Norms and distances

The history of science is the history of *measurement*, for instnace, see the **history of timekeeping**. In a very real sense sciences advances have been enabled by more precise ways of performing measurements. The industrial revolution was created not just by coal and steam and steel, but also by better measurements. The Greeks (Hero of Alexandria, maybe) invented a proto-steam engine. Alt. history buffs often speculate, “What if they didn’t treat it like a toy?” But they couldn’t. A real steam engine, one that delivers useful work, need precisely milled components. It requires measurement.

Data *science* is no different from science. We need measurement, but our objects of study are usually digital, mathematical and statistical. Luckily mathematicians have not be laggardly when inventing ways of measuring such objects. Hence we get to norms and distances.

A quick cheat sheet on norms, inner products and distances follows:

Defining properties

Name	Inner product $< \mathbf{u}, \mathbf{v} >: V \times V \to F$	Norm $\|\|\mathbf{u}\|\| : V \to \mathbb{R}$	Distance $d(\mathbf{u}, \mathbf{v}) : V \times V \to \mathbb{R}$
Positive definiteness, non-negativity, point separating	$< \mathbf{u}, \mathbf{u} > \geq 0$, with equality if and only if $\mathbf{u} = 0$	$\|\|\mathbf{u}\|\| \geq 0$, with equality if and only if $\mathbf{u} = 0$	$d(\mathbf{u}, \mathbf{v}) = 0$ if and only if $\mathbf{u} = \mathbf{v}$
(Conjugate) symmetry	$< \mathbf{u}, \mathbf{v} >= \overline{< \mathbf{v}, \mathbf{u} >}$, which means $< \mathbf{u}, \mathbf{u} >$ is real.		$d(\mathbf{u}, \mathbf{v}) = d(\mathbf{v}, \mathbf{u})$
Linearity or homogeneity	$< a\mathbf{u} + b\mathbf{v}, \mathbf{z} >= a < \mathbf{u}, \mathbf{z} > +b < \mathbf{v}, \mathbf{z} >$	$\|\|x\mathbf{u}\|\| = \|x\| \times \|\|\mathbf{u}\|\|$	
Triangle inequality, subadditivity		$\|\|\mathbf{u} + \mathbf{v}\|\| \leq \|\|\mathbf{u}\|\| + \|\|\mathbf{v}\|\|$	$d(\mathbf{u}, \mathbf{w}) \leq d(\mathbf{u}, \mathbf{v}) + d(\mathbf{v}, \mathbf{w})$

- V is a vector space
- F is $\mathbb{R}$ or $\mathbb{C}$.
- Linearity in first arg could be linearity in second (and conj linearity in the other arg) for inner products.

From inner product to norm to distance to similarity

We can sometimes start from an inner product and derive the others, $e.g.$,

$$< \mathbf{u}, \mathbf{v} > \quad \Rightarrow \quad ||\mathbf{u}|| =< \mathbf{u}, \mathbf{v} >^{1/2} \quad \Rightarrow \quad d(\mathbf{u}, \mathbf{v}) = ||\mathbf{u} - \mathbf{v}|| \quad \Rightarrow \quad s(\mathbf{u}, \mathbf{v}) = 1/d(\mathbf{u}, \mathbf{v})$$

The chain doesn’t work in the other directions without extra conditions (and isn’t necessarily unique).

A Table of common norms/inner products/distances

Show the common cases and their relationships

Name	Space	inner product	Norm	Distance	Similarity
L_p for $p \geq 1$	$\mathbb{R}^n$ (vectors)		$(\sum_i \lvert x_i \rvert^p)^{1/p}$	$(\sum_i \lvert x_i - y_i \rvert^p)^{1/p}$	
L_2	$\mathbb{R}^n$ (vectors)	$\sum_i x_i y_i$	$(\sum_i \lvert x_i \rvert^2)^{1/2}$	$(\sum_i \lvert x_i - y_i \rvert^2)^{1/2}$	
L_0	$\mathbb{R}^n$ (vectors)		$\sum_i I(\lvert x_i \rvert > 0)$	$\sum_i I(\lvert x_i - y_i \rvert > 0)$	
$L_\infty = \lim_{p \to \infty} L_p$	$\mathbb{R}^n$ (vectors)		$\max_i \lvert x_i \rvert$	$\max_i \lvert x_i - y_i \rvert$	
Cosine	$\mathbb{R}^n$ (vectors)			$1 - C(\mathbf{u}, \mathbf{v})$	$C(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\lvert\lvert \mathbf{u} \rvert\rvert_2 \times \lvert\lvert \mathbf{v} \rvert\rvert_2}$ $= \lvert\lvert \mathbf{u} \rvert\rvert_2 \lvert\lvert \mathbf{v} \rvert\rvert_2 \cos(\theta)$
L_p	$\mathbb{R}^{n \times m}$ (matrices)		$\sup_{x \neq 0} \frac{\lvert\lvert Ax \rvert\rvert_p}{\lvert\lvert x \rvert\rvert_p}$	$\lvert\lvert A - B \rvert\rvert$	
L_2	$\mathbb{R}^{n \times m}$ (matrices)		largest singular value, σ_1	$\lvert\lvert A - B \rvert\rvert$	
L_1	$\mathbb{R}^{n \times m}$ (matrices)		max abs. col. sum	$\lvert\lvert A - B \rvert\rvert$	
L_∞	$\mathbb{R}^{n \times m}$ (matrices)		max abs. row sum	$\lvert\lvert A - B \rvert\rvert$	
Entry-wise $\lvert\lvert A \rvert\rvert_{p,q}$	$\mathbb{R}^{n \times m}$ (matrices)		$\left[\sum_j (\sum_i \lvert a_{ij} \rvert^p)^{q/p}\right]^{1/q}$	$\lvert\lvert A - B \rvert\rvert$	
Schatten p-norm	$\mathbb{R}^{n \times m}$ (matrices)		$(\sum_i \lvert \sigma_i \rvert^p)^{1/p}$		
Frobenius $\lvert\lvert A \rvert\rvert_F$	$\mathbb{R}^{n \times m}$ (matrices)	$\mathrm{trace}(A^T B) = \sum_{i,j} a_{ij} b_{ij}$	$\sqrt{\sum_{ij} \lvert a_{ij} \rvert^2} = \sqrt{\sum_k \sigma_k^2}$	$\sqrt{\sum_{ij} \lvert a_{ij} - b_{ij} \rvert^2}$	
Nuclear $\lvert\lvert A \rvert\rvert_*$	$\mathbb{R}^{n \times m}$ (matrices)		trace($\sqrt{A^* A}$) $= \sum_i \sigma_i$	$\lvert\lvert A - B \rvert\rvert$	
$\lvert\lvert A \rvert\rvert_{max}$	$\mathbb{R}^{n \times m}$ (matrices)		$\max_{ij} \lvert a_{ij} \rvert$	$\lvert\lvert A - B \rvert\rvert$	
all of the above	Tensors				
Kolmogorov-Smirnov	Probability distribution function			$\sup_x \lvert F(x) - G(x) \rvert$	
Jaccard	Sets			$1 - J(A, B)$	$J(A, B) = \frac{A \cap B}{A \cup B}$
Hamming	Strings length n			# of different symbols	
Levenshtein	Strings			# of edits	

Empty elements of the table indicate something that is either not possible to define, or at least not commonly used.

Many of these norms and distances have other synonymous names, *e.g.,*

- Manhattan or taxicab or … = L_1
- Euclidean = L_2
- The L_p norms (distances, …) have equivalents for function spaces, involving integrals instead of sums
- Chebyshev or maximum = L_∞
- Minkowski $= L_p$
- Levenshtein = Edit (although that is not unique)
- Total variation distance is related to L_1
- Spectral norm (for matrices) $= L_2$
- Entry-wise (when $p = q$) gives a vectorized version of L_p norm
- The nuclear norm, which is related to the rank, is the Schatten norm for $p = 1$
- The nuclear norm is also the “convex envelope” of rank(A), so can be used to move towards rank minimisation.

- The Frobenius norm is the Schatten norm for $p = 2$

Note: $|A| = \sqrt{A^* A} = B$ such that $BB = A^* A$, which works because $A^* A$ is positive definite. The singular values of A are the eigenvalues of $\sqrt{A^* A}$.

Others

There are so, so many norms and distances. Here are a few more examples:

- Mahalanobis
- Wasserstein](https://en.wikipedia.org/wiki/Leonid_Vaseršteĭn) distance** or Kantorovich–Rubinstein **metric**
- Hellinger distance
- Cut norms (for matrices), Grothendieck norm
- Dual norm

- Logarithmic norm
- Structural similarity (SSIM) for images, https://ece.uwaterloo.ca/~z70wang/research/ssim/

Non-metrics

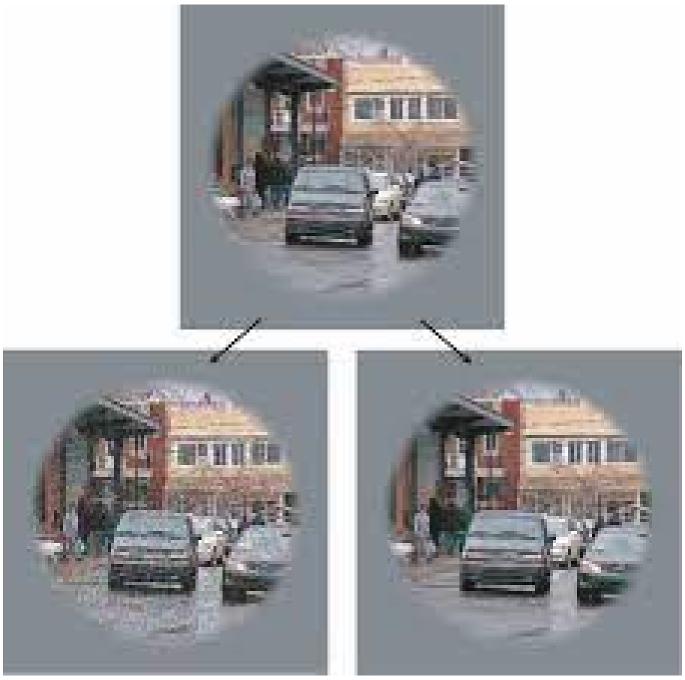
It is common, particularly for "distances" that one or more properties of a formal metric are not valid. We can still use such things, but with a little more care (please).

- Jaro–Winkler (Strings)
- Kullback-Leibler
- Shannon-Jensen

They commonly are given names like pseudo-metrics or divergences. Some of these are derived from a starting point of a similarity metric, but that doesn't have the same mathematical niceties as a distance.

Pictures of contours of common vector norms

There are so many norms, particularly for matrices, but none (of the above) are very good at capturing perceptual differences.



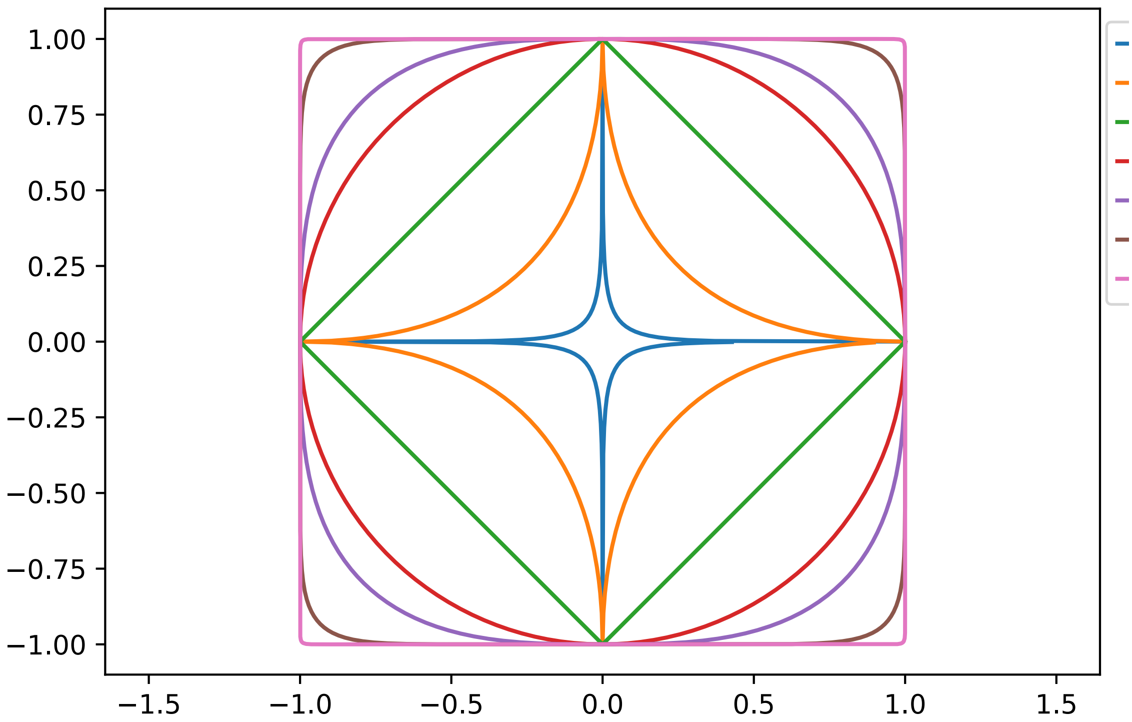
From "Does spatial invariance result from insensitivity to change?", Kingdom, Field and Olmos, Journal of Vision (2007) 7(14):11, 1–13, http://redwood.psych.cornell.edu/papers/kingdom_field_olmos_2007.pdf

Some useful theorems

- Cauchy-Schwarz inequality
- Hölder's inequality.
- $trace(A) = \sum_i \lambda_i = \sum_i |\sigma_i|$
-

To do

Plot level curves (contours) of the L_p vector norms in 2D.



Links

- https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa
- https://towardsdatascience.com/importance-of-distance-metrics-in-machine-learning-modelling-e51395ffe60d
- https://dsp.stackexchange.com/questions/188/what-distance-metric-can-i-use-for-comparing-images
- "Does spatial invariance result from insensitivity to change?", Kingdom, Field and Olmos, Journal of Vision (2007) 7(14):11, 1–13, http://redwood.psych.cornell.edu/papers/kingdom_field_olmos_2007.pdf
- https://chrischoy.github.io/research/matrix-norms/
- https://math.ntnu.edu.tw/~jschen/Papers/schatten-p-norm-JNCA.pdf