Curse of dimensionality

From Wikipedia, the free encyclopedia

The **curse of dimensionality** refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience.

There are multiple phenomena referred to by this name in domains such as numerical analysis, sampling, combinatorics, machine learning, data mining and databases. The common theme of these problems is that when the dimensionality increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance. In order to obtain a statistically sound and reliable result, the amount of data needed to support the result often grows exponentially with the dimensionality. Also organizing and searching data often relies on detecting areas where objects form groups with similar properties; in high dimensional data however all objects appear to be sparse and dissimilar in many ways which prevents common data organization strategies from being efficient.

The term *curse of dimensionality* was coined by Richard E. Bellman when considering problems in dynamic optimization.[1][2]

Contents

The "curse of dimensionality" depends on the algorithm

The "curse of dimensionality" is not a problem of high-dimensional data, but a joint problem of the data and the algorithm being applied. It arises when the algorithm does not scale well to high-dimensional data, typically due to needing an amount of time or memory that is exponential in the number of dimensions of the data.

When facing the curse of dimensionality, a good solution can often be found by changing the algorithm, or by pre-processing the data into a lower-dimensional form. For example, the notion of intrinsic dimension refers to the fact that any low-dimensional data space can trivially be turned into a higher-dimensional space by adding redundant (e.g. duplicate) or randomized dimensions, and in turn many high-dimensional data sets can be reduced to lower-dimensional data without significant information

loss. This is also reflected by the effectiveness of dimension reduction methods such as principal component analysis in many situations. Algorithms that are based on distance functions or nearest neighbor search can also work robustly on data having many spurious dimensions, depending on the statistics of those dimensions.

Curse of dimensionality in different domains

Combinatorics

In some problems, each variable can take one of several discrete values, or the range of possible values is divided to give a finite number of possibilities. Taking the variables together, a huge number of combinations of values must be considered. This effect is also known as the combinatorial explosion. Even in the simplest case of d binary variables, the number of possible combinations already is $O(2^d)$, exponential in the dimensionality. Naively, each additional dimension doubles the effort needed to try all combinations.

Sampling

There is an exponential increase in volume associated with adding extra dimensions to a mathematical space. For example, 10^2=100 evenly-spaced sample points suffice to sample a unit interval (a "1-dimensional cube") with no more than 10^{-2}=0.01 distance between points; an equivalent sampling of a 10-dimensional unit hypercube with a lattice that has a spacing of 10^{-2}=0.01 between adjacent points would require 10^{20}[=$(10^2)^{10}$] sample points. In general, with a spacing distance of 10^{-n} the 10-dimensional hypercube appears to be a factor of $10^{n(10-1)}$[=$(10^n)^{10}/(10^n)$] "larger" than the 1-dimensional hypercube, which is the unit interval. In the above example n=2: when using a sampling distance of 0.01 the 10-dimensional hypercube appears to be 10^{18} "larger" than the unit interval. This effect is a combination of the combinatorics problems above and the distance function problems explained below.

Optimization

When solving dynamic optimization problems by numerical backward induction, the objective function must be computed for each combination of values. This is a significant obstacle when the dimension of the "state variable" is large.

Machine learning

In machine learning problems that involve learning a "state-of-nature" (maybe an infinite distribution) from a finite number of data samples in a high-dimensional feature space with each feature having a number of possible values, an enormous amount of training data are required to ensure that there are several samples with each combination of values. With a fixed number of training samples, the predictive power reduces as the dimensionality increases, and this is known as the *Hughes effect*[3] or *Hughes phenomenon* (named after Gordon F. Hughes).[4][5]

Bayesian statistics

The curse of dimensionality has often been a difficulty with Bayesian statistics, for which the posterior distributions often have many parameters.

However, this problem has been largely overcome by the advent of simulation-based Bayesian inference, especially using Markov chain Monte Carlo methods, which suffices for many practical problems. Of course, simulation-based methods converge slowly and therefore are not a panacea for high-dimensional problems.

Distance functions

When a measure such as a Euclidean distance is defined using many coordinates, there is little difference in the distances between different pairs of samples.

One way to illustrate the "vastness" of high-dimensional Euclidean space is to compare the proportion of an inscribed hypersphere with radius r and dimension d, to that of a hypercube with edges of length $2r$. The volume of such a sphere is: $\frac{2r^d\pi^{d/2}}{d\,\Gamma(d/2)}$. The volume of the cube would be: $(2r)^d$. As the dimension d of the space increases, the hypersphere becomes an insignificant volume relative to that of the hypercube. This can clearly be seen by comparing the proportions as the dimension d goes to infinity:

$\frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \to 0$ as $d \to \infty$. Furthermore, the distance between the center and the corners is $r\sqrt{d}$, which increases without bound for fixed r.

In this sense, nearly all of the high-dimensional space is "far away" from the centre. To put it another way, the high-dimensional unit hypercube can be said to consist almost entirely of the "corners" of the hypercube, with almost no "middle".

This also helps to understand the chi-squared distribution. Indeed, the (non-central) chi-squared distribution associated to a random point in the interval *[-1,1]* is the same as the distribution of the length-squared of a random point in the *d*-cube. By the law of large numbers, this distribution concentrates itself in a narrow band around *d* times the standard deviation squared (σ^2) of the original derivation. This illuminates the chi-squared distribution and also illustrates that most of the volume of the *d*-cube concentrates near the surface of a sphere of radius $\sqrt{d}\sigma$.

A further development of this phenomenon is as follows. Any fixed distribution on *R* induces a product distribution on points in R^d. For any fixed *n*, it turns out that the minimum and the maximum distance between a random reference point *Q* and a list of *n* random data points $P_1,\ldots,P_n$ become indiscernible compared to the minimum distance:[6]

$$\lim_{d\to\infty} E\left(\frac{\mathrm{dist}_{\max}(d) - \mathrm{dist}_{\min}(d)}{\mathrm{dist}_{\min}(d)}\right) \to 0.$$

This is often cited as distance functions losing their usefulness (for the nearest-neighbor criterion in feature-comparison algorithms, for example) in high dimensions. However, recent research has shown this to only hold in the artificial scenario when the one-dimensional distributions *R* are independent and identically distributed.[7] When attributes are correlated, data can become easier and provide higher distance contrast and the signal-to-noise ratio was found to play an important role, thus feature selection should be used.[7]

Nearest neighbor search

The effect complicates nearest neighbor search in high dimensional space. It is not possible to quickly reject candidates by using the difference in one coordinate as a lower bound for a distance based on all the dimensions.[8][9]

However, it has recently been observed that the mere number of dimensions does not necessarily result in difficulties,[10] since *relevant* additional dimensions can also increase the contrast. In addition, for the resulting ranking it remains useful to discern close and far neighbors. Irrelevant ("noise") dimensions, however, reduce the contrast in the manner described above. In time series analysis, where the data are inherently high-dimensional, distance functions also work reliably as long as the signal-to-noise ratio is high enough.[11]

k-nearest neighbor classification

Another effect of high dimensionality on distance functions concerns *k*-nearest neighbor (*k*-NN) graphs constructed from a data set using a distance function. As the dimension increases, the indegree distribution of the *k*-NN digraph becomes skewed with a peak on the right because of the emergence of a disproportionate number of **hubs**, that is, data-points that appear in many more *k*-NN lists of other data-points than the average. This phenomenon can have a considerable impact on various techniques for classification (including the *k*-NN classifier), semi-supervised learning, and clustering,[12] and it also affects information retrieval.[13]

Anomaly detection

In a recent survey, Zimek et al. identified the following problems when searching for anomalies in high-dimensional data:[7]

1. Concentration of scores and distances: derived values such as distances become numerically similar
2. Irrelevant attributes: in high dimensional data, a significant number of attributes may be irrelevant
3. Definition of reference sets: for local methods, reference sets are often nearest-neighbor based
4. Incomparable scores for different dimensionalities: different subspaces produce incomparable scores
5. Interpretability of scores: the scores often no longer convey a semantic meaning
6. Exponential search space: the search space can no longer be systematically scanned
7. Data snooping bias: given the large search space, for every desired significance an hypothesis can be found
8. Hubness: certain objects occur more frequently in neighbor lists than others.

Many of the analyzed specialized methods tackle one or another of these problems, but there remain many open research questions.

See also

- Bellman equation
- Backwards induction
- Cluster analysis
- Clustering high-dimensional data
- Combinatorial explosion
- Concentration of measure
- Dimension reduction

- Dynamic programming
- Fourier-related transforms
- High-dimensional space
- Linear least squares
- Multilinear PCA
- Multilinear subspace learning
- Principal component analysis
- Quasi-random
- Singular value decomposition
- Stereology
- Time series
- Wavelet

References

1. Richard Ernest Bellman; Rand Corporation (1957). *Dynamic programming* (http://books.google.it/books?id=wdtoPwAACAAJ). Princeton University Press. ISBN 978-0-691-07951-6.,
Republished: Richard Ernest Bellman (2003). *Dynamic Programming* (http://books.google.com/books?id=fyVtp3EMxasC). Courier Dover Publications. ISBN 978-0-486-42809-3.
2. Richard Ernest Bellman (1961). *Adaptive control processes: a guided tour* (http://books.google.com/books?id=POAmAAAAMAAJ). Princeton University Press.
3. Oommen, T. .; Misra, D. .; Twarakavi, N. K. C.; Prakash, A. .; Sahoo, B. .; Bandopadhyay, S. . (2008). "An Objective Analysis of Support Vector Machine Based Classification for Remote Sensing". *Mathematical Geosciences* **40** (4): 409. doi:10.1007/s11004-008-9156-6 (https://dx.doi.org/10.1007%2Fs11004-008-9156-6).
4. Hughes, G.F. (January 1968). "On the mean accuracy of statistical pattern recognizers" (http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=1054102). *IEEE Transactions on Information Theory* **14** (1): 55–63. doi:10.1109/TIT.1968.1054102 (https://dx.doi.org/10.1109%2FTIT.1968.1054102).
5. Not to be confused with the unrelated, but similarly named, *Hughes effect in electromagnetism* (named after Declan C. Hughes (http://spiedl.aip.org/vsearch/servlet/VerityServlet?KEY=SPIEDL&possible1=Hughes%2C+Declan+C.&possible1zone=author&maxdisp=25&smode=strresults&pjournals=OPEGAR%2CJBOPFO%2CPSISDG%2CJEIME5%2CJMMMGF%2CJARSC4%2CJNOACQ&deliveryType=spiedl&aqs=true)) which refers to an asymmetry in the hysteresis curves of laminated cores made of certain magnetic materials, such as permalloy or mu-metal, in alternating magnetic fields.
6. Beyer, K.; Goldstein, J.; Ramakrishnan, R.; Shaft, U. (1999). "When is "Nearest Neighbor" Meaningful?". *Proc. 7th International Conference on Database Theory - ICDT'99*. LNCS **1540**: 217–235. doi:10.1007/3-540-49257-7_15 (https://dx.doi.org/10.1007%2F3-540-49257-7_15). ISBN 978-3-540-65452-0.
7. Zimek, A.; Schubert, E.; Kriegel, H.-P. (2012). "A survey on unsupervised outlier detection in high-dimensional numerical data". *Statistical Analysis and Data Mining* **5** (5): 363–387. doi:10.1002/sam.11161 (https://dx.doi.org/10.1002%2Fsam.11161).
8. Marimont, R.B.; Shapiro, M.B. (1979). "Nearest Neighbour Searches and the Curse of Dimensionality" (http://imamat.oxfordjournals.org/content/24/1/59.short). *IMA J Appl Math* **24** (1): 59–70. doi:10.1093/imamat/24.1.59 (https://dx.doi.org/10.1093%2Fimamat%2F24.1.59).
9. Chávez, Edgar; Navarro, Gonzalo; Baeza-Yates, Ricardo; Marroquín, José Luis (2001). "Searching in Metric Spaces". *ACM Computing Surveys* **33** (3): 273–321. doi:10.1145/502807.502808 (https://dx.doi.org/10.1145%2F502807.502808). CiteSeerX: 10.1.1.100.7845 (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.100.7845).
10. Houle, M. E.; Kriegel, H. P.; Kröger, P.; Schubert, E.; Zimek, A. (2010). *Can Shared-Neighbor Distances Defeat the Curse of Dimensionality?* (http://www.dbs.ifi.lmu.de/~zimek/publications/SSDBM2010/SNN-SSDBM2010-preprint.pdf) (PDF). Scientific and Statistical Database Management. Lecture Notes in Computer Science **6187**. p. 482. doi:10.1007/978-3-642-13818-8_34 (https://dx.doi.org/10.1007%2F978-3-642-13818-8_34). ISBN 978-3-642-13817-1.
11. Bernecker, T.; Houle, M. E.; Kriegel, H. P.; Kröger, P.; Renz, M.; Schubert, E.; Zimek, A. (2011). *Quality of Similarity Rankings in Time Series*. Symposium on Spatial and Temporal Databases. Lecture Notes in

Computer Science **6849**. p. 422. doi:10.1007/978-3-642-22922-0_25 (https://dx.doi.org/10.1007%2F978-3-642-22922-0_25). ISBN 978-3-642-22921-3.
12. Radovanović, Miloš; Nanopoulos, Alexandros; Ivanović, Mirjana (2010). "Hubs in space: Popular nearest neighbors in high-dimensional data" (http://www.jmlr.org/papers/volume11/radovanovic10a/radovanovic10a.pdf) (PDF). *Journal of Machine Learning Research* **11**: 2487–2531.
13. Radovanović, M.; Nanopoulos, A.; Ivanović, M. (2010). *On the existence of obstinate results in vector space models*. 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10. p. 186. doi:10.1145/1835449.1835482 (https://dx.doi.org/10.1145%2F1835449.1835482). ISBN 9781450301534.

Categories: Numerical analysis | Dynamic programming | Machine learning

---