Introduction: Linear Algebra

Introduction to Artificial Intelligence with Mathematics
Lecture Notes

Ganguk Hwang

Department of Mathematical Sciences
KAIST

Metric Space

Metrics generalize the notion of distance from Euclidean space.

Definition 1

A metric on a set S *is a function* $d : S \times S \longrightarrow \mathbb{R}$ *that satisfies*

- $d(x, y) \geq 0$ *and the equality holds if and only if* $x = y$
- $d(x, y) = d(y, x)$
- $d(x, y) \leq d(x, z) + d(z, y)$ *(the triangle inequality)*

for all $x, y, z \in S$*.*

Normed Space

Norms generalize the notion of length from Euclidean space.

Definition 2

A norm on a real vector space V is a function $||\cdot|| : V \longrightarrow \mathbb{R}$ that satisfies

- $||\mathbf{x}|| \geq 0$ *and the equality holds if and only if* $\mathbf{x} = \mathbf{0}$
- $||\alpha\mathbf{x}|| = |\alpha| \; ||\mathbf{x}||$
- $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$ *(triangle inequalty)*

for all $\mathbf{x}, \mathbf{y} \in V$ and all $\alpha \in \mathbb{R}$.

A vector space with a norm is called a normed vector space or a normed space.

Examples of norms

Let $\mathbf{x} = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$.

- $||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|$
- $||\mathbf{x}||_2 = \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}$
- $||\mathbf{x}||_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$
- $||\mathbf{x}||_\infty = \max_{1 \le i \le n} |x_i|$

Inner Product Spaces

Definition 3

An inner product on a real vector space V is a function $< \cdot, \cdot >: V \times V \longrightarrow \mathbb{R}$ that satisfies

- $< \mathbf{x}, \mathbf{x} > \geq 0$ *and the equality holds if and only if* $\mathbf{x} = \mathbf{0}$
- $< \mathbf{x} + \mathbf{y}, \mathbf{z} > = < \mathbf{x}, \mathbf{y} > + < \mathbf{x}, \mathbf{z} >$ *and* $< \alpha\mathbf{x}, \mathbf{y} > = \alpha < \mathbf{x}, \mathbf{y} >$
- $< \mathbf{x}, \mathbf{y} > = < \mathbf{y}, \mathbf{x} >$

for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ *and all* $\alpha \in \mathbb{R}$.

A vector space with an inner product is called an inner product space. Note that an inner product on V induces a norm on V as follows:

$$||\mathbf{x}|| = \sqrt{< \mathbf{x}, \mathbf{x} >}.$$

- Two vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if $<\mathbf{x}, \mathbf{y}>= 0$.
- If two orthogonal vectors $\mathbf{x}$ and $\mathbf{y}$ satisfy $||\mathbf{x}|| = ||\mathbf{y}|| = 1$, then they are called orthonormal.
- The standard inner product in $\mathbb{R}^n$ is

$$<\mathbf{x}, \mathbf{y}>= \sum_{i=1}^{n} x_i y_i$$

where $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and $\mathbf{y} = (y_1, y_2, \cdots, y_n)$.

Pythagorean Theorem

If two vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal,

$$||\mathbf{x}+\mathbf{y}||^2 = ||\mathbf{x}||^2 + ||\mathbf{y}||^2.$$

Cauchy-Schwarz Inequality

$$| < \mathbf{x}, \mathbf{y} > |^2 \leq ||\mathbf{x}||^2 \; ||\mathbf{y}||^2$$

c.f. $0 \leq ||\mathbf{x} - \lambda \mathbf{y}||^2$ with $\lambda = \frac{<\mathbf{x},\mathbf{y}>}{||\mathbf{y}||^2}$

Orthogonal complements and projections

Let V be an inner product space and S be a finite-dimensional subspace of V. Then $\mathbf{v} \in V$ can be written uniquely as

$$\mathbf{v} = \mathbf{v}_S + \mathbf{v}_\perp$$

where $\mathbf{v}_S \in S$ and $\mathbf{v}_\perp \in S^\perp$.

This implies that

$$V = S \oplus S^\perp.$$

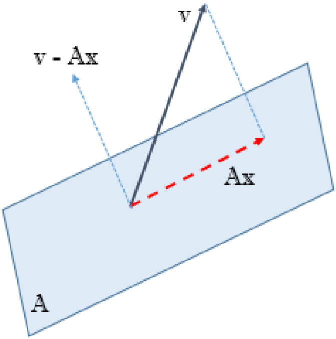
Moreover, we can define an orthogonal projection $P_S \colon V \longrightarrow S$ defined by

$$P_S(\mathbf{v}) = \mathbf{v}_S.$$

Let $\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_k$ form a basis of a subspace S of $\mathbb{R}^n$, and $\mathbf{A}$ denote the $n \times k$ matrix with these vectors as columns, then the orthogonal projection P_S is given by

$$P_S(\mathbf{v}) = \mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top \mathbf{v}.$$

c.f. consider $\mathbf{A}^\top(\mathbf{v} - \mathbf{A}\mathbf{x}) = \mathbf{0}$ and find $\mathbf{x}$.



Eigenvalues and Eigenvectors

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a square matrix. We say that a nonzero vector $\mathbf{x} \in \mathbb{R}^n$ is an eigenvector of $\mathbf{A}$ corresponding to eigenvalue λ if

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x}.$$

The following are the properties of eigenvalues and eigenvectors.

- For any $\gamma \in \mathbb{R}$, $\mathbf{x}$ is an eigenvector of $\mathbf{A} + \gamma \mathbf{I}$ with eigenvalue $\lambda + \gamma$.
- If $\mathbf{A}$ is invertible, then $\mathbf{x}$ is an eigenvector of $\mathbf{A}^{-1}$ with eigenvalue λ^{-1}.
- $\mathbf{A}^k \mathbf{x} = \lambda^k \mathbf{x}$ for any $k \in \{0, 1, 2, \cdots\}$.

Trace

The trace of an $n \times n$ square matrix $\mathbf{A} = (a_{ij})$ is the sum of its diagonal elements, i.e.,

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

The nice properties of the trace are

- $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$
- $\text{tr}(\alpha\mathbf{A}) = \alpha\text{tr}(\mathbf{A})$
- $\text{tr}(\mathbf{A}^\top) = \text{tr}(\mathbf{A})$
- $\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{BCA}) = \text{tr}(\mathbf{CAB})$ (invariant under *cyclic* permutations)
- The trace of $\mathbf{A}$ is equal to the sum of its eigenvalues (repeated according to multiplicity).

Determinant

We skip the definitionof the determinant of a squre matrix, but we need its properties.

- $\det(\mathbf{I}) = 1$
- $\det(\mathbf{A}^\top) = \det(\mathbf{A})$
- $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$
- $\det(\mathbf{A}^{-1}) = \det(\mathbf{A})^{-1}$
- $\det(\alpha\mathbf{A}) = \alpha^n\det(\mathbf{A})$
- The determinant of a matrix is equal to the product of its eigenvalues (repeated according to multiplicity).

Symmetric Matrices

A square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is said to be symmetric if $\mathbf{A}^\top = \mathbf{A}$.

- If $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric, then there exists an orthonormal basis for $\mathbb{R}^n$ consisting of eigenvectors of $\mathbf{A}$.
- Let $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n$ denote the orthonormal basis and $\lambda_1, \lambda_2, \cdots, \lambda_n$ be their eigenvalues. Let $\mathbf{U}$ be the matix with $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n$ as its columns, and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$. Then

$$\mathbf{A}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}, \text{ equivalently } \mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top.$$

Positive Semi-definite Matrices

A symmetric matrix $\mathbf{A}$ is positive semi-definite (positive definite, resp.) if for every non-zero vector $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0 (> 0$, resp.).

- A symmetric matrix is positive semi-definite if and only if all of its eigenvalues are nonnegative, and positive definite if and only if its eigenvalues are positive.
- For $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{A}^\top \mathbf{A}$ is positive semi-definite. If null$(\mathbf{A}) = \{\mathbf{0}\}$, then $\mathbf{A}^\top \mathbf{A}$ is positive definite.
- If $\mathbf{A}$ is positive semi-definite and $\epsilon > 0$, then $\mathbf{A} + \epsilon \mathbf{I}$ is positive definite.
- If $\mathbf{A}$ is positive semi-definite, it satisfies $\mathbf{A} = \mathbf{A}^{\frac{1}{2}} \mathbf{A}^{\frac{1}{2}}$ where $\mathbf{A}^{\frac{1}{2}} = \mathbf{U} \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{U}^\top$ and $\mathbf{\Lambda}^{\frac{1}{2}} = \mathrm{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_n})$.

Singular Value Decomposition

Sigular value decomposition is one of important tools in linear algebra and machine learning. Its strength stems from the fact that every matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ has a singular value decomposion.

For $\mathbf{A} \in \mathbb{R}^{m \times n}$, it can be decomposed as

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

where $\mathbf{U} \in \mathbb{R}^{m \times m}$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$ are orthogonal matrices and $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ is a diagonal matrix with the sigular values of $\mathbf{A}$, denoted by σ_i on its diagonal.

When the rank of $\mathbf{A}$ is r, only the first r singular values are nonzero in the increasing order, i.e.,

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq \sigma_{r+1} = \cdots = \sigma_{\min\{m,n\}} = 0.$$

Another way to represent the singular value decomposition is

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are the i-th column vectors of $\mathbf{U}$ and $\mathbf{V}$, called the left singular vectors and the right singular vectors of $\mathbf{A}$, respectively.

It is easy to see that $\mathbf{A}^\top\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$ are positive semi-definite, so their eigenvalues are nonnegative.

Observing

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

we see that the columns of $\mathbf{V}$ are eigenvectors of $\mathbf{A}^\top\mathbf{A}$ and the columns of $\mathbf{U}$ are eigenvector of $\mathbf{A}\mathbf{A}^\top$.

Note that, even though two matrices $\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ are not necessarily the same size, they are diagonal with σ_i^2 and some zeros. That is, the singular values of $\mathbf{A}$ are the square roots of the eigenvalues of $\mathbf{A}^\top\mathbf{A}$ or $\mathbf{A}\mathbf{A}^\top$.

Matrix Differentiation

We use the following convention.

$$\mathbf{x} = (x_1, x_2, \cdots, x_n)^\top, \mathbf{y} = (y_1, y_2, \cdots, y_m)^\top$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}.$$

- If $\mathbf{y} = \mathbf{A}\mathbf{x}$ and $\mathbf{A}$ is independent of $\mathbf{x}$, then

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}.$$

- If $\mathbf{y} = \mathbf{A}\mathbf{x}$, $\mathbf{x} = (x_1(\mathbf{z}), \cdots, x_n(\mathbf{z}))$ for some $\mathbf{z}$, and $\mathbf{A}$ is independent of $\mathbf{z}$, then

$$\frac{\partial \mathbf{y}}{\partial \mathbf{z}} = \mathbf{A}\frac{\partial \mathbf{x}}{\partial \mathbf{z}}.$$

- If $\alpha = \mathbf{y}^\top \mathbf{A}\mathbf{x}$ and $\mathbf{A}$ is independent of $\mathbf{x}$ and $\mathbf{y}$, then

$$\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{y}^\top \mathbf{A}, \frac{\partial \alpha}{\partial \mathbf{y}} = \mathbf{x}^\top \mathbf{A}^\top.$$

- If $\alpha = \mathbf{x}^\top \mathbf{A}\mathbf{x}$ and $\mathbf{A}$ is independent of $\mathbf{x}$, then

$$\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top).$$

- If $\alpha = \mathbf{y}^\top \mathbf{x}$, $\mathbf{x} = (x_1(\mathbf{z}), \cdots, x_n(\mathbf{z}))^\top$, and $\mathbf{y} = (y_1(\mathbf{z}), \cdots, y_n(\mathbf{z}))^\top$ for some $\mathbf{z}$, then
$$\frac{\partial \alpha}{\partial \mathbf{z}} = \mathbf{x}^\top \frac{\partial \mathbf{y}}{\partial \mathbf{z}} + \mathbf{y}^\top \frac{\partial \mathbf{x}}{\partial \mathbf{z}}.$$
- If $\alpha = \mathbf{y}^\top \mathbf{A}\mathbf{x}$, $\mathbf{x} = (x_1(\mathbf{z}), \cdots, x_n(\mathbf{z}))^\top$, $\mathbf{y} = (y_1(\mathbf{z}), \cdots, y_n(\mathbf{z}))^\top$ for some $\mathbf{z}$ and $\mathbf{A}$ is independent of $\mathbf{z}$, then
$$\frac{\partial \alpha}{\partial \mathbf{z}} = \mathbf{y}^\top \mathbf{A} \frac{\partial \mathbf{x}}{\partial \mathbf{z}} + \mathbf{x}^\top \mathbf{A}^\top \frac{\partial \mathbf{y}}{\partial \mathbf{z}}.$$
- If $\alpha = \mathbf{x}^\top \mathbf{A}\mathbf{x}$, $\mathbf{x} = (x_1(\mathbf{z}), \cdots, x_n(\mathbf{z}))^\top$ for some $\mathbf{z}$ and $\mathbf{A}$ is independent of $\mathbf{z}$, then
$$\frac{\partial \alpha}{\partial \mathbf{z}} = \mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top) \frac{\partial \mathbf{x}}{\partial \mathbf{z}}.$$