf{x}}$ (chain rule)

sv4 $\frac{\partial \mathbf{u}^T\mathbf{v}}{\partial \mathbf{x}} = \mathbf{u}^T\frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}^T\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$ (product rule)

sv5 $\frac{\partial \mathbf{u}^T A\mathbf{v}}{\partial \mathbf{x}} = \mathbf{u}^T A\frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}^T A^T\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$, where A is not a function of $\mathbf{x}$ (product rule)

Properties of matrix derivatives cont.

Properties for scalar by matrix

sm1 $\frac{\partial au}{\partial X} = a\frac{\partial u}{\partial X}$, where a is not a function of x.

sm2 $\frac{\partial (u+v)}{\partial X} = \frac{\partial u}{\partial X} + \frac{\partial v}{\partial X}$

sm3 $\frac{\partial uv}{\partial X} = u\frac{\partial v}{\partial X} + v\frac{\partial u}{\partial X}$ (product rule)

sm4 $\frac{\partial f(g(u))}{\partial X} = \frac{\partial f(g)}{\partial g}\frac{\partial g(u)}{\partial u}\frac{\partial u}{\partial X}$ (chain rule)

Properties for vector by vector

vv1 $\frac{\partial (a\mathbf{u}+\mathbf{v})}{\partial \mathbf{x}} = a\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}}$, where a is not a function of x.

vv2 $\frac{\partial A\mathbf{u}}{\partial \mathbf{x}} = A\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$, where A is not a function of $\mathbf{x}$.

vv3 $\frac{\partial f(g(\mathbf{u}))}{\partial \mathbf{x}} = \frac{\partial f(g)}{\partial g}\frac{\partial g(u)}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$ (chain rule)

Linear regression

Problem

Given a set of n points (x_i, y_i) on a scatterplot, find the relationship between x and y: $\widehat{y_i} = \beta_0 + \beta_1 x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$.

- We can write linear regression in matrix form:

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 + \epsilon_1 \\ \beta_0 + \beta_1 x_2 + \epsilon_2 \\ \dots \\ \beta_0 + \beta_1 x_n + \epsilon_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{bmatrix}$$

- Rewrite as $Y_{n\times1} = X_{n\times2}\beta_{n\times1} + \epsilon_{n\times1}$, thus residuals are $\epsilon = Y - X\beta$. We would like to minimize sum of squared residuals $\epsilon^T\epsilon = (Y - X\beta)^T(Y - X\beta)$
- $\frac{\partial}{\partial\beta}(Y - X\beta)^T(Y - X\beta) = -2X^T(Y - X\beta) = 0$, thus $\beta = (X^TX)^{-1}X^TY$

Eigenvalue and eigenvector

Definition

Given a matrix $A \in \mathbb{R}^{n\times n}$, and non-zero column vector v, if $Av = \lambda v$, then λ is an eigenvalue of A and v is an eigenvector corresponding to eigenvalue λ.

- Transformation: a matrix A acts on vectors x like a function does, with input x and output Ax. Eigenvectors are vectors for which Ax is parallel to x. In other words: $Ax = \lambda x$.
- If M is Hermitian, then all the eigenvalues of M are real. Note that a real symmetric matrix is always Hermitian, i.e., $A^T = A$.
- The adjacency matrix of an undirected graph is symmetric, and this implies that its eigenvalues are all real.

How to find eigenvalues and eigenvectors

Method 1

$Av = \lambda v \Leftrightarrow (A - \lambda I)v = 0$. For non-zero vector v, which is equivalent to $det(A - \lambda I) = 0$, where the equation is called characteristic equation .

- For example, $A = \begin{bmatrix} 3 & -2 \\ 1 & 0 \end{bmatrix}$, we have $det(A - \lambda I) = \lambda^2 - 3\lambda + 2 = 0$. We find eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 2$. We can find eigenvector via solving the linear equation $(A - \lambda_i I)v = 0$. That is $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$.
- This approach is not scalable.

Method 2: the power method

- The power method is an iterative algorithm which has the following basic form for generating a single eigenvalue and eigenvector of A.
  - Initial nonzero vector $x^{(0)} \in \mathbb{R}^n$, such that $\|x^{(0)}\| = 1$.
  - For $k = 0, 1, \cdots$, we let $x^{(k+1)}$ be a nonzero multiple of $Ax^{(k)}$, typically $x^{(k+1)} = Ax^{(k)} / \|Ax^{(k)}\|$.

The power method

Analysis

- Why is it correct? Suppose $x^{(0)}$ is in the subspace generated the eigenvectors, i.e., $x^{(0)} = \sum_{i=1}^{n} c_i v_i$ with $c_1 \neq 0$. Then $x^{(k)}$ converges to the dominant eigenvector v_1 because

$$\lim_{k\to\infty} A^k x^{(0)} = \lim_{k\to\infty} \sum_{i=1}^{n} c_i A^k v_i = \lim_{k\to\infty} \sum_{i=1}^{n} c_i \lambda_i^k v_i$$
$$= \lim_{k\to\infty} c_1 \lambda_1^k \Big[v_1 + \sum_{i=2}^{n} \frac{c_i}{c_1} \frac{\lambda_i}{\lambda_1}^k v_i \Big] = c_1 \lambda_1^k v_1$$

- It will find only one eigenvalue (the one with the greatest absolute value), and it may converge only slowly.
- It can fail if there is not a single largest eigenvalue, i.e., $\lambda_1 = \lambda_2$.

The power method cont.

Extension

- Inverse power method: it operates with A^{-1} rather than A since the eigenvalues of A^{-1} are $\frac{1}{\lambda_i}$. It gives a way of finding the smallest (in absolute value) eigenvalue of a matrix.
- Spectral shift: using the fact that the eigenvalues of $A - \alpha I$ are $\lambda_i - \alpha$. If we find the largest eigenvalue λ_1, we can find the largest in absolute value of $\lambda_i - \lambda_1$. However, it is not clear how it could be implemented in general to find all the eigenvalues of matrix A.
- Symmetric power method: we can repeatedly find all eigenvalues of a symmetric matrix A with distinct eigenvalues because

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^T,$$

$$A^{(i)} = A^{(i-1)} - \lambda_i v_i v_i^T$$

Diagonalization

Definition of similar matrices

If matrices $A, B \in \mathbb{R}^{n\times n}$ are said to be similar if there is an invertible $n \times n$ matrix P such that $A = PBP^{-1}$.

- Theorem: If $n \times n$ matrices are similar, then they have the same characteristic of polynomial and hence the same enigenvalues (with the same multiplicities).

Definition of diagonalizable

A square matrix A is said to be diagonalizable if it is similar to a diagonal matrix, i.e., there exists an invertible matrix P and a diagonal matrix D such that $A = PDP^{-1}$.

- Why is this useful? If A is diagonalizable, then $A^3 = PD^3P^{-1}$.
- How to find diagonal matrix? If $v_1, \cdots, v_n$ are linearly independent eigenvectors of A and λ_i are their corresponding eignevalues, then $A = PDP^{-1}$, where $P = [v_1 \ \cdots \ v_n]$ and $D = Diag(\lambda_1, \cdots, \lambda_n)$.

Diagonalization

Analysis

$$\begin{bmatrix} 1.5 & 0.5 \\ 0.5 & 1.5 \end{bmatrix} \text{ with } U = \begin{bmatrix} 0.85 & -0.53 \\ 0.53 & 0.85 \end{bmatrix} \text{ and } 1.81, 0.69$$

Method 2: the power method

- The power method is an iterative algorithm which has the following basic form for generating a single eigenvalue and eigenvector of A.
  - Initial nonzero vector $x^{(0)} \in \mathbb{R}^n$, such that $\|x^{(0)}\| = 1$.
  - For $k = 0, 1, \cdots$, we let $x^{(k+1)}$ be a nonzero multiple of $Ax^{(k)}$, typically $x^{(k+1)} = Ax^{(k)} / \|Ax^{(k)}\|$.
- This approach is not scalable.

Take-home messages

- Markov chain
- Graphical model
  - Directed model
  - Undirected model
- Tail bounds