The Curse of Dimensionality

Raúl Rojas

February 15, 2015

Abstract

A k-nearest neighbors classifier has a simple structure and can help to bootstrap a classification project with little effort. However, the "nearest neighbor" concept breaks down when the dimensionality of the feature space is high. In a high-dimensional space most points, taken from a random finite set of N data points inside a finite volume, are far away from each other.

The "curse of dimensionality" refers to the problem of finding structure in data embedded in a highly dimensional space. The more features we have, the more data points we need in order to fill space.

Take for example a hypercube with side length equal to 1, in an n-dimensional space. The volume of the hypercube ist 1. If we want to allocate that volume among N smaller cubes (each containing a data point) distributed more or less homogeneously in the n-dimensional hypercube, each small cube will have a volume equal to $1/N$. Their side length d would be

$$d = \left(\frac{1}{N}\right)^{(1/n)}$$

For a finite N, d converges to 1 when n goes to infinity. That is, the new smaller cubes have each "almost" the same volume as the bigger cube. In

infinite dimensional space you can put N cubes of volume 1 inside a cube of volume 1! Keeping the volume of the smaller cubes fixed to, let's say, $d = 0.1$, requires an exponentially growing number of data points N, when the dimension increases.

Another way of making this computation would be to take a sphere of radius 1, in n-dimensional space. Its volume is $V = k_n 1^n$, where k_n is a proportionality constant which depends on the value of n. If we want to take N spheres of radius d to cover the same volume as the sphere of radius 1, we need spheres of volume $k_n d^n$ each. Making N constant, we require a radius

$$d = \left(\frac{k_n}{N k_n}\right)^{(1/n)} = \left(\frac{1}{N}\right)^{(1/n)},$$

just as in the previous case. The radius of the smaller spheres approaches the radius of the larger sphere, when the dimension goes to infinity.

A more intuitive result can be obtained if we think of N data points spread uniformly in a sphere of radius 1. If we ask what is the distance of the origin to its nearest neighbor, this distance will change according to the data set that we take. We can perform many experiments with randomly distributed data points inside the sphere, and then ask what is the median of the distance of the origin to its nearest neighbor. If we call this median d, half of the time the nearest neighbor will be farther away than d and half of the time it will be nearer than the distance d. If the nearest neighbor is farther away than d, then *all* N data points are farther away than the distance d from the origin. That is, they are all in the region between a sphere of radius d and the sphere of radius 1. The probability of this happening is $1/2$, since d is the median of the nearest neighbors. The volume between the sphere of radius 1 and the sphere of radius d is $k_n 1^n - k_n d^n$. The probability that all N data points, generated independently, fall between both spheres is then

$$\frac{1}{2} = \left(\frac{k_n - k_n d^n}{k_n}\right)^N$$

From this we conclude that for the expected value of the median d the following expression is valid

$$d = \left(1 - \left(\frac{1}{2}\right)^{1/N}\right)^{1/n}$$

Since for any positive number x, such that $0 < x < 1$, the function $x^{1/n}$ converges to 1 when n goes to infinity, we conclude that the median d of the nearest neighbor distances converges to 1. That is, all points seem to be far away from the origin, sitting on the surface of the sphere of dimension n.

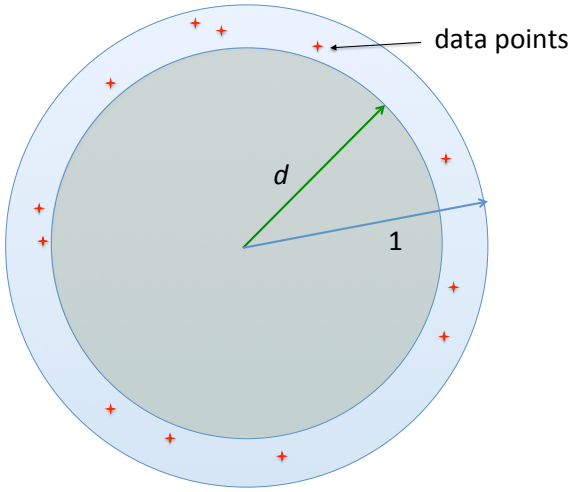

Figure 1: In a highly dimensional space all homogeneously distributed data points seem to be near the shell of the sphere

There is no easy way out of the curse of dimensionality. The best approach is trying to generalize the "shape" of the data point clouds, computing approximations to the distribution functions (for example Gaussians), or putting "fences" between the different classes, as decision surfaces do.

References

[1] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, New York, 2001.