Cross Validated

Why is Euclidean distance not a good metric in high dimensions?

Asked 10 years, 10 months ago Modified 3 years ago Viewed 167k times

363

I read that 'Euclidean distance is not a good distance in high dimensions'. I guess this statement has something to do with the curse of dimensionality, but what exactly? Besides, what is 'high dimensions'? I have been applying hierarchical clustering using Euclidean distance with 100 features. Up to how many features is it 'safe' to use this metric?

machine-learning clustering distance-functions metric high-dimensional

Share Cite Edit Follow

edited May 18, 2014 at 22:14
gung - Reinstate Monica
148k 90 410 725

asked May 18, 2014 at 17:50
teaLeef
3,847 3 15 11

6 Closely related: [Euclidean distance is usually not good for sparse data?](#) as pointed out by [facuq](#). – cardinal May 19, 2014 at 2:22

6 This is likely too basic for you; I wrote a series of blog posts on the subject of the Euclidean metric in higher dimensions and how that impacts searching vector spaces for nearest matches. [blogs.msdn.com/b/ericlippert/archive/tags/...](#) – Eric Lippert May 19, 2014 at 3:02

1 @HorstGrünbusch see answers below for some references. Variance of distances becomes small compare to the average. So at some point, you run into trouble choosing thresholds, weights, ordering; and you may even get numerical precision problems, too. But if your data is sparse, it likely is of much lower *intrinsic* dimensionality. – Has QUIT--Anony-Mousse May 19, 2014 at 8:17

3 "high dimensions" seems to be a misleading term - some answers are treating 9-12 as "high dimensions", but in other areas high dimensionality would mean thousands or a million dimensions (say, measuring angles between bag-of-words vectors where each dimension is the frequency of some word in a dictionary), and 100 dimensions would be called low, not high. – Peteris May 20, 2014 at 7:58

2 This question could really do with some context. *Not good for what?* – Szabolcs May 20, 2014 at 21:04

8 Answers

Sorted by: Highest score (default)

381

A great summary of non-intuitive results in higher dimensions comes from "[A Few Useful Things to Know about Machine Learning](#)" by Pedro Domingos at the University of Washington:

+50

> [O]ur intuitions, which come from a three-dimensional world, often do not apply in high-dimensional ones. In high dimensions, most of the mass of a multivariate Gaussian distribution is not near the mean, but in an increasingly distant "shell" around it; and most of the volume of a high-dimensional orange is in the skin, not the pulp. If a constant number of examples is distributed uniformly in a high-dimensional hypercube, beyond some dimensionality most examples are closer to a face of the hypercube than to their nearest neighbor. And if we approximate a hypersphere by inscribing it in a hypercube, in high dimensions almost all the volume of the hypercube is outside the hypersphere. This is bad news for machine learning, where shapes of one type are often approximated by shapes of another.

The article is also full of many additional pearls of wisdom for machine learning.

Another application, beyond machine learning, is nearest neighbor search: given an observation of interest, find its nearest neighbors (in the sense that these are the points with the smallest distance from the query point). But in high dimensions, a curious phenomenon arises: the ratio between the nearest and farthest points approaches 1, i.e. the points essentially become uniformly distant from each other. This phenomenon can be observed for wide variety of distance metrics, but it is more pronounced for the Euclidean metric than, say, Manhattan distance metric. The premise of nearest neighbor search is that "closer" points are more relevant than "farther" points, but if all points are essentially uniformly distant from each other, the distinction is meaningless.

From Charu C. Aggarwal, Alexander Hinneburg, Daniel A. Keim, "On the Surprising Behavior of Distance Metrics in High Dimensional Space":

> It has been argued in [Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, Uri Shaft, "When Is 'Nearest Neighbor' Meaningful?"] that under certain reasonable assumptions on the data distribution, the ratio of the distances of the nearest and farthest neighbors to a given target in high dimensional space is almost 1 for a wide variety of data distributions and distance functions. In such a case, the nearest neighbor problem becomes ill defined, since the contrast between the distances to diferent data points does not exist. In such cases, even the concept of proximity may not be meaningful from a qualitative perspective: a problem which is even more fundamental than the performance degradation of high dimensional algorithms.
>
> ... Many high-dimensional indexing structures and algorithms use the [E]uclidean distance metric as a natural extension of its traditional use in two- or three-dimensional spatial applications. ... In this paper we provide some surprising theoretical and experimental results in analyzing the dependency of the L_k norm on the value of k. More specifically, we show that the relative contrasts of the distances to a query point depend heavily on the L_k metric used. This provides considerable evidence that the meaningfulness of the L_k norm worsens faster within increasing dimensionality for higher values of k. Thus, for a given problem with a fixed (high) value for the dimensionality d, it may be preferable to use lower values of k. This

> means that the L_1 distance metric (Manhattan distance metric) is the most preferable for high dimensional applications, followed by the Euclidean metric (L_2). ...

The authors of the "Surprising Behavior" paper then propose using L_k norms with $k < 1$. They produce some results which demonstrate that these "fractional norms" exhibit the property of increasing the contrast between farthest and nearest points. However, later research has concluded against fractional norms. See: "Fractional norms and quasinorms do not help to overcome the curse of dimensionality" by Mirkes, Allohibi, & Gorban (2020). (Thanks to michen00 for the comment and helpful citation.)

Share Cite Edit Follow

edited Feb 14, 2022 at 20:48

answered May 18, 2014 at 22:43
Sycorax ♦
94.6k 23 238 392

7 A good follow-up reading to Aggarwal, Hinneburg, & Keim (2001) is "Fractional norms and quasinorms do not help to overcome the curse of dimensionality" by Mirkes, Allohibi, & Gorban (2020). doi.org/10.3390/e22101105 – michen00 Aug 24, 2021 at 3:48

157

+50

The notion of Euclidean distance, which works well in the two-dimensional and three-dimensional worlds studied by Euclid, has some properties in higher dimensions that are contrary to our (maybe just *my*) geometric intuition which is also an extrapolation from two and three dimensions.

Consider a 4×4 square with vertices at $(\pm 2, \pm 2)$. Draw four unit-radius circles centered at $(\pm 1, \pm 1)$. These "fill" the square, with each circle touching the sides of the square at two points, and each circle touching its two neighbors. For example, the circle centered at $(1, 1)$ touches the sides of the square at $(2, 1)$ and $(1, 2)$, and its neighboring circles at $(1, 0)$ and $(0, 1)$. Next, draw a small circle *centered at the origin* that touches all four circles. Since the line segment whose endpoints are the centers of two osculating circles passes through the point of osculation, it is easily verified that the small circle has radius $r_2 = \sqrt{2} - 1$ and that it touches touches the four larger circles at $(\pm r_2/\sqrt{2}, \pm r_2/\sqrt{2})$. Note that the small circle is "completely surrounded" by the four larger circles and thus is also completely inside the square. Note also that the point $(r_2, 0)$ lies on the small circle. Notice also that from the origin, one cannot "see" the point $(2, 0)$ on the edge of the square because the line of sight passes through the point of osculation $(1, 0)$ of the two circles centered at $(1, 1)$ and $(1, -1)$. Ditto for the lines of sight to the other points where the axes pass through the edges of the square.

Next, consider a $4 \times 4 \times 4$ cube with vertices at $(\pm 2, \pm 2, \pm 2)$. We fill it with 8 osculating unit-radius spheres centered at $(\pm 1, \pm 1, \pm 1)$, and then put a smaller osculating sphere centered at the origin. Note that the small sphere has radius $r_3 = \sqrt{3} - 1 < 1$ and the point $(r_3, 0, 0)$ lies on the surface of the small sphere. But notice also that in three dimensions, one *can* "see" the point $(2, 0, 0)$ from the origin; there are no bigger bigger spheres blocking the view as happens in two dimensions. These clear lines of sight from the origin to the points where the axes pass through the surface of the cube occur in all larger dimensions as well.

Generalizing, we can consider a n-dimensional hypercube of side 4 and fill it with 2^n osculating unit-radius hyperspheres centered at $(\pm 1, \pm 1, \ldots, \pm 1)$ and then put a "smaller" osculating sphere of radius

$$r_n = \sqrt{n} - 1 \tag{1}$$

at the origin. The point $(r_n, 0, 0, \ldots, 0)$ lies on this "smaller" sphere. But, notice from (1) that when $n = 4$, $r_n = 1$ and so the "smaller" sphere has unit radius and thus really does not deserve the soubriquet of "smaller" for $n \geq 4$. Indeed, it would be better if we called it the "larger sphere" or just "central sphere". As noted in the last paragraph, there is a clear line of sight from the origin to the points where the axes pass through the surface of the hypercube. Worse yet, when $n > 9$, we have from (1) that $r_n > 2$, and thus the point $(r_n, 0, 0, \ldots, 0)$ on the central sphere *lies outside the hypercube of side 4 even though it is "completely surrounded" by the unit-radius hyperspheres that "fill" the hypercube (in the sense of packing it).* The central sphere "bulges" outside the hypercube in high-dimensional space. I find this very counter-intuitive because my mental translations of the notion of Euclidean distance to higher dimensions, using the geometric intuition that I have developed from the 2-space and 3-space that I am familiar with, do not describe the reality of high-dimensional space.

My answer to the OP's question "Besides, what is 'high dimensions'?" is $n \geq 9$.

Share Cite Edit Follow

edited Dec 8, 2021 at 13:54
Benjamin Wang
99 10

answered May 19, 2014 at 19:53
Dilip Sarwate
48.2k 4 126 236

20 Related: History of the high-dimensional volume paradox. – cardinal May 20, 2014 at 0:17

11 @stackoverflowuser2010: If this answer's completely incomprehensible, how can you tell whether it addresses or attempts to address the original question? A more constructive approach might be to ask for elucidation of any points you find unclear rather than dismissing the whole thing out of hand. – Scortchi ♦ Jul 12, 2016 at 8:21

8 @stackoverflowuser2010 Since this answer has many dozens of upvotes, it would appear that many people feel that it's both reasonably comprehensible and responds in some acceptable way to the question. Perhaps you could attempt a more constructive criticism -- how, specifically do you think this answer would be improved? What should it include that it does not? – Glen_b Jul 12, 2016 at 9:35

1 @Scortchi: Maybe I'm expecting too much, but a clear-cut answer to this question that could help the community would be something like "Euclidean distance is not a good metric because <X>". – stackoverflowuser2010 Jul 13, 2016 at 0:08

9 @stackoverflow2010 You won't ever see a "good" answer like that *because* <things are much more complicated than if-then statements>. If you want an easy answer, it's most likely false. Just like damn Brexit liars, they were good at offering easy answers (false, but easy). – Has QUIT--Anony-Mousse Jul 14, 2016 at 19:32

53

It is a matter of **signal-to-noise**. Euclidean distance, due to the squared terms, is particular sensitive to noise; but even Manhattan distance and "fractional" (non-metric) distances suffer.

I found the studies in this article very enlightening:

Zimek, A., Schubert, E. and Kriegel, H.-P. (2012),
A survey on unsupervised outlier detection in high-dimensional numerical data.
Statistical Analy Data Mining, 5: 363–387. doi: 10.1002/sam.11161

It revisits the observations made in e.g. On the Surprising Behavior of Distance Metrics in High Dimensional Space by Aggarwal, Hinneburg and Keim mentioned by @Pat. But it also shows how out synthetic experiments are misleading and that in fact **high-dimensional data *can* become easier**. If you have a lot of (redundant) signal, and the new dimensions add little noise.

The last claim is probably most obvious when considering duplicate dimensions. Mapping your data set $x, y \to x, y, x, y, x, y, x, y, \ldots, x, y$ increases representative dimensionality, but does not at all make Euclidean distance fail. (See also: intrinsic dimensionality)

So in the end, it still depends on your data. If you have a lot of useless attributes, Euclidean distance will become useless. If you could easily embed your data in a low-dimensional data space, then Euclidean distance should also work in the full dimensional space. In particular for *sparse* data, such as TF vectors from text, this does appear to be the case that the data is of much lower dimensionality than the vector space model suggests.

Some people believe that cosine distance is better than Euclidean on high-dimensional data. I do not think so: cosine distance and Euclidean distance are *closely* related; so we must expect them to suffer from the same problems. However, textual data where cosine is popular is usually *sparse*, and cosine is faster on data that is sparse - so for sparse data, there are good reasons to use cosine; and because the data is sparse the intrinsic dimensionality is much much less than the vector space dimension.

See also this reply I gave to an earlier question:
https://stats.stackexchange.com/a/29647/7828

Share Cite Edit Follow

edited Jun 11, 2020 at 14:32
Community Bot
1

answered May 19, 2014 at 8:13
Has QUIT--Anony-Mousse
43k 7 69 111

The average angle of randomly placed points in $[-1, 1]^n$ is always close to 90° for big n (see plots here) – Martin Thoma Apr 1, 2017 at 13:10

2 And what would be the conclusion from that? On [-1;1]^d one shouldn't use Cosine because it is not defined at 0, the average doesn't tell us anything about the curse, and uniform data is unrealistic. – Has QUIT--Anony-Mousse Apr 1, 2017 at 14:07

I didn't try it by now, but I guess that the angles look similar for real data. The fact that it is not defined at 0 should not really matter as it is just a single point. My conclusion is similar to yours: Cosine distance is not well-suited for high-dimensional spaces (though there might be domains were it still works) – Martin Thoma Apr 1, 2017 at 14:16

A more realistic scenario would be points on the nonnegative unit sphere. And the measure of interest would likely be variance, not mean. – Has QUIT--Anony-Mousse Apr 1, 2017 at 14:29

1 No, that is not a unit sphere. The construction is taking the usual unit sphere, and intersecting it with nonnegative space. You want: A) all vectors have unit length, and B) no vector entry is negative. As used with text, where cosine is most popular. – Has QUIT--Anony-Mousse Apr 2, 2017 at 9:11

37

The best place to start is probably to read *On the Surprising Behavior of Distance Metrics in High Dimensional Space* by Aggarwal, Hinneburg and Keim . There is a currently working link [here (pdf)], but it should be very google-able if that breaks. In short, as the number of dimensions grows, the relative euclidean distance between a point in a set and its closest neighbour, and between that point and its furthest neighbour, changes in some non-obvious ways. Whether or not this will badly affect your results depends a great deal on what you're trying to achieve and what your data's like.

Share Cite Edit Follow

edited Jun 2, 2015 at 15:32
samthebrand
107 5

answered May 18, 2014 at 20:10
Pat
2,599 1 19 15

See also "Fractional norms and quasinorms do not help to overcome the curse of dimensionality" (Mirkes, Allohibi, & Gorban 2020) doi.org/10.3390/e22101105 – michen00 Aug 24, 2021 at 3:41

11

Euclidean distance is very rarely a good distance to choose in Machine Learning and this becomes more obvious in higher dimensions. This is because most of the time in Machine Learning you are not dealing with a Euclidean Metric Space, but a Probabilistic Metric Space and therefore you should be using probabilistic and information theoretic distance functions, e.g. entropy based ones.

Humans like euclidean space because it's easy to conceptualize, furthermore it's mathematically easy because of linearity properties that mean we can apply linear algebra. If we define distances in terms of, say Kullback-Leibler Divergence, then it's harder to visualize and work with mathematically.

Share Cite Edit Follow

answered May 20, 2014 at 9:44
samthebest
933 5 11

3 It can be problematically, as KL Divergence isn't a metric. :-) – Carlos Agarie May 21, 2014 at 21:52

2 If one needs symmetry, you can use Mutual Information, which as hinted at, can be defined in terms of KL. – samthebest May 22, 2014 at 14:53

5

As an analogy, imagine a circle centred at the origin. Points are distributed evenly. Suppose a randomly-selected point is at (x1, x2). The Euclidean distance from the origin is ((x1)^2 + (x2)^2)^0.5

Now, imagine points evenly distributed over a sphere. That same point (x1, x2) will now probable be (x1, x2, x3). Since, in an even distribution, only a few points have one of the coordinates as zero, we shall assume that [x3 != 0] for our randomly-selected evenly-distributed point. Thus, our random point is most likely (x1, x2, x3) and not (x1, x2, 0).

The effect of this is: any random point is now at a distance of ((x1)^2 + (x2)^2 + (x3)^2)^0.5 from the origin of the 3-D sphere. This distance is larger than that for a random point near the origin of a 2-D circle. This problem gets worse in higher dimensions, which is why we choose metrics other than Euclidean dimensions to work with higher dimensions.

EDIT: There's a saying which I recall now: "Most of the mass of a higher-dimensional orange is in the skin, not the pulp", meaning that in higher dimensions **evenly** distributed points are more "near" (Euclidean distance) the boundary than the origin.

Side note: Euclidean distance is not TOO bad for real-world problems due to the 'blessing of non-uniformity', which basically states that for real data, your data is probably NOT going to be distributed evenly in the higher dimensional space, but will occupy a small clusted subset of the space. This makes sense intuitively: if you're measuring 100 quantities about humans like height, weight, etc, an even distribution over the dimension space just does not make sense, e.g. a person with (height=65 inches, weight=150 lbs, avg_calorie_intake=4000) which is just not possible in the real world.

Share Cite Edit Follow

edited Jan 6, 2016 at 5:25

answered Jan 6, 2016 at 5:12

Abhishek Divekar
764 1 9 17

1 If any future readers are interested in the "orange/pulp" quote, or the "blessing of non-uniformity" remark, both appear in "A few useful things to learn about machine learning," which is linked to in my answer on this thread. – Sycorax ♦ Jul 17, 2016 at 21:45

1

Another facet of this question is this:

Very often high dimensions in (machine-learning/statistical) problems are a result of over-constrained features.

Meaning the dimensions are NOT independent (or uncorrelated), but Euclidean metrics assume (at-least) un-correlation and thus may not produce best results

So to answer your question the number of "high dimensions" is related to how many features are inter-depedennt or redundant or over-constrained

Additionally: It is a theorem of [Csiszar (et al.), "Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems"] that Euclidean metrics are "natural" candidates for inference when the features are of certain forms:

> An attempt is made to determine the logically consistent rules for selecting a vector from any feasible set defined by linear constraints, when either all n-vectors or those with positive components or the probability vectors are permissible. Some basic postulates are satisfied if and only if the selection rule is to minimize a certain

function which, if a "prior guess" is available, is a measure of distance from the prior guess. Two further natural postulates restrict the permissible distances to the author's f-divergences and Bregman's divergences, respectively. As corollaries, axiomatic characterizations of the methods of least squares and minimum discrimination information are arrived at. Alternatively, the latter are also characterized by a postulate of composition consistency. As a special case, a derivation of the method of maximum entropy from a small set of natural axioms is obtained.

Share Cite Edit Follow edited Jul 9, 2020 at 6:41 answered May 22, 2014 at 23:39

Nikos M.
245 1 5

3 Euclidean metrics do not "assume... un-correlation". Euclidean distances work worst in high dimensions with uncorrelated variables. Consider the extreme case: you have very many dimensions that are all perfectly correlated, r=1, now your data are in fact uni-dimensional, & Euclidean distance works fine w/ uni-dimensional data. – gung - Reinstate Monica May 23, 2014 at 0:04

No i dont think so, Euclidean distance by definition assumes un-correllated data (except if using generalized Euclidean distance with correllation matrix) – Nikos M. May 23, 2014 at 0:18

Features with total correlation (r=1) is a trivial example and equivalent to a "trivial correlation matrix", but maybe i'm wrong – Nikos M. May 23, 2014 at 0:21

@gung You can interpret a Euclidean loss as a cross entropy loss of Gaussians with fixed unit isotropic variance matrix. I think this is a good point, but it could be better explained. – Neil G Jul 13, 2016 at 2:11

1 @NeilG, I have no idea what that even means. Consider 2 points on a plane: $(0,0)$ & $(1,1)$. The Euclidean distance between them is defined as: $d_E = \sqrt{\sum_j (x_{2j} - x_{1j})^2}$; here, $\sqrt{2}$. Now imagine more points exist on this plane, but that all lie along the line $X_1 = X_2$, which makes the correlation b/t the 2 variables 1. What is the Euclidean distance between those 2 points? It's still $\sqrt{2}$. Now imagine that the additional points are arranged uniformly st $\text{cor}(X_1, X_2) = 0$. What is the Euclidean distance between those 2 points? It's still $\sqrt{2}$. – gung - Reinstate Monica Jul 13, 2016 at 2:48

0

This paper may help you too "Improved sqrt-cosine similarity measurement" visit https://journalofbigdata.springeropen.com/articles/10.1186/s40537-017-0083-6 This paper explains why Euclidean distance is not a good metric in high dimensional data and what is the best replacement for Euclidean distance in high dimensional data. Euclidean distance is L2 norm and by decreasing the value of k in Lk norm we can alleviate the problem of distance in high dimensional data. You can find the references in this paper as well.

Share Cite Edit Follow

edited Apr 17, 2019 at 2:14

answered Apr 17, 2019 at 1:40

Sahar
33 6

6 Welcome to the site. We are trying to build a permanent repository of high-quality statistical information in the form of questions & answers. Thus, we're wary of link-only answers, due to linkrot. Can you post a full citation & a summary of the information at the link, in case it goes dead?
– gung - Reinstate Monica Apr 17, 2019 at 1:51