Linear Algebra in a Nutshell

Part 2: Norms, Kernels and Dimensions

Stefan Evert

Institute of Cognitive Science
University of Osnabrück, Germany
stefan.evert@uos.de

Rovereto, 20 March 2007

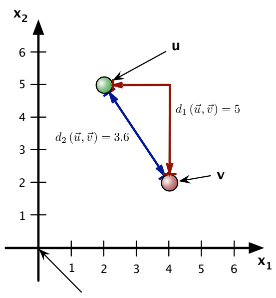
What's missing?

- We know (almost `:-)`) everything about vector spaces and the methods of linear algebra now
- But we need something else in order to perform clustering or find dimensions of major variance . . .
  - Can you guess what is missing?

☞ We need a notion of distance!

Measuring distance

- **distance** between vectors $\vec{u}, \vec{v} \in \mathbb{R}^n$ ➜ (dis)**similarity** of data points
  - $\vec{u} = (u_1, \dots, u_n)$
  - $\vec{v} = (v_1, \dots, v_n)$
- **Euclidean** distance $d_2(\vec{u}, \vec{v})$
- "**city block**" distance $d_1(\vec{u}, \vec{v})$
- both are special cases of the p-**distance** $d_p(\vec{u}, \vec{v})$ (for $p \in [1, \infty]$)



$$d_p(\vec{x}, \vec{y}) := \left(|u_1 - v_1|^p + \cdots + |u_n - v_n|^p\right)^{1/p}$$

$$d_\infty(\vec{x}, \vec{y}) = \max\{|u_1 - v_1|, \dots, |u_n - v_n|\}$$

Metric: a measure of distance

- A general measure of the distance $d(\vec{u}, \vec{v})$ between points $\vec{u}$ and $\vec{v}$ is called a **metric** and must satisfy the following **axioms**:
  - $d(\vec{u}, \vec{v}) = d(\vec{v}, \vec{u})$
  - $d(\vec{u}, \vec{v}) > 0$ for $\vec{u} \neq \vec{v}$
  - $d(\vec{u}, \vec{u}) = 0$
  - $d(\vec{u}, \vec{w}) \leq d(\vec{u}, \vec{v}) + d(\vec{v}, \vec{w})$ (**triangle inequality**)
- Metrics are very broad class of distance measures, some of which do not fit well into vector spaces
- E.g., metrics need not be **translation-invariant**

$$d(\vec{u} + \vec{x}, \vec{v} + \vec{x}) \neq d(\vec{u}, \vec{v})$$

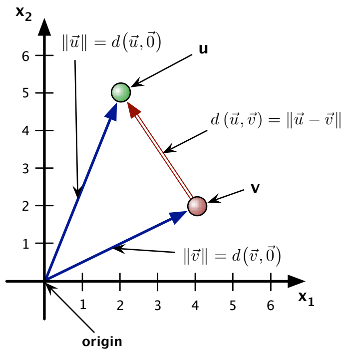
- Another unintuitive example is the **discrete metric**

$$d(\vec{u}, \vec{v}) = \begin{cases} 0 & \vec{u} = \vec{v} \\ 1 & \vec{u} \neq \vec{v} \end{cases}$$

☞ exercise: show that discrete metric satisfies axioms

Distance & norm

- Intuitively, **distance** $d(\vec{u}, \vec{v})$ should correspond to **length** $\|\vec{u} - \vec{v}\|$ of vector $\vec{u} - \vec{v}$
  - $d(\vec{u}, \vec{v})$ is a **metric**
  - $\|\vec{u} - \vec{v}\|$ is a **norm**
  - $\|\vec{u}\| = d(\vec{u}, \vec{0})$
- Such a metric is always **translation-invariant**



- $d_p(\vec{u}, \vec{v}) = \|\vec{v} - \vec{u}\|_p$
- p-**norm** for $p \in [1, \infty]$:

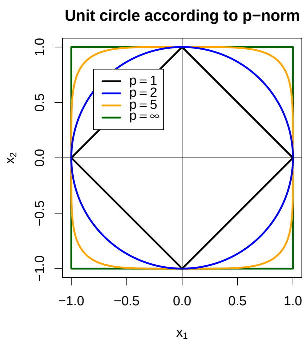
$$\|\vec{u}\|_p \coloneqq \left(|u_1|^p + \cdots + |u_n|^p\right)^{1/p}$$

Norm: a measure of length

- A general **norm** $\|\vec{u}\|$ for the length of a vector $\vec{u}$ must satisfy the following **axioms**:
  - $\|\vec{u}\| > 0$ for $\vec{u} \neq \vec{0}$
  - $\|\lambda\vec{u}\| = |\lambda| \cdot \|\vec{u}\|$ (**homogeneity**, not req'd for metric)
  - $\|\vec{u} + \vec{v}\| \leq \|\vec{u}\| + \|\vec{v}\|$ (**triangle inequality**)
- every norm defines a translation-invariant metric

$$d(\vec{u}, \vec{v}) \coloneqq \|\vec{u} - \vec{v}\|$$

Norm: a measure of length



- Visualisation of norms in $\mathbb{R}^2$ by plotting **unit circle** for each norm, i.e. points $\vec{u}$ with $\|\vec{u}\| = 1$
- Here: p-norms $\|\cdot\|_p$ for different values of p
- Triangle inequality $\Longleftrightarrow$ unit circle is **convex**
- This shows that p-norms with $p < 1$ would violate the triangle inequality

Operator and matrix norm

- The **norm** of a linear map (or "operator") $f: U \to V$ between normed vector spaces U and V is defined as

$$\|f\| \coloneqq \max\left\{\|f(\vec{u})\| \;\middle|\; \vec{u} \in U, \|\vec{u}\| = 1\right\}$$

  - $\|f\|$ depends on the norms chosen in U and V!
- The definition of the operator norm implies

$$\|f(\vec{u})\| \leq \|f\| \cdot \|\vec{u}\|$$

- norm of a matrix A = norm of corresponding map f
  - NB: this is not the same as a p-norm of A in $\mathbb{R}^{k \cdot n}$
  - **spectral norm** induced by Euclidean vector norms in U and V = largest **singular value** of A (➜ SVD)

Cave canem!

- Discussion about which norm to use for measuring distributional similarity in word space models
- Measures of **distance** between points:
  - "natural" Euclidean norm $\|\cdot\|_2$
  - city-block ("Manhattan") distance $\|\cdot\|_1$
  - maximum distance $\|\cdot\|_\infty$
  - and many other formulae ...
- Measures of the **similarity** of arrows:
  - "cosine distance" $\sim u_1 v_1 + \cdots + u_n v_n$
  - Dice coefficient (matching non-zero coordinates)
  - and, of course, many other formulae ...
  - ☞ these measures determine **angles** between arrows
- **Don't do this!** – the Euclidean norm induces a much richer and more intuitive geometric structure
  - ☞ There's a trick to make Euclidean norms more flexible

Euclidean norm & inner product

- The Euclidean norm $\|\vec{u}\|_2 = \sqrt{\langle \vec{u}, \vec{u} \rangle}$ is special because it can be derived from the **inner product**:

$$\langle \vec{u}, \vec{v} \rangle \coloneqq \vec{x}^T \vec{y} = x_1 y_1 + \cdots + x_n y_n$$

where $\vec{u} \equiv_E \vec{x}$ and $\vec{v} \equiv_E \vec{y}$ are the standard coordinates of $\vec{u}$ and $\vec{v}$ (certain other coordinate systems also work)

- The inner product is a **positive definite** and **symmetric bilinear form** with the following properties:
  - $\langle \lambda\vec{u}, \vec{v} \rangle = \langle \vec{u}, \lambda\vec{v} \rangle = \lambda \langle \vec{u}, \vec{v} \rangle$
  - $\langle \vec{u} + \vec{u}', \vec{v} \rangle = \langle \vec{u}, \vec{v} \rangle + \langle \vec{u}', \vec{v} \rangle$
  - $\langle \vec{u}, \vec{v} + \vec{v}' \rangle = \langle \vec{u}, \vec{v} \rangle + \langle \vec{u}, \vec{v}' \rangle$
  - $\langle \vec{u}, \vec{v} \rangle = \langle \vec{v}, \vec{u} \rangle$ (**symmetric**)
  - $\langle \vec{u}, \vec{u} \rangle = \|\vec{u}\|^2 > 0$ for $\vec{u} \neq \vec{0}$ (**positive definite**)
  - also called **dot product** or **scalar product**

Angles and orthogonality

- The Euclidean inner product has an important **geometric interpretation**: it can be used to define angles and orthogonality
- **Cauchy-Schwarz inequality**:

$$\left| \langle \vec{u}, \vec{v} \rangle \right| \leq \|\vec{u}\| \cdot \|\vec{v}\|$$

- **Angle** ϕ between vectors $\vec{u}, \vec{v} \in \mathbb{R}^n$:

$$\cos\phi \coloneqq \frac{\langle \vec{u}, \vec{v} \rangle}{\|\vec{u}\| \cdot \|\vec{v}\|}$$

  - $\cos\phi$ is the "cosine distance" measure of similarity
- $\vec{u}$ and $\vec{v}$ are **orthogonal** iff $\langle \vec{u}, \vec{v} \rangle = 0$
  - the **shortest connection** between a point $\vec{u}$ and a subspace U is orthogonal to all vectors $\vec{v} \in U$

Cartesian coordinates

- A set of vectors $\vec{b}^{(1)}, \ldots, \vec{b}^{(n)}$ is called **orthonormal** if the vectors are pairwise orthogonal and of unit length:
  - $\langle \vec{b}^{(j)}, \vec{b}^{(k)} \rangle = 0$ for $j \neq k$
  - $\langle \vec{b}^{(k)}, \vec{b}^{(k)} \rangle = \|\vec{b}^{(k)}\|^2 = 1$
- An orthonormal basis and the corresponding coordinates are called **Cartesian**
- Cartesian coordinates are particularly intuitive, and the inner product has the same form wrt. every Cartesian basis B: for $\vec{u} \equiv_B \vec{x}'$ and $\vec{v} \equiv_B \vec{y}'$, we have

$$\langle \vec{u}, \vec{v} \rangle = (\vec{x}')^T \vec{y}' = x'_1 y'_1 + \cdots + x'_n y'_n$$

- NB: the column vectors of the matrix B are orthonormal
  - recall that the columns of B specify the standard coordinates of the vectors $\vec{b}^{(1)}, \ldots, \vec{b}^{(n)}$

Orthogonal projection

- Cartesian coordinates $\vec{u} \equiv_B \vec{x}$ can easily be computed:

$$\left\langle \vec{u}, \vec{b}^{(k)} \right\rangle = \left\langle \sum_{j=1}^{n} x_j \vec{b}^{(j)}, \vec{b}^{(k)} \right\rangle = \sum_{j=1}^{n} x_j \underbrace{\left\langle \vec{b}^{(j)}, \vec{b}^{(k)} \right\rangle}_{=\delta_{jk}} = x_k$$

  - Kronecker delta: $\delta_{jk} = 1$ for $j = k$ and 0 for $j \neq k$

- **Orthogonal projection** $P_V : \mathbb{R}^n \to V$ to subspace $V \coloneqq \mathrm{sp}\left(\vec{b}^{(1)}, \dots, \vec{b}^{(k)}\right)$ (for $k < n$) is given by

$$P_V \vec{u} \coloneqq \sum_{j=1}^{k} \vec{b}^{(j)} \left\langle \vec{u}, \vec{b}^{(j)} \right\rangle$$

Hyperplanes & normal vectors

- A hyperplane $U \subseteq \mathbb{R}^n$ through the origin $\vec{0}$ can be characterized by the equation

$$U = \{\vec{u} \in \mathbb{R}^n \mid \langle \vec{u}, \vec{n} \rangle = 0\}$$

  for a suitable $\vec{n} \in \mathbb{R}^n$ with $\|\vec{n}\| = 1$
- $\vec{n}$ is called the **normal vector** of U
- The orthogonal projection P_U into U is given by

$$P_U \vec{v} \coloneqq \vec{v} - \vec{n} \langle \vec{v}, \vec{n} \rangle$$

- An arbitrary hyperplane $\Gamma \subseteq \mathbb{R}^n$ can analogously be characterized by

$$\Gamma = \{\vec{u} \in \mathbb{R}^n \mid \langle \vec{u}, \vec{n} \rangle = a\}$$

  where $a \in \mathbb{R}$ is the (signed) **distance** of Γ from $\vec{0}$

Orthogonal matrices

- A matrix A whose column vectors are orthonormal is called an **orthogonal** matrix
- A^T is orthogonal iff A is orthogonal

- The **inverse** of an orthogonal matrix is simply its transpose, i.e. $A^{-1} = A^T$
  - it is easy to show $A^T A = I$ by matrix multiplication, since the columns of A are orthonormal
  - since A^T is also orthogonal, it follows that $AA^T = (A^T)^T A^T = I$
  - side remark: the transposition operator $\cdot^T$ is called an **involution** because $(A^T)^T = A$

Isometric maps

- An endomorphism $f : \mathbb{R}^n \to \mathbb{R}^n$ is called an **isometry** iff $\langle f(\vec{u}), f(\vec{v}) \rangle = \langle \vec{u}, \vec{v} \rangle$ for all $\vec{u}, \vec{v} \in \mathbb{R}^n$
- Geometric interpretation: isometries preserve angles and distances (which are defined in terms of $\langle \cdot, \cdot \rangle$)
- f is an isometry iff its matrix A is orthogonal
- Coordinate transformations between Cartesian systems are isometric (because B and $B^{-1} = B^T$ are orthogonal)
- Every isometric endomorphism of $\mathbb{R}^n$ can be written as a combination of **planar rotations** and **axial reflections** in a suitable Cartesian coordinate system

$$R_\phi^{(1,3)} = \begin{bmatrix} \cos\phi & 0 & -\sin\phi \\ 0 & 1 & 0 \\ \sin\phi & 0 & \cos\phi \end{bmatrix}, \quad Q^{(2)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

General inner products

- General **inner products** can be defined by

$$\langle \vec{u}, \vec{v} \rangle_B \coloneqq (\vec{x}')^T \vec{y}' = x'_1 y'_1 + \cdots + x'_y y'_n$$

wrt. non-Cartesian basis B ($\vec{u} \equiv_B \vec{x}'$, $\vec{v} \equiv_B \vec{y}'$)

- $\langle \cdot, \cdot \rangle_B$ can be expressed in standard coordinates $\vec{u} \equiv_E \vec{x}$, $\vec{v} \equiv_E \vec{y}$ using the transformation matrix B:

$$\begin{aligned} \langle \vec{u}, \vec{v} \rangle_B &= (\vec{x}')^T \vec{y}' = (B^{-1}\vec{x})^T (B^{-1}\vec{y}) \\ &= \vec{x}^T (B^{-1})^T B^{-1} \vec{y} \eqqcolon \vec{x}^T C \vec{y} \end{aligned}$$

UNIVERSITÄT OSNABRÜCK

General inner products

Linear Algebra in a Nutshell: Norms, Kernels & Dimensions
S. Evert
Distance
Metric spaces
Vector norms
Euclidean geometry
Normal vector
Isometry
General inner product
Kernel trick

- The coefficient matrix $C \coloneqq (B^{-1})^T B^{-1}$ of the general inner product is **symmetric**

$$C^T = (B^{-1})^T ((B^{-1})^T)^T = (B^{-1})^T B^{-1} = C$$

and **positive definite**

$$\vec{x}^T C \vec{x} = (B^{-1}\vec{x})^T (B^{-1}\vec{x}) = (\vec{x}')^T \vec{x}' \geq 0$$

UNIVERSITÄT OSNABRÜCK

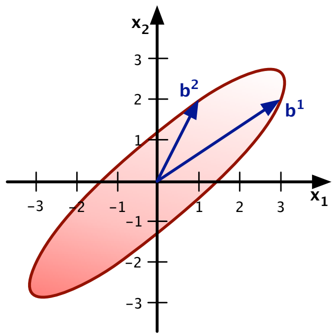
General inner products

Linear Algebra in a Nutshell: Norms, Kernels & Dimensions
S. Evert
Distance
Metric spaces
Vector norms
Euclidean geometry
Normal vector
Isometry
General inner product
Kernel trick

An example:

- $\vec{b}^{(1)} = (3,2)$, $\vec{b}^{(2)} = (1,2)$
- $B = \begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix}$
- $B^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} \\ -\frac{1}{2} & \frac{3}{4} \end{bmatrix}$
- $C = \begin{bmatrix} .5 & -.5 \\ -.5 & .625 \end{bmatrix}$
- graph shows **unit circle** of the inner product C, i.e. points $\vec{x}$ with

$$\vec{x}^T C \vec{x} = 1$$

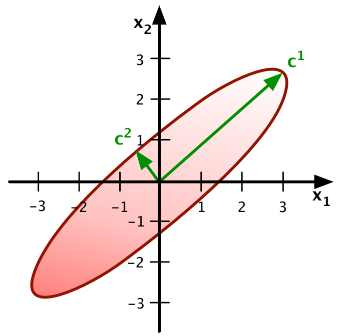


UNIVERSITÄT OSNABRÜCK

General inner products

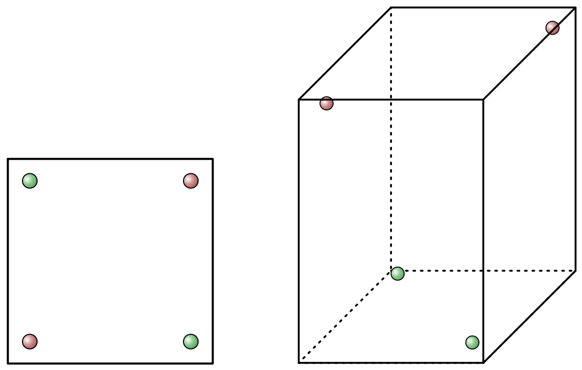
Linear Algebra in a Nutshell: Norms, Kernels & Dimensions
S. Evert
Distance
Metric spaces
Vector norms
Euclidean geometry
Normal vector
Isometry
General inner product
Kernel trick

- C is a symmetric matrix
- There is always an orthonormal basis so that C has diagonal form
- ➥ "standard" dot product with additional scaling factors (wrt. this orthonormal basis)
- Intuitively, the unit circle is a squashed and rotated circle



The kernel trick

Tweak your space, don't tweak your norms ...



The kernel trick

- Use standard inner products, but map data to higher-dimensional space before applying them
- ➥ All methods of Euclidean geometry are still available
- Non-linear mappings can drastically change the geometry of the original vector space
- The **kernel trick** allows efficient computation of inner products and distances without an explicit high-dimensional representation

$$\langle \vec{u}, \vec{v} \rangle = f(\vec{u}, \vec{v})$$

where f must satisfy the properties of an inner product

Kernelisation

- Kernelised versions of all algorithms from linear algebra and normed vector spaces can be formulated
- A hyperplane in a kernelised space corresponds to a **non-linear classifier** in the original space
- ➥ this is the principle behind support vector machines

I think that's enough for today ...