Curse of Dimensionality

Dror Baron

This supplement adds information about the curse of dimensionality.

Curse of dimensionality: The nearest neighbors (NN) method no doubt seemed quite appealing based on our numerical results (see the binary classification example adapted from Chapter 2 in Hastie et al. and our corresponding Matlab implementation). Unfortunately, NN suffers from what is called the curse of dimensionality. In particular, as the dimension p grows, the number of data points N required to map out the space thoroughly grows exponentially with p, and realistically speaking there is a scarcity of data points in high dimensions.

To see why this curse of dimensionality occurs, consider that the density of points in p dimensions scales as $N^{1/p}$. For example, if $N = 10^6$ (possibly not "big data" but still a substantial data set), then even a relatively modest $p = 20$ yields $N^{1/p} = (10^6)^{0.05} \approx 2$. That is, along each dimension we can expect to sample roughly 2 points. If points are spread around uniformly in the $p = 20$ dimensional hypercube $[-1, +1]^{20}$, then the nearest "neighbor" will typically be roughly one unit away in each dimension, and it probably doesn't seem like much of a neighbor any more.