Mathematics for Machine Learning

Rithvik Rao

MLH Fellowship

26 June 2020

Based on these notes from UC Berkeley CS 189

Table of Contents

Vector Spaces

Definition (Vector Space)

A *vector space* V is a set of vectors for which addition and scalar multiplication are defined. V satisfies:

- *(Additive identity)* $\mathbf{x} + \mathbf{0} = \mathbf{x}$ $\quad \forall \mathbf{x} \in V$
- *(Additive inverse)* $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$ $\quad \forall \mathbf{x} \in V$
- *(Multiplicative identity)* $1\mathbf{x} = \mathbf{x}$ $\quad \forall \mathbf{x} \in V$
- *(Commutativity)* $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ $\quad \forall \mathbf{x}, \mathbf{y} \in V$
- *(Associativity)*
  $$(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z}) \qquad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in V$$
  $$\alpha(\beta\mathbf{x}) = (\alpha\beta\mathbf{x}) \qquad \forall \mathbf{x} \in V, \ \alpha, \beta \in \mathbb{R}$$
- *(Distributivity)*
  $$\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y} \qquad \forall \mathbf{x}, \mathbf{y} \in V, \ \alpha \in \mathbb{R}$$
  $$(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x} \qquad \forall \mathbf{x} \in V, \ \alpha, \beta \in \mathbb{R}$$

Vector Spaces

Definition (Linear Independence)

Vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are *linearly independent* if $\alpha_1 \mathbf{v}_1 + \cdots + \alpha_n \mathbf{v}_n = \mathbf{0}$ implies $\alpha_1 = \cdots = \alpha_n = 0$.

Vector Spaces

Definition (Linear Independence)

Vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are *linearly independent* if $\alpha_1 \mathbf{v}_1 + \cdots + \alpha_n \mathbf{v}_n = \mathbf{0}$ implies $\alpha_1 = \cdots = \alpha_n = 0$.

Definition (Span)

The *span* of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is the set of all vectors that can be produced by a linear combination of these vectors.
$\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} = \{\mathbf{v} \in V : \exists \alpha_1, \ldots, \alpha_n \text{ s.t. } \alpha_1 \mathbf{v}_1 + \cdots + \alpha_n \mathbf{v}_n = \mathbf{v}\}$

Vector Spaces

Definition (Linear Independence)

Vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are *linearly independent* if $\alpha_1 \mathbf{v}_1 + \cdots + \alpha_n \mathbf{v}_n = \mathbf{0}$ implies $\alpha_1 = \cdots = \alpha_n = 0$.

Definition (Span)

The *span* of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is the set of all vectors that can be produced by a linear combination of these vectors.
$\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} = \{\mathbf{v} \in V : \exists \alpha_1, \ldots, \alpha_n \text{ s.t. } \alpha_1 \mathbf{v}_1 + \cdots + \alpha_n \mathbf{v}_n = \mathbf{v}\}$

Definition (Basis)

Linearly independent vectors which span the vector space form a *basis*.

Vector Spaces

Definition (Linear Independence)

Vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are *linearly independent* if $\alpha_1 \mathbf{v}_1 + \cdots + \alpha_n \mathbf{v}_n = \mathbf{0}$ implies $\alpha_1 = \cdots = \alpha_n = 0$.

Definition (Span)

The *span* of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is the set of all vectors that can be produced by a linear combination of these vectors.
$\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} = \{\mathbf{v} \in V : \exists \alpha_1, \ldots, \alpha_n \text{ s.t. } \alpha_1 \mathbf{v}_1 + \cdots + \alpha_n \mathbf{v}_n = \mathbf{v}\}$

Definition (Basis)

Linearly independent vectors which span the vector space form a *basis*.

Definition (Dimension)

The number of vectors in a basis for a finite-dimensional vector space. Denoted $\dim V$.

Euclidean Space

$\mathbb{R}^n$, where vectors take the form $\mathbf{x} = (x_1, \dots, x_n)$, like points in n-dimensional space.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix} \qquad \alpha\mathbf{x} = \begin{bmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{bmatrix}$$

Subspaces

Definition (Subspace)

$S \subseteq V$ is a *subspace* of V if:

- $\mathbf{0} \in S$
- S is closed under addition: $\mathbf{x}, \mathbf{y} \in S \implies \mathbf{x} + \mathbf{y} \in S$.
- S is closed under scalar multiplication: $\mathbf{x} \in S \implies \alpha\mathbf{x} \in S, \ \forall \alpha \in \mathbb{R}$.

For any two subspaces of V, U and W:

$$\dim(U + W) = \dim U + \dim W - \dim(U \cap W)$$

Linear Maps

Definition (Linear Map)

A *linear map* is a function $T : V \to W$, where v and W are vector spaces, which satisifies

- $T(\mathbf{x} + \mathbf{y}) = T\mathbf{x} + T\mathbf{y} \qquad \forall \mathbf{x}, \mathbf{y} \in V$
- $T(\alpha \mathbf{x}) = \alpha T\mathbf{x} \qquad \forall \mathbf{x} \in V, \ \alpha \in \mathbb{R}$

A linear map is a **homomorphism**. If its inverse is also a linear map, then it is an **isomorphism**. An isomorphism between V and W implies they are **isomorphic**, or $V \cong W$.

Matrices of Linear Maps

Given finite-dimensional vector spaces V, W with bases $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and $\mathbf{w}_1, \ldots, \mathbf{w}_m$, and linear map $T : V \to W$, we define the matrix of T as having entries A_{ij}, where $i = 1, \ldots, m$ and $j = 1, \ldots, n$. The jth column of $\mathbf{A}$ consists of:

$$T\mathbf{v}_j = A_{1j}\mathbf{w}_1 + \cdots + A_{mj}\mathbf{w}_m$$

Every matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ induces a linear map $T : \mathbb{R}^n \to \mathbb{R}^m$ given by $T\mathbf{x} = \mathbf{A}\mathbf{x}$.

A matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ has a **transpose** $\mathbf{A}^\mathbf{T} \in \mathbb{R}^{n \times m}$, where $A^T_{ij} = A_{ji}\ \forall (i, j)$.

Null Space, Range

Let $T : V \to W$ be a linear map.

Definition (Null Space, or Kernel)

$$\text{null}(T) = \{\mathbf{v} \in V \mid T\mathbf{v} = \mathbf{0}\}$$

Definition (Range)

$$\text{range}(T) = \{\mathbf{w} \in W \mid \exists \mathbf{v} \in V \text{ s.t. } T\mathbf{v} = \mathbf{w}\}$$

The **column space** of a matrix is the span of its columns, while the **row space** is the span of its rows. The column space is the range of the linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ induced by $\mathbf{A}$, or range($\mathbf{A}$).
The dimension of the row space of $\mathbf{A}$ is called the **rank** of $\mathbf{A}$, and:

$$\text{rank}(\mathbf{A}) = \dim \text{range}(\mathbf{A})$$

Metric Spaces

Definition (Metric Space)

A *metric space* is any set together with a *metric* on that set, which generalizes the concept of *distance*. A metric on a set S is a function $d : S \times S \to \mathbb{R}$ which satisfies, for all $x, y, z \in S$:

- *(Weakly positive)* $d(x, y) \geq 0$, with equality iff $x = y$
- *(Invariant to ordering of points)* $d(x, y) = d(y, x)$
- *(Triangle inequality)* $d(x, z) \leq d(x, y) + d(y, z)$

Motivation: a sequence $\{x_n\} \subseteq S$ converges to the limit x if for all $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $d(x_n, x) < \epsilon$ for all $n \geq N$.

Normed Spaces

Definition (Normed Space)

A *normed space* is a vector space together with a *norm*, which generalizes the concept of *length*. A norm on a real vector space V is a function $\|\cdot\| : V \to \mathbb{R}$ that satisfies

- *(Weakly positive)* $\|\mathbf{x}| \geq 0$, with equality if and only if $\mathbf{x} = \mathbf{0}$
- *(Scalar multiplication)* $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$
- *(Triangle inequality)* $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

Any normed space is a metric space. (Why?)

$$\underbrace{\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}}_{\text{the } p\text{-norm, for } p \geq 1} \qquad \underbrace{\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|}_{\text{the infinity-norm}}$$

Inner Product Spaces

Definition (Inner Product Space)

An *inner product space* is a vector space endowed with an *inner product*, which is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ satisfying

- *(Weakly positive)* $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, with equality if and only if $\mathbf{x} = \mathbf{0}$
- *(Scalar multiplication, distributivity)* $\langle \alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z} \rangle = \alpha\langle \mathbf{x}, \mathbf{z} \rangle + \beta\langle \mathbf{y}, \mathbf{z} \rangle$
- *(Order-invariance)* $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$

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

Any inner product on V induces a norm: $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.
Two vectors $\mathbf{x}$ and $\mathbf{y}$ are **orthogonal** ($\mathbf{x} \perp \mathbf{y}$) if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$, and if $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$ also, then the vectors are **orthonormal**.
The standard inner product on $\mathbb{R}^n$, the **dot product**, is given by:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i = \mathbf{x}^T \mathbf{y}$$

Useful Results on Inner Product Spaces

Theorem (Pythagorean Theorem)

If $\mathbf{x} \perp \mathbf{y}$, *then* $\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$.

Theorem (Cauchy-Schwarz Inequality)

$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|$ *for all* $\mathbf{x}, \mathbf{y} \in V$.

Orthogonal Complements and Projections

If $S \subseteq V$, and V is an inner product space, then the **orthogonal complement** of S, $S^\perp$, is the set of all vectors in V that are orthogonal to every element of S:

$$S^\perp = \{\mathbf{v} \in V \mid \mathbf{v} \perp \mathbf{s},\ \forall \mathbf{s} \in S\}$$

Every $\mathbf{v} \in V$ can be written uniquely in the form $\mathbf{v} = \mathbf{v}_S + \mathbf{v}_\perp$, where $\mathbf{v}_S \in S$ and $\mathbf{v}_\perp \in S^\perp$.
I omit discussion of orthogonal projections here. Note that a **projection** is any linear map P that satisfies $P^2 = P$, and that, and that the **orthogonal projection** is used to find the closest point in S to a given $\mathbf{v} \in V$.

Table of Contents

Eigen-stuff

For any square $\mathbf{A} \in \mathbb{R}^{n \times n}$:

Definition (Eigen{vector, value})

$\mathbf{x} \in \mathbb{R}^n$ is a (right-hand) *eigenvector* of $\mathbf{A}$ with *eigenvalue* λ if:

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

Theorem (Useful Eigen-stuff)

- *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 $k \in \mathbb{Z}$ (where $\mathbf{A}^0 = \mathbf{I}$ by definition).*

(These are good exercises!)

Trace

Definition (Trace)

The *trace* of a square matrix is the sum of its diagonal entries, and conveniently, also the sum of its eigenvalues:

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} A_{ii} \qquad \mathrm{tr}(\mathbf{A}) = \sum_{i} \lambda_i(\mathbf{A})$$

Theorem (Trace Properties)

- $\mathrm{tr}(\mathbf{A} + \mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$
- $\mathrm{tr}(\alpha \mathbf{A}) = \alpha\, \mathrm{tr}(\mathbf{A})$
- $\mathrm{tr}(\mathbf{A}^T) = \mathrm{tr}(\mathbf{A})$
- $\mathrm{tr}(\mathbf{ABCD}) = \mathrm{tr}(\mathbf{BCDA}) = \mathrm{tr}(\mathbf{CDAB}) = \mathrm{tr}(\mathbf{DABC})$

Determinant

Definition (Determinant)

Many things ... volume scaling factor of linear transformation, volume of n-dimensional parallelepiped spanned by rows or columns. But always a scalar value. Computed recursively using the *minor expansion formula*, or:

$$\det(\mathbf{A}) = \prod_i \lambda_i(\mathbf{A})$$

Theorem (Determinant Properties)

- $\det(\mathbf{I}) = 1$
- $\det(\mathbf{A}^T) = \det(\mathbf{A})$
- $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$
- $\det(\mathbf{A}^{-1}) = \det(\mathbf{A})^1$
- $\det(\alpha\mathbf{A}) = \alpha^n \det(\mathbf{A})$

Orthogonal Matrices

Definition (Orthogonal Matrix)

$\mathbf{Q} \in \mathbb{R}^{n \times n}$ is *orthogonal* if its columns are pairwise orthonormal. This implies:

$$\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^{\mathbf{T}} = \mathbf{I} \iff \mathbf{Q}^T = \mathbf{Q}^{-1}$$

Orthogonal matrices preserve inner products and 2-norms.

Symmetric Matrices

Definition (Symmetric Matrix)

$\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric if it equals its own transpose, i.e. $\mathbf{A} = \mathbf{A}^T$.

Theorem (Spectral Theorem)

If $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric, there exists an orthonormal basis for $\mathbb{R}^n$ consisting of eigenvectors of $\mathbf{A}$.

Definition (Spectral Decomposition / Eigendecomposition)

Let the orthonormal basis of eigenvectors be $\mathbf{q}_1, \ldots, \mathbf{q}_n$, with eigenvalues $\lambda_1, \ldots, \lambda_n$. Let $\mathbf{Q}$ be an orthogonal matrix with columns $\mathbf{q}_1, \ldots, \mathbf{q}_n$, and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Then:

$$\mathbf{A}\mathbf{q}_i = \lambda_i \mathbf{q}_i,\ \forall i \implies \mathbf{AQ} = \mathbf{Q\Lambda} \implies \mathbf{A} = \mathbf{Q\Lambda Q^T}$$

Rayleigh Quotients

For symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is called a **quadratic form**.

Definition (Rayleigh Quotient)

$$R_{\mathbf{A}}(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$$

Connects the quadratic form of a symmetric matrix with its eigenvalues.

Theorem (Properties of Rayleigh Quotient)

- *Scale invariance: for any vector* $\mathbf{x} \neq \mathbf{0}$, *any scalar* $\alpha \neq 0$, $R_{\mathbf{A}}(\mathbf{x}) = R_{\mathbf{A}}(\alpha\mathbf{x})$.
- *If* $\mathbf{x}$ *is an eigenvector of* $\mathbf{A}$ *with eigenvalue* λ, *then* $R_{\mathbf{A}}(\mathbf{x}) = \lambda$.

Theorem (Min-max Theorem)

For all $\mathbf{x} \neq \mathbf{0}$, $\lambda_{\min}(\mathbf{A}) \leq R_{\mathbf{A}}(\mathbf{x}) \leq \lambda_{\max}(\mathbf{A})$.

Positive (Semi-)Definite Matrices

Definition (Positive Semi-Definite, or $\mathbf{A} \succeq 0$)

Symmetric matrix $\mathbf{A}$ is *positive semi-definite* if for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$.

Definition (Positive Definite, or $\mathbf{A} \succ 0$)

Symmetric matrix $\mathbf{A}$ is *positive definite* if for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$.

Theorem (Results for Positive (Semi-)Definite Matrices)

- *Symmetric $\mathbf{A}$ is positive semi-definite iff all eigenvalues are nonnegative, and positive definite iff all eigenvalues positive.*
- *For $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{A}^T\mathbf{A}$ is positive semi-definite. If* $\text{null}(\mathbf{A}) = \{\mathbf{0}\}$, $\mathbf{A}^T\mathbf{A}$ *is positive definite.*
- *If $\mathbf{A}$ is positive semi-definite and $\epsilon > 0$, $\mathbf{A} + \epsilon\mathbf{I}$ is positive definite.*

Table of Contents

Singular Value Decomposition

Definition (Singular Value Decomposition)

Any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ has an SVD: $\mathbf{A} = \mathbf{U\Sigma V}^\top$. $\mathbf{U} \in \mathbb{R}^{m \times m}$ 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 *singular values* of $\mathbf{A}$ (σ_i) on its diagonal.

Used for ordinary least-squares regression, among many other things which require pseudo-inverses!

Fundamental Theorem of Linear Algebra

Theorem (Fundamental-ish Theorem of Linear Algebra)

If $\mathbf{A} \in \mathbb{R}^{m \times n}$*, then*

- $\text{null}(\mathbf{A}) = \text{range}(\mathbf{A}^\top)^\perp$
- $\text{null}(\mathbf{A}) \oplus \text{range}(\mathbf{A}^\top) = \mathbb{R}^n$
- $\underbrace{\dim \text{range}(\mathbf{A})}_{\text{rank}(\mathbf{A}) = r} + \dim \text{null}(\mathbf{A}) = n$
- *If* $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ *is the SVD of* $\mathbf{A}$*, the columns of* $\mathbf{U}$ *and* $\mathbf{V}$ *form orthonormal bases for the four "fundamental subspaces" of* $\mathbf{A}$*:*

Subspace	*Columns*
$\text{range}(\mathbf{A})$	*The first* r *columns of* $\mathbf{U}$
$\text{range}(\mathbf{A}^\top)$	*The first* r *columns of* $\mathbf{V}$
$\text{null}(\mathbf{A}^\top)$	*The last* $m - r$ *columns of* $\mathbf{U}$
$\text{null}(\mathbf{A})$	*The last* $n - r$ *columns of* $\mathbf{V}$

Low-Rank Approximation

Given some matrix, we are sometimes interested in finding another matrix of the same dimension but lower rank that is as close as possible to the original matrix.

Theorem (Eckart-Young-Mirsky Theorem)

Let $\|\cdot\|$ be a unitary invariant matrix norm. Suppose $\mathbf{A} \in \mathbb{R}^{m\times n}$, where $m \geq n$, has SVD $\mathbf{A} = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. Then the best rank-k approximation to A, where $k \leq \mathrm{rank}(\mathbf{A})$, is given by:

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

in the sense that:

$$\|\mathbf{A} - \mathbf{A}_k\| \leq \|\mathbf{A} - \tilde{\mathbf{A}}\|$$

for any $\tilde{\mathbf{A}} \in \mathbb{R}^{m\times n}$ with $\mathrm{rank}(\tilde{\mathbf{A}}) \leq k$.

Moore-Penrose Pseudoinverse

$\mathbf{A} \in \mathbb{R}^{m \times n}$ is only invertible if $m = n$.

Definition (Moore-Penrose Pseudoinverse)

If $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\top}$, then the *Moore-Penrose pseudoinverse* $\mathbf{A}^{+} \in \mathbb{R}^{n \times m}$ is given by:

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

It satisfies:

- $\mathbf{A}\mathbf{A}^{+}\mathbf{A} = \mathbf{A}$
- $\mathbf{A}^{+}\mathbf{A}\mathbf{A}^{+} = \mathbf{A}^{+}$
- $\mathbf{A}\mathbf{A}^{+}$ is symmetric
- $\mathbf{A}^{+}\mathbf{A}$ is symmetric