Optimization Techniques for Machine Learning

AMLZC326 · #01 Analytic Geometry

Anshid Aboobacker

Motivation

- Mathematics is the language of Machine Learning.
- You have already gone through
  - Linear Algebra
  - Calculus
  - Discrete Maths
  - Probability & Statistics
- This course will take you through **Optimization**

Learning Objectives

Before optimizing functions, we must understand the space itself.
In this session we will:

- Recapitulation of vector spaces and bases
- Work with norms and inner products
- Interpret angles and orthogonality
- Recognize orthonormal bases and matrices

Table of Contents

Vector Spaces

A vector space $(V, +, \cdot)$ over real numbers is a set V with two operations: Vector addition $+ : V \times V \to V$, and Scalar multiplication $\cdot : \mathbb{R} \times V \to V$, which satisfies the following:

1. $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}$ (Commutativity)
2. $(\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{z} = \boldsymbol{x} + (\boldsymbol{y} + \boldsymbol{z})$ (Associativity)
3. $\boldsymbol{x} + \boldsymbol{0} = \boldsymbol{x}$ (Existence of zero element)
4. $\boldsymbol{x} + (-\boldsymbol{x}) = \boldsymbol{0}$ (Existence of additive inverse)
5. $a \cdot (\boldsymbol{x} + \boldsymbol{y}) = a \cdot \boldsymbol{x} + a \cdot \boldsymbol{y}$
6. $(a + b) \cdot \boldsymbol{x} = a \cdot \boldsymbol{x} + b \cdot \boldsymbol{x}$
7. $(ab) \cdot \boldsymbol{x} = a \cdot (b \cdot \boldsymbol{x})$
8. $1 \cdot \boldsymbol{x} = \boldsymbol{x}$

Examples

- $\mathbb{R}^n$
- Polynomials
  $P_n = \{p(x) | p(x) \text{ is a polynomial of degree } \leq n\}$
- Matrices $\mathbb{R}^{m \times n}$

Basis of a Vector Space

A set $B = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is a basis of V if:

- B is a *spanning* set: Every $\boldsymbol{x} \in V$ can be written as

$$\boldsymbol{x} = a_1 \boldsymbol{v}_1 + \cdots + a_n \boldsymbol{v}_n$$

- B is a *linearly independent* set:

$$a_1 \boldsymbol{v}_1 + \cdots + a_n \boldsymbol{v}_n = \boldsymbol{0} \;\Rightarrow\; a_1 = \cdots = a_n = 0$$

Example: Standard basis for $\mathbb{R}^3$

$$[1, 0, 0]^T,\ [0, 1, 0]^T,\ [0, 0, 1]^T$$

Basis is NOT Unique

In $\mathbb{R}^2$, both of the following sets are bases:

- $B_1 = \{[1, 0]^T, [0, 1]^T\}$
- $B_2 = \{[1, 1]^T, [1, -1]^T\}$

Dimension

The dimension of V, written $\dim(V)$, is the number of vectors in any basis of V.

Examples:

- $\dim(\mathbb{R}^n) = n$
- $\dim(P_2) = 3$

Norms

- A norm on a vector space is a function $\|.\| : V \to \mathbb{R}$, $\boldsymbol{x} \to \|\boldsymbol{x}\|$ which assigns to each vector $\boldsymbol{x}$ a length $\|\boldsymbol{x}\|$ such that for all $\lambda \in \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in V$ the following properties hold:
  - Absolutely homogeneous: $\|\lambda \boldsymbol{x}\| = |\lambda| \|\boldsymbol{x}\|$
  - Triangle inequality: $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$
  - Positive definite: $\|\boldsymbol{x}\| \geq 0$ and $\|\boldsymbol{x}\| = 0 \implies \boldsymbol{x} = 0$

Examples of Norms

Manhattan norm : $\|\boldsymbol{x}\|_1 = \sum_{i=1}^{i=n} |x_i|$

Euclidean norm : $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{i=n} x_i^2}$.

Spectral norm: For a matrix $A \in \mathbb{R}^{n \times m}$,

$$\|A\|_2 = \max_{\boldsymbol{x} \neq 0} \frac{\|A\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \sigma_{\max}(A)$$

where $\sigma_{\max}(A)$ is the largest singular value of A.

Table of Contents

Inner products

- Dot product in $\mathbb{R}^n$ is given by $\boldsymbol{x}^T\boldsymbol{y} = \sum_{i=1}^{i=n} x_i y_i$
- A bilinear mapping Ω is a mapping with two arguments and is linear in both arguments: Let V be a vector space such that $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in V$, and let $\lambda, \psi \in \mathbb{R}$. Then we have $\Omega(\lambda\boldsymbol{x} + \psi\boldsymbol{y}, \boldsymbol{z}) = \lambda\Omega(\boldsymbol{x}, \boldsymbol{z}) + \psi\Omega(\boldsymbol{y}, \boldsymbol{z})$, and $\Omega(\boldsymbol{x}, \lambda\boldsymbol{y} + \psi\boldsymbol{z}) = \lambda\Omega(\boldsymbol{x}, \boldsymbol{y}) + \psi\Omega(\boldsymbol{x}, \boldsymbol{z})$.
- Let V be a vector space and $\Omega : V \times V \to \mathbb{R}$ be a bilinear mapping that takes two vectors as arguments and returns a real number. Then Ω is called symmetric if $\Omega(\boldsymbol{x}, \boldsymbol{y}) = \Omega(\boldsymbol{y}, \boldsymbol{x})$. Also Ω is called positive-definite if $\forall x \in V \setminus \{0\}, \;\; \Omega(\boldsymbol{x}, \boldsymbol{x}) > 0$ and $\Omega(\boldsymbol{0}, \boldsymbol{0}) = 0$.

Inner products

- A positive-definite, symmetric bilinear mapping $\Omega : V \times V \to \mathbb{R}$ is called an inner product. To denote an inner product on V we generally write $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$.
- The pair $(V, \langle ., . \rangle)$ is called an inner product space.
- Next we introduce the concept of symmetric, positive-definite matrices and show we can express an inner product using such matrices.
- We recall that in a vector space V any vector $\boldsymbol{x}$ can be written as linear combination of the basis vectors. We use this to express an inner product in terms of a matrix.

Symmetric, positive-definite matrices

- A matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is *symmetric* if $\boldsymbol{A} = \boldsymbol{A}^T$.
- An $n \times n$ symmetric real matrix $\boldsymbol{A}$ is said to be *positive-definite* if $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} > 0$ for all non-zero $\boldsymbol{x} \in \mathbb{R}^n$.
- For a real-valued, finite-dimensional vector space V and an ordered basis B of V, it holds that $\langle ., . \rangle : V \times V \to \mathbb{R}$ is an inner product if and only if there exists a symmetric, positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \hat{\boldsymbol{x}}^T \boldsymbol{A} \hat{\boldsymbol{y}}$.

Lengths and distances

- Inner products and norms are closely related in the sense that any inner product induces a norm: $\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$
- Not every norm is induced by an inner product, for example the Manhattan norm.
- For an inner product vector space $(V, \langle ., . \rangle)$, the induced norm $\|.\|$ satisfies the Cauchy-Schwarz inequality: $|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|$.

Metric space

- Consider an inner product space $(V, \langle ., . \rangle)$. Define $d(\boldsymbol{x}, \boldsymbol{y})$ the distance between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ to be $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\| = \sqrt{\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y} \rangle}$.
- If we use the dot product as the inner product, then the distance is called the Euclidean distance.
- The mapping $d : V \times V \to \mathbb{R}$ is called a metric.

Properties of a metric space

A metric d has the following properties:

- d is positive-definite which means $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0 \;\; \forall \boldsymbol{x}, \boldsymbol{y} \in V$. $d(\boldsymbol{x}, \boldsymbol{y}) = 0 \implies \boldsymbol{x} = \boldsymbol{y}$.
- d is symmetric which means $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}) \;\; \forall \boldsymbol{x}, \boldsymbol{y} \in V$.
- d obeys the triangle inequality as follows: $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z}) \;\; \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in V$

Inner products and metrics seem to be very similar in terms of their properties - however there is one important difference. When $\boldsymbol{x}$ and $\boldsymbol{y}$ are close to each other the inner product is large but the distance metric is small. On the other hand when $\boldsymbol{x}$ and $\boldsymbol{y}$ are far apart, then the inner product is small but the distance metric is large.

Angle between two vectors

- In addition to being able to capture the lengths of vectors and the distance between vectors, inner products can also capture the **angle** ω between two vectors and can thus capture the geometry of a vector space.
- The key to using the inner product to characterize the angle between two vectors is the Cauchy-Schwarz inequality.
- Assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ are not the $\mathbf{0}$ vector. Then the Cauchy-Schwarz inequality tells us that

$$-1 \leq \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|} \leq 1 \tag{1}$$

Angle between two vectors

- Since the Cauchy-Schwarz ratio lies between -1 and 1 we can set it equal to the cosine of a unique angle $\omega \in [0, \pi]$ such that

$$\cos(\omega) = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|} \tag{2}$$

- The angle ω is the angle between two vectors.
- The notion of angle captures similarity of orientation between two vectors. When the dot product is close to zero, the vectors are more or less pointing in orthogonal directions and $\omega \approx \pi/2$.

Table of Contents

Orthogonality

- A key feature of the inner product is that we can use it to characterize vectors that are orthogonal.
- Two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal if and only if the inner product between them is 0. For an orthogonal pair of vectors $\boldsymbol{x}$, $\boldsymbol{y}$ we can write $\boldsymbol{x} \perp \boldsymbol{y}$.
- If additionally $\|\boldsymbol{x}\| = 1 = \|\boldsymbol{y}\|$, then $\boldsymbol{x}$ and $\boldsymbol{y}$ are said to be *orthonormal*.
- By the above definition the $\mathbf{0}$-vector is orthogonal to all vectors.
- Vectors which are orthogonal with respect to one inner product need not be orthogonal with respect to another inner product.

Example

- Consider the vectors $\boldsymbol{x} = [1, 1]^T$ and $\boldsymbol{y} = [-1, 1]^T$
- With respect to the inner product defined as a dot product we see that $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y} = 1 * -1 + 1 * 1 = 0$.
- With respect to the inner product $\boldsymbol{x}^T \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \boldsymbol{y}$, the angle between the two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ becomes

$$\cos(\omega) = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}$$

Example

- $$\begin{aligned}\cos(\omega) &= \frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{y}}{\sqrt{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}\sqrt{\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}}} \\ &= \frac{2x_1y_1 + x_2y_2}{\sqrt{(2x_1^2 + x_2^2)(2y_1^2 + y_2^2)}} \\ &= \frac{-1}{3}\end{aligned}$$

  where $\boldsymbol{A} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$.
- Thus, with respect to the second inner product the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are no longer orthogonal.

Orthonormal matrix

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is an orthogonal matrix if and only if its columns are orthonormal:

$$\boldsymbol{A}^T \boldsymbol{A} = \boldsymbol{I} = \boldsymbol{A}\boldsymbol{A}^T$$
$$\text{i.e., } \boldsymbol{A}^T = \boldsymbol{A}^{-1}$$

- Transformations by an orthonormal matrix preserve lengths. This can be seen as follows, using the dot product as the definition of the inner product:
$\|\boldsymbol{A}\boldsymbol{x}\|^2 = (\boldsymbol{A}\boldsymbol{x})^T \boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{I}\boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{x} = \|\boldsymbol{x}\|^2$.

Orthonormal matrix

- Also the angle between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ does not change after transformation by an orthonormal matrix. This can be seen as follows:

$$\begin{aligned}
\cos(\omega) &= \frac{(\boldsymbol{Ax})^T \boldsymbol{Ay}}{\|\boldsymbol{Ax}\| \ \|\boldsymbol{Ay}\|} \\
&= \frac{\boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{Ay}}{\|\boldsymbol{x}\| \ \|\boldsymbol{y}\|} \\
&= \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\| \ \|\boldsymbol{y}\|}
\end{aligned}$$

Example

- 2D-rotation matrix: $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ where θ is the angle of rotation.

Orthonormal basis

- We already looked at the concept of a basis of a vector space, and found that for the vector space $\mathbb{R}^n$ we need n basis vectors.
- Our basis vectors needed only to be linearly independent - we can ensure linear independence by ensuring that our basis vectors point in different directions, so that a linear combination of $n-1$ basis vectors cannot cancel out the nth basis vector.
- Now we will look at a special case of a basis where the vectors are all mutually orthogonal in the sense of the inner product, and each vector is of unit length. We call such a basis an orthonormal basis.

Orthonormal basis

- Question: Can you immediately think of an orthonormal basis for $\mathbb{R}^n$? Is an orthonormal basis for a vector space unique?
- Formal definition of an orthonormal basis: Consider an n-dimensional vector space V and n basis vectors $\{\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots \boldsymbol{b}_n\}$. If it is true that $\forall i, j = 1, \ldots n, i \neq j \ \langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle = 0$ and $\langle \boldsymbol{b}_i, \boldsymbol{b}_i \rangle = 1$, then the basis is called an orthonormal basis.
- If the basis vectors are only mutually orthogonal but not of length unity, then we have an orthogonal basis.

Gram-Schmidt Process

- Given a set of basis vectors for a vector space, can we convert the given basis into an orthogonal basis?
- **Idea:** We construct new vectors step by step by *removing* from each vector its projections onto the previously constructed vectors.
- Starting with linearly independent vectors $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$:
  - Keep $\boldsymbol{u}_1 = \boldsymbol{v}_1$
  - Subtract from $\boldsymbol{v}_2$ its projection onto $\boldsymbol{u}_1$ to get

$$\boldsymbol{u}_2 = \boldsymbol{v}_2 - \frac{\langle \boldsymbol{v}_2, \boldsymbol{u}_1 \rangle}{\langle \boldsymbol{u}_1, \boldsymbol{u}_1 \rangle} \boldsymbol{u}_1$$

Gram-Schmidt Process

- Subtract from $\boldsymbol{v}_k$ its projections onto $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-1}$ to get

$$\boldsymbol{u}_k = \boldsymbol{v}_k - \sum_{j=1}^{k-1} \frac{\langle \boldsymbol{v}_k, \boldsymbol{u}_j \rangle}{\langle \boldsymbol{u}_j, \boldsymbol{u}_j \rangle} \boldsymbol{u}_j$$

- Each step ensures the new vector is orthogonal to all previous ones.
- Normalizing each $\boldsymbol{u}_k$ gives an orthonormal basis:

$$\boldsymbol{e}_k = \frac{\boldsymbol{u}_k}{\|\boldsymbol{u}_k\|}$$

Gram-Schmidt Process: Example

Consider the vectors in $\mathbb{R}^2$: $\boldsymbol{v}_1 = [1, 2]^T \qquad \boldsymbol{v}_2 = [3, 4]^T$

- Step 1: $\boldsymbol{u}_1 = \boldsymbol{v}_1 = [1, 2]^T$
- Step 2: Subtract the projection of $\boldsymbol{v}_2$ onto $\boldsymbol{u}_1$

$$\boldsymbol{u}_2 = \boldsymbol{v}_2 - \frac{\langle \boldsymbol{v}_2, \boldsymbol{u}_1 \rangle}{\langle \boldsymbol{u}_1, \boldsymbol{u}_1 \rangle} \boldsymbol{u}_1$$

$$= \begin{bmatrix} 3 \\ 4 \end{bmatrix} - \frac{(3)(1) + (4)(2)}{(1)^2 + (2)^2} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix} - \frac{11}{5} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

- Resulting orthogonal vectors: $\boldsymbol{u}_1 = [1, 2]^T, \quad \boldsymbol{u}_2 = [\frac{4}{5}, -\frac{2}{5}]^T$

Key Takeaways

- A vector space $(V, +, \cdot)$ is closed under addition and scalar multiplication; a basis provides unique coordinates for every vector in the space
- A norm $\|\cdot\|$ measures the length of a vector; an inner product $\langle \cdot, \cdot \rangle$ encodes both length and angle (and hence orthogonality)
- Gram-Schmidt orthogonalisation converts any linearly independent set into an orthonormal basis by iteratively subtracting projections
- These geometric foundations — norms, inner products, orthonormal bases — underlie every matrix decomposition and optimisation algorithm in the course

Thank you :)