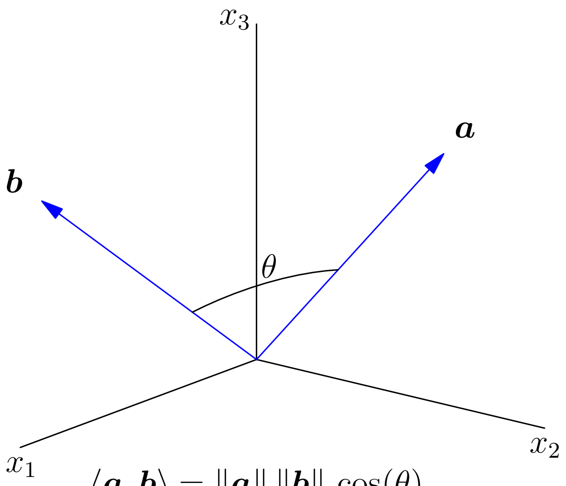
Advanced Machine Learning

Inner Product Spaces

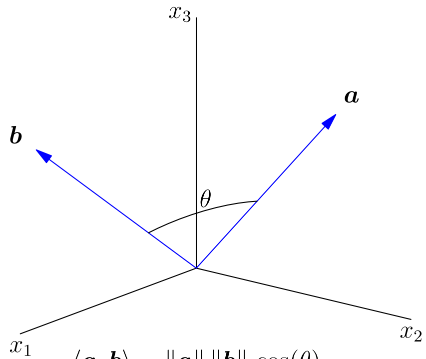


$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \|\boldsymbol{a}\| \, \|\boldsymbol{b}\| \, \cos(\theta)$$

Inner products, operators

Outline

1. Inner Products
2. Operators



$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \|\boldsymbol{a}\| \, \|\boldsymbol{b}\| \, \cos(\theta)$$

Recap

- We have looked at vector space (closed sets where we can add elements and multiply them by a scalar)
- Recall that vector spaces don't just apply to normal vectors ($\mathbb{R}^n$), but to matrices, functions, sequences, random variables, . . .
- Proper distances or metrics, $d(\boldsymbol{x},\boldsymbol{y})$, allow us to construct ideas about geometry of the vector space
- Norms, $\|\boldsymbol{x}\|$, that allow us to reason about the size of vector
- Norm induce a distance, $d(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|$

Inner Products

- We will often consider objects with an *inner product*
- For vectors in $\mathbb{R}^n$

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

- For functions

$$\langle f,g \rangle = \int_{x \in \mathcal{I}} f(x) g(x) \mathrm{d}x$$

- For $m \times n$ matrices

$$\langle \mathbf{A},\mathbf{B} \rangle = \mathrm{Tr} \mathbf{A}^{\mathsf{T}} \mathbf{B} = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij}$$

Axioms of Inner Products

- An inner product satisfies
  1. $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$ for all $\boldsymbol{x} \in \mathcal{V}$
  2. $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0}$
  3. $\langle \alpha \boldsymbol{x}, \boldsymbol{y} \rangle = \alpha \langle \boldsymbol{x}, \boldsymbol{y} \rangle$
  4. $\langle \boldsymbol{x}, \boldsymbol{y} + \boldsymbol{z} \rangle = \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \langle \boldsymbol{x}, \boldsymbol{z} \rangle$
  5. $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$
- We can show that $\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$ satisfies the axioms of a norm, so that an inner-product space is a normed space
- The norm associated with the inner-product for vectors in $\mathbb{R}^n$ (i.e. $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\mathsf{T} \boldsymbol{y}$) is the Euclidean norm $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^\mathsf{T} \boldsymbol{x}}$

Cauchy-Schwarz Inequality

- One of the most important results of inner-product spaces, known as the **Cauchy-Schwarz inequality** is that

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 \leq \langle \boldsymbol{x}, \boldsymbol{x} \rangle \langle \boldsymbol{y}, \boldsymbol{y} \rangle = \|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2$$

- Or

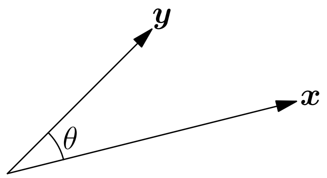
$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\| \, \|\boldsymbol{y}\|$$

- This is a very general result so for example

$$\left| \int f(x) g(x) \mathrm{d}x \right| \leq \sqrt{\left(\int f^2(x) \mathrm{d}x \right) \left(\int g^2(x) \mathrm{d}x \right)}$$

Angles Between Vectors

- A natural interpretation of the inner product is in providing a measure of the angle between vectors



$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\mathsf{T} \boldsymbol{y} = \|\boldsymbol{x}\| \, \|\boldsymbol{y}\| \, \cos(\theta)$$

- Vectors are orthogonal if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$
- We can extend this idea to functions

$$\langle f(x), g(x) \rangle = \int_{x \in \mathcal{I}} f(x) g(x) \mathrm{d}x = \|f(x)\| \, \|g(x)\| \cos(\theta)$$

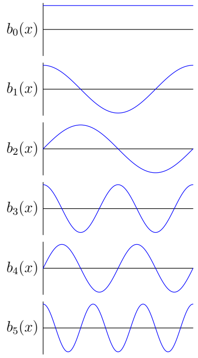
- Note that $\sin(x)$ and $\cos(x)$ are orthogonal in the interval $[0, 2\pi]$

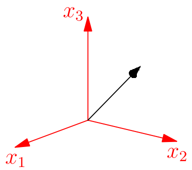
Basis Functions

- Any set of vectors $\{\boldsymbol{b}_i | i = 1, \ldots\}$ that span the space can be used as a basis or coordinate system
- The simplest and most useful case is when the vectors are orthogonal and normalised (i.e. $\|\boldsymbol{b}_i\| = 1$)
- In $\mathbb{R}^3$ we could use $\boldsymbol{b}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, $\boldsymbol{b}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, $\boldsymbol{b}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$
- This is not unique as we can rotate our basis vectors
- For an orthogonal basis we can write any vector as $\hat{\boldsymbol{x}} = \begin{pmatrix} \boldsymbol{x}^\mathsf{T} \boldsymbol{b}_1 \\ \boldsymbol{x}^\mathsf{T} \boldsymbol{b}_2 \\ \boldsymbol{x}^\mathsf{T} \boldsymbol{b}_3 \end{pmatrix}$

Orthogonal Functions

- For functions we can use any ortho-normal set of functions as a basis
- The most familiar are the Fourier functions $\sin(n\theta)$ and $\cos(n\theta)$
- Any function in $C(0,2\pi)$ can be represented by a point $\boldsymbol{f} = \begin{pmatrix} \langle f(x), b_0(x) \rangle \\ \langle f(x), b_1(x) \rangle \\ \vdots \end{pmatrix}$
- There might be an infinite number of components
- This is analogous to points in $\mathbb{R}^n$ (for large n)



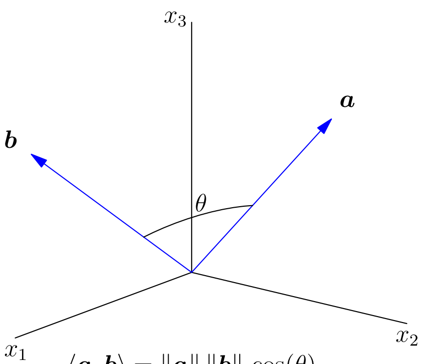


Algebraic Structure

- We have gone to these lengths as we want to show that many properties of vectors are shared by other objects (matrices, functions, etc.)
- The notions of distance (geometry), norms (size of vectors) and inner products (angles between vectors) provides a very rich set of concepts
- Vectors form the backbone of objects we will use repeated in machine learning
- The next piece of the jigsaw is to understand how we can transform these objects

Outline

1. Inner Products
2. **Operators**



$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \|\boldsymbol{a}\| \, \|\boldsymbol{b}\| \, \cos(\theta)$$

Operators

- In machine learning we are interested in transforming our input vectors into some output predictions
- To accomplish this we will apply some mapping or operators on the vector $\mathcal{T} : \mathcal{V} \to \mathcal{V}'$
- This says that $\mathcal{T}$ maps some object $\boldsymbol{x} \in \mathcal{V}$ to an object $\boldsymbol{y} = \mathcal{T}[\boldsymbol{x}]$ in a new vector space $\mathcal{V}'$
- This new vector space may or may not be the same as the original vector space
- Our objects may be any object in a vector space such as a function

Linear Operators

- Operators are in general very complicated, but a particular nice set of operators are linear operators
- $\mathcal{T}$ is a linear operator if
  1. $\mathcal{T}[a\boldsymbol{x}] = a\mathcal{T}[\boldsymbol{x}]$
  2. $\mathcal{T}[\boldsymbol{x}+\boldsymbol{y}] = \mathcal{T}[\boldsymbol{x}] + \mathcal{T}[\boldsymbol{y}]$
- For normal vectors ($\boldsymbol{x} \in \mathbb{R}^n$) the most general linear operation is

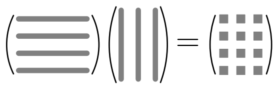
$$\mathcal{T}[\boldsymbol{x}] = \mathbf{M}\boldsymbol{x}$$

  where $\mathbf{M}$ is a matrix

Matrix multiplication

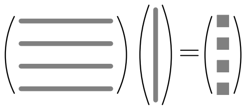
- For an $\ell \times m$ matrix $\mathbf{A}$ and an $m \times n$ matrix $\mathbf{B}$ we can compute the $(\ell \times n)$ product, $\mathbf{C} = \mathbf{AB}$, such that

$$C_{ij} = \sum_{k=1}^{m} A_{ik} B_{kj}$$



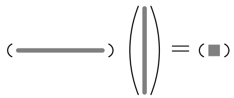
- Treating the vector $\boldsymbol{x}$ as a $n \times 1$ matrix then

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{x} \quad \Rightarrow \quad y_i = \sum_j M_{ij} x_j$$



- Using the same matrix notation we can define the inner product as

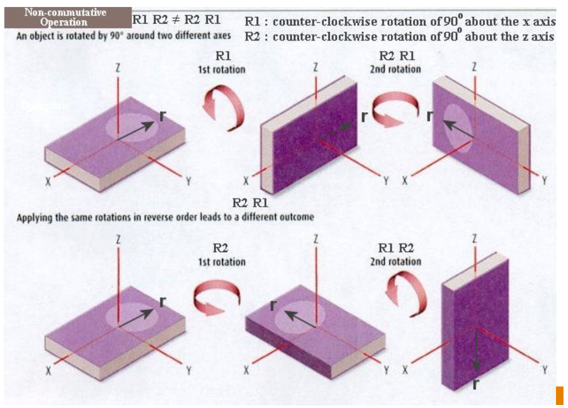
$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^{\mathsf{T}} \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$$



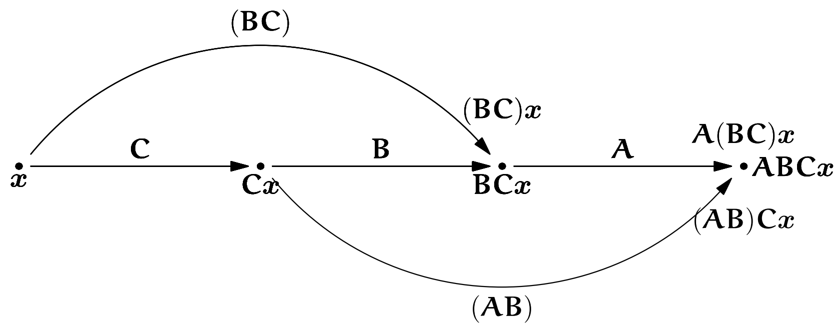
Non-commutativity

- In general $\mathbf{AB} \neq \mathbf{BA}$

$$\begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \qquad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \end{pmatrix}$$



Associativity of Mappings



- For all $\boldsymbol{x}$ we have $\mathbf{A}(\mathbf{BC})\boldsymbol{x} = (\mathbf{AB})\mathbf{C}\boldsymbol{x}$
- This implies $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$

Kernels

- The equivalent of a matrix for functions (i.e. a linear operator) is known as a kernel $K(x,y)$

$$g(x) = \mathcal{T}[f] = \int_{y\in\mathcal{I}} K(x,y)\, f(y)\, \mathrm{d}y$$

- Our domain does not need to be one dimensional, e.g.

$$g(\boldsymbol{x}) = \mathcal{T}[f] = \int_{\boldsymbol{y}\in\mathcal{I}} K(\boldsymbol{x},\boldsymbol{y})\, f(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}$$

- We shall soon see examples of high-dimensional kernels

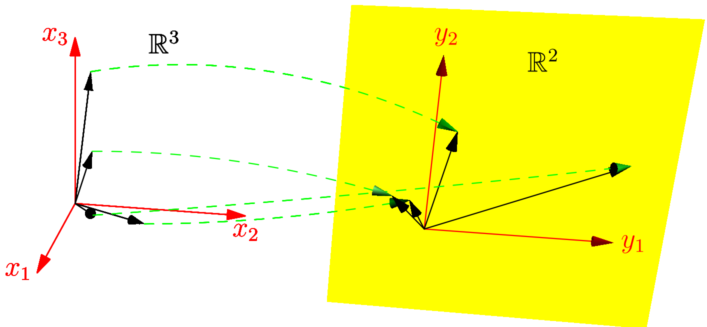
Kernels in Machine Learning

- Kernels are used heavily in machine learning
- In kernel methods such as SVM, SVR, Kernel-PCA
- They are also used in Gaussian Processes
- In all these cases we consider symmetric, positive semi-definite kernels
- Sometimes they can be interpreted as covariance between random functions

$$K(\boldsymbol{x},\boldsymbol{y}) = \mathbb{E}_{f\sim\mathcal{P}}\left[\left(f(\boldsymbol{x}) - \mu(\boldsymbol{x})\right)\left(f(\boldsymbol{y}) - \mu(\boldsymbol{y})\right)\right]$$

General Linear Mappings

- In general a linear operator will map vectors between different vector spaces
- E.g. $\mathbb{R}^3 \to \mathbb{R}^2$



Square Matrices

- We will spend a lot of time on operators that map from a vector space onto itself $\mathcal{T}: \mathcal{V} \to \mathcal{V}$
- For vectors in $\mathbb{R}^n$ such linear operators are represented by square matrices
- When there is a one-to-one mapping then we have a unique inverse
- We will study such mappings in detail in the next lecture

Summary

- We haven't covered much machine learning as such—sorry
- But mathematics is the language of machine learning and you have to get used to it
- Mathematics is like programming, if you don't understand the syntax and you can't write it down then its meaningless
- We've taken a high level view of inner product spaces and operator, this will pay us back later as we look at kernel methods