Mathematics for Machine Learning

Prof Willie Brink

Applied Mathematics, Stellenbosch University

Lecture 2: Analytic Geometry

Contents of the module

3.1 Norms

A norm on a vector space V is a function which assigns each vector $\boldsymbol{x}$ its length $\|\boldsymbol{x}\| \in \mathbb{R}$, such that for all $\lambda \in \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in V$,

- $\|\lambda \boldsymbol{x}\| = |\lambda| \, \|\boldsymbol{x}\|$
- $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ *triangle inequality*
- $\|\boldsymbol{x}\| \geq 0$, and $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0}$

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

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

3.2 Inner products

The dot product in $\mathbb{R}^n$: $\boldsymbol{x}^\mathsf{T}\boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$

In general, $\Omega : V \times V \to \mathbb{R}$ is an inner product on vector space V if it is bilinear*, symmetric ($\Omega(\boldsymbol{x}, \boldsymbol{y}) = \Omega(\boldsymbol{y}, \boldsymbol{x})$), and positive definite ($\forall \boldsymbol{x} \in V \setminus \{\boldsymbol{0}\} : \Omega(\boldsymbol{x}, \boldsymbol{x}) > 0$, and $\Omega(\boldsymbol{0}, \boldsymbol{0}) = 0$).

We often write $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ instead of $\Omega(\boldsymbol{x}, \boldsymbol{y})$. The pair $(V, \langle \cdot, \cdot \rangle)$ is an inner product space. If using the dot product, we call $(V, \langle \cdot, \cdot \rangle)$ a Euclidean vector space.

Consider a vector space V with inner product $\langle \cdot, \cdot \rangle$, and basis $B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$ of V. It then follows that for any $\boldsymbol{x}, \boldsymbol{y} \in V$, $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \hat{\boldsymbol{x}}^\mathsf{T} \boldsymbol{A} \hat{\boldsymbol{y}}$, with $A_{i,j} = \langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle$ and $\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}$ the coordinate vectors of $\boldsymbol{x}, \boldsymbol{y}$ w.r.t. B.

* $\Omega(\lambda \boldsymbol{x} + \psi \boldsymbol{y}, \boldsymbol{z}) = \lambda \Omega(\boldsymbol{x}, \boldsymbol{z}) + \psi \Omega(\boldsymbol{y}, \boldsymbol{y})$ and $\Omega(\boldsymbol{x}, \lambda \boldsymbol{y} + \psi \boldsymbol{z}) = \lambda \Omega(\boldsymbol{x}, \boldsymbol{y}) + \psi \Omega(\boldsymbol{x}, \boldsymbol{z})$

Positive definite matrices

Since the inner product is symmetric and positive definite, $\boldsymbol{A}$ from the previous slide is symmetric, and $\forall \boldsymbol{x} \in V \setminus \{\boldsymbol{0}\} : \boldsymbol{x}^\mathsf{T} \boldsymbol{A} \boldsymbol{x} > 0$.

We say $\boldsymbol{A}$ is (symmetric) positive definite.

If $\boldsymbol{x}^\mathsf{T} \boldsymbol{A} \boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in V \setminus \{\boldsymbol{0}\}$, we say $\boldsymbol{A}$ is positive semidefinite.

For vector space V with basis B, $\langle \cdot, \cdot \rangle$ is an inner product if and only if there exists a positive definite matrix $\boldsymbol{A}$ such that $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \hat{\boldsymbol{x}}^\mathsf{T} \boldsymbol{A} \hat{\boldsymbol{y}}$, where $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ are the coordinate representations of $\boldsymbol{x}$ and $\boldsymbol{y}$ in V w.r.t. basis B.

Note: the null space of $\boldsymbol{A}$ is only $\boldsymbol{0}$, since $\boldsymbol{x}^\mathsf{T} \boldsymbol{A} \boldsymbol{x} > 0$ for all $\boldsymbol{x} \neq \boldsymbol{0}$;

the diagonal elements of $\boldsymbol{A}$ are positive, since $a_{i,i} = \boldsymbol{e}_i^\mathsf{T} \boldsymbol{A} \boldsymbol{e}_i > 0$ with $\boldsymbol{e}_i$ the ith vector of the canonical basis in $\mathbb{R}^n$.

3.3 Lengths and distances

Any inner product induces a norm: $\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$

The induced norm satisfies the Cauchy-Schwarz inequality: $|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|$

The distance between $\boldsymbol{x}$ and $\boldsymbol{y}$: $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\| = \sqrt{\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y} \rangle}$

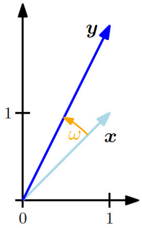
If using the dot product, we call it the Euclidean distance.

Distance is a metric, satisfying:

- $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ for all $\boldsymbol{x}, \boldsymbol{y} \in V$, and $d(\boldsymbol{x}, \boldsymbol{y}) = 0$ if and only if $\boldsymbol{x} = \boldsymbol{y}$
- $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in V$
- $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in V$

Similar vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ will result in a large inner product and a small distance.

3.4 Angles and orthogonality



According to Cauchy-Schwarz, $-1 \leq \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} \leq 1$.

There exists a unique angle $\omega \in [0, \pi]$ such that $\cos\omega = \dfrac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}$.

Vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal, $\boldsymbol{x} \perp \boldsymbol{y}$, if and only if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$.

If $\boldsymbol{x} \perp \boldsymbol{y}$ and $\|\boldsymbol{x}\| = \|\boldsymbol{y}\| = 1$, we say $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthonormal.

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

Then $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}_n$ which implies that $\boldsymbol{A}^{-1} = \boldsymbol{A}^T$.

Orthogonal matrices preserve length: $\|\boldsymbol{A}\boldsymbol{x}\|^2 = (\boldsymbol{A}\boldsymbol{x})^T(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{x}^T(\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{x} = \|\boldsymbol{x}\|^2$

Orthogonal matrices also preserve angles between vectors: $\frac{\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{y} \rangle}{\|\boldsymbol{A}\boldsymbol{x}\|\|\boldsymbol{A}\boldsymbol{y}\|} = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}$

3.5 Orthonormal basis

The basis $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n\}$ of vector space V is called an orthonormal basis (ONB) of V if $\langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle = 0$ for $i \neq j$, and $\langle \boldsymbol{b}_i, \boldsymbol{b}_i \rangle = 1$.

Gram-Schmidt process of building an ONB from a set $\tilde{\boldsymbol{b}}_1, \ldots, \tilde{\boldsymbol{b}}_n$ of basis vectors:

1. concatenate the vectors into matrix $\tilde{\boldsymbol{B}} = [\,\tilde{\boldsymbol{b}}_1 \ \cdots \ \tilde{\boldsymbol{b}}_n\,]$
2. apply Gaussian elimination to the augmented matrix $[\,\tilde{\boldsymbol{B}}\tilde{\boldsymbol{B}}^T \mid \tilde{\boldsymbol{B}}\,]$

3.6 Orthogonal complement

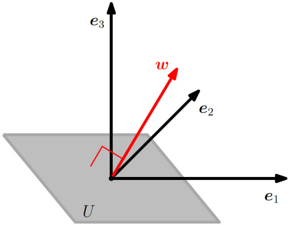
Consider a D-dimensional vector space V, and an M-dimensional subspace $U \subseteq V$.

The orthogonal complement $U^\perp$ of U is a $(D-M)$-dimensional subspace of V, and contains all vectors in V that are ortogonal to every vector in U.

$U \cap U^\perp = \{\mathbf{0}\}$, and any vector $\boldsymbol{x} \in V$ can be uniquely written as

$$\boldsymbol{x} = \sum_{i=1}^{M} \lambda_i \boldsymbol{b}_i + \sum_{j=1}^{D-M} \psi_j \boldsymbol{b}_j^\perp$$

with $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M)$ and $(\boldsymbol{b}_1^\perp, \ldots, \boldsymbol{b}_{D-M}^\perp)$ the bases of U and $U^\perp$



Example: if U describes a plane in 3D, its complement is the span of the plane's normal vector.

3.8 Orthogonal projections

A linear mapping π from V to $U \subseteq V$ is called a projection if $\pi^2 = \pi \circ \pi = \pi$.

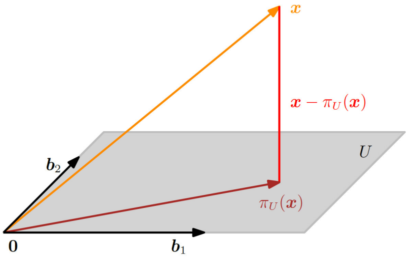
A projection matrix $\boldsymbol{P}_\pi$ has the property that $\boldsymbol{P}_\pi^2 = \boldsymbol{P}_\pi$.

The projection of vector $\boldsymbol{x} \in \mathbb{R}^n$ onto a lower-dimensional subspace U with basis $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m)$, is necessarily a linear combination of those basis vectors of U:

$$\pi_U(\boldsymbol{x}) = \lambda_1 \boldsymbol{b}_1 + \ldots + \lambda_m \boldsymbol{b}_m = \boldsymbol{B\lambda} \quad \text{with } \boldsymbol{B} = [\, \boldsymbol{b}_1 \ \cdots \ \boldsymbol{b}_m \,]$$

Three-step procedure to find $\boldsymbol{P}_\pi$:

1. find $\lambda_1, \ldots, \lambda_m$ such that $\boldsymbol{B\lambda}$ is closest to $\boldsymbol{x}$
   $\implies$ solve the normal eqn $\boldsymbol{B}^T\boldsymbol{B\lambda} = \boldsymbol{B}^T\boldsymbol{x}$
2. $\pi_U(\boldsymbol{x}) = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{x}$
3. then $\boldsymbol{P}_\pi = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T$



Projections $\pi_U(\boldsymbol{x})$ are still vectors in $\mathbb{R}^n$, but lie in a subspace of dimension m, requiring only m coordinates $\lambda_1, \ldots, \lambda_m$ to be represented in terms of the basis $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m)$.

For $\boldsymbol{Ax} = \boldsymbol{b}$ when $\boldsymbol{b}$ is not in the column space of $\boldsymbol{A}$, we may approximate a solution by projecting $\boldsymbol{b}$ to that column space $\implies$ the least-squares solution

If $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m)$ is an ONB, the proj. matrix simplifies to $\boldsymbol{P}_\pi = \boldsymbol{B}\boldsymbol{B}^T$, and $\boldsymbol{\lambda} = \boldsymbol{B}^\top \boldsymbol{x}$.

Gram-Schmidt orthogonalisation iteratively constructs an orthogonal basis $(\boldsymbol{u}_1 \ldots, \boldsymbol{u}_n)$ from any basis $(\boldsymbol{b}_1 \ldots, \boldsymbol{b}_n)$:

$$\boldsymbol{u}_1 = \boldsymbol{b}_1, \text{ and } \boldsymbol{u}_k = \boldsymbol{b}_k - \pi_{\text{span}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-1}]}(\boldsymbol{b}_k), \quad k = 2, \ldots, n$$

Projecting onto an affine subspace $L = \boldsymbol{x}_0 + U$:

$$\pi_L(\boldsymbol{x}) = \pi_U(\boldsymbol{x} - \boldsymbol{x}_0) + \boldsymbol{x}_0$$

3.9 Rotations

A class of linear mappings with orthogonal transformation matrices (length and angle preserving).

Rotation in $\mathbb{R}^2$: $\boldsymbol{R}(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$

Rotation in $\mathbb{R}^3$: combine rotations about the three standard basis vectors

$$\boldsymbol{R}_1(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}, \quad \boldsymbol{R}_2(\phi) = \begin{bmatrix} \cos\phi & 0 & \sin\phi \\ 0 & 1 & 0 \\ -\sin\phi & 0 & \cos\phi \end{bmatrix}, \quad \boldsymbol{R}_3(\psi) = \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Rotation in $\mathbb{R}^n$: fix $n-2$ dimensions and restrict the rotation to a 2D plane in $\mathbb{R}^n$
(this is called Givens rotation)