Insights, resources, and application of high-dimensional data

BY YALIN LIAO

Before, I knew our intuition from two or three dimensional space may be wrong in high dimensional space. For example, if many samples are uniformly placed inside a unit hypercube in a high-dimensional space, it is proven that most of these samples are closer to a face of hypercube than to their nearest neighbors[1]. But I don't understand why. Later, I learned that the curse of dimmensionality is that to achieve an error rate the minimum number of training samples required by some algorithms increases exponentailly with the dimensionality of the space. That is to say, leaning efficiency could expentially descrease as the dimensionality increases. After reading these papers, I have more perspectives to look at the curse of dimensionality.

The curse of dimensionality can be understood from the underlying metric. In paper [2], authors systematically study the distance concentration effect and conclude that the fractional distance metric provides more meaningful results in clustering algorithms. But the introduction says "in high dimensional space the data becomes sparse". I feel confused at the beginning as one the one hand the data are distance concentrated and on the other hand they are sparse. Recall the example in the first paragraph, it is more understandable. As most samples are located in the face (corners), they are naturally sparse. Since they are mostly in corners of the unit hypercube in $\mathbb{R}^d$, with high probability the distance of two points is roughly equal to $\sqrt{d}$ (in the sense of L_∞ norm). This inituitivelly explains the distance concentration phenomenon and there is no conflict between data sparse and data distance concentration in high dimensional space. The distance concentration does not means samples in high dimensional data centered in some points but the distance of any two points almost share same distance in high dimensional space. Therefore, clustering high-dimensional data points using distance metrics is very challenging.

Why are most samples close to a face of hypercube than to their nesrest neighbors? Consing L_∞ norm, understanding it is simple. A hypercube having length $2r$ centered at the origin in $\mathbb{R}^d$ is a ball r-radius ball centered at the origin using L_∞ norm, i.e., $B_\infty(0,r)$. Let ε be an arbitray small positive real-valued number ($0<\varepsilon<r$). Then[3]

$$\frac{V_d(B_\infty(0,r))}{V_d(B_\infty(0,r-\varepsilon))}=\frac{r^d}{(r-\varepsilon)^d}=\left(1+\frac{\varepsilon}{r-\varepsilon}\right)^d\to\infty \text{ as } d\to\infty$$

That means when d is very large,

$$V_d(B_\infty(0,r))\gg V_d(B_\infty(0,r-\varepsilon))$$

no matter how small ε is. Further, we have

$$V_d(B_\infty(0,r))-V_d(B_\infty(0,r-\varepsilon))\gg V_d(B_\infty(0,r-\varepsilon))$$

since

$$\frac{V_d(B_\infty(0,r))-V_d(B_\infty(0,r-\varepsilon))}{V_d(B_\infty(0,r-\varepsilon))}=\left(1+\frac{\varepsilon}{r-\varepsilon}\right)^d-1\to\infty \text{ as } d\to\infty$$

1. Machine Learing Fundamentals by Hui Jiang.
2. On the Surprising Behavior of Distance Metrics in High Dimensional Space
3. Use V_d to denote volume of balls in $\mathbb{R}^d$.

That is to say, the mass of the ball is centered around the thin bell close to the surface (width ε). If all data points are drawn from the uniform distribution over $B_\infty(0,r)$, then

$$P(B_\infty(0,r))=1 \text{ and } \frac{P(B_\infty(0,r-\varepsilon))}{P(B_\infty(0,r))}=\frac{V_d(B_\infty(0,r-\varepsilon))}{V_d(B_\infty(0,r))}=\left(1-\frac{\varepsilon}{r}\right)^d$$

and thus $P(B_\infty(0,r-\varepsilon))=\left(1-\frac{\varepsilon}{r}\right)^d$ and $P(B_\infty(0,r)\backslash B_\infty(0,r-\varepsilon))=1-\left(1-\frac{\varepsilon}{r}\right)^d$ which goes to 1 if d goes to ∞. This explains why samples in high dimensional space are closer to the face only using high school math. Can we explain the case L_2 in a similar way? The answer is yes, but we need the following extral knowldge from Wikipedia[4]

$$V_d(B_2(0,r))=\frac{2(2\pi)^{(n-1)/2}}{d!!}r^d$$

where $B_2(0,r)$ denotes the ball centered at 0 with radius r. Similarly, we have

$$P(B_2(0,r)\backslash B_2(0,r-\varepsilon))=1-\left(1-\frac{\varepsilon}{r}\right)^d\to 1 \text{ as } d\to\infty$$

if all samples are drawn from the uniform distribution over $B_2(0,r)$. Again, we see with high probability samples in high dimensional space are concentrated on the hyperball surface. This coincides with the explainations - the choice of a certain radius ε to select a neighborhood (a fundamental step in many outlier detection methods) is notoriously rather sensitive to dimensionality; small changes in the radius may decide whether everything or nothing is selected in a given dimensionality from [5]. Note there is a difference between L_2 and L_∞ on the volume of balls in high dimensional space

$$V_d(B_2(0,r))=\frac{2(2\pi)^{(n-1)/2}}{d!!}r^d\to 0 \text{ as } d\to\infty, \forall r>0$$

vs

$$V_d(B_\infty(0,r))=r^d\to\begin{cases}0 & 0<r<1\\ 1 & r=1\\ \infty & r>1\end{cases} \text{ as } d\to\infty$$

Simply saying, the volume of L_2 ball shrinks to zero no matter how large its radius r is[6] while that of L_∞ ball depends on its radius.

Now we turn to understand the theorem to describle the distance concentration. The theorem is summarized as follows:

$$\lim_{d\to\infty}\text{var}\left(\frac{\|X_d\|_k}{\mathbb{E}[\|X_d\|_k]}\right)=0\Rightarrow\frac{\max_N\{\|X_d\|_k\}-\min_N\{\|X_d\|_k\}}{\min_N\{\|X_d\|_k\}}\to^p 0 \text{ as } d\to\infty$$

How do we understand the formula intuituvely? As $\{X_d\}$ are i.i.d, $\{\|X_d\|\}$ are i.i.d. Let $\mu=\mathbb{E}[\|X_d\|_k]$. Then

$$\begin{aligned}\text{var}\left(\frac{\|X_d\|_k}{\mathbb{E}[\|X_d\|_k]}\right) &= \mathbb{E}\left[\left(\frac{\|X_d\|_k}{\mathbb{E}[\|X_d\|_k]}-\mathbb{E}\left[\frac{\|X_d\|_k}{\mathbb{E}[\|X_d\|_k]}\right]\right)^2\right]\\ &= \mathbb{E}\left[\left(\frac{\|X_d\|_k}{\mu}-\mathbb{E}\left[\frac{\|X_d\|_k}{\mu}\right]\right)^2\right]\\ &= \mathbb{E}\left[\left(\frac{\|X_d\|_k}{\mu}-1\right)^2\right]\to 0 \text{ as } d\to\infty\end{aligned}$$

4. Volume of an n ball

5. A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data.

6. The survey paper uses the volume of L_2 ball vanishes in high dimensional space to demonsrate the sensitivity of the radius choice. Maybe it's more reasonable to use the volume ratio as it is vanishing for both L_2 and L_∞, while that the volume of L_∞ ball vanishes does not hold. (May not correct for $L_p, p>2$ norm)

That is to say, $\{\| X_d\|_k\}$ are concentrated at μ or with high probability $\| X_d\|_k \approx \mu$ so as to

$$\max_N \{\| X_d\|_k\} \approx \mu$$

$$\min_N \{\| X_d\|_k\} \approx \mu$$

Therefore,

$$\frac{\max_N \{\| X_d\|_k \} - \min_N \{\| X_d\|_k \}}{\min_N \{\| X_d\|_k \}} \approx \frac{\mu - \mu}{\mu} = 0$$

In one word, since $\{\| X_d\|_k\}$ are highly centered around μ (if variance is zero, all points are identical) the difference between the maximum distance and the minimum distance is close to zero. The rigorous proof is built on the law of large number theory[7] but it follows the same idea using Slutsky's theorem (continuous functions preserve probability convergence proproperty).

7. When Is "Nearest Neighbor" Meaningful?