Machine Learning

Nirdosh Bhatnagar

High-Dimensional Data: An Introduction

1. Introduction

– Especial care is necessary in analysing high-dimensional data points, because their characteristics generally defy normal intuition.

– The behavior of high-dimensional data is typically nonintuitive because we have been conditioned to the fact that the number of available data points is much larger than the number of attributes (dimension) of a data point.

2. **Curse of Dimensionality**

Suppose that we are given a set of $n \in \mathbb{P}$ data points, and each data point is a vector of size $t \in \mathbb{P}$.

- It is generally assumed while analysing data sets that both n and t are fixed.
- It is also assumed in several cases that $n > t$, and occasionally that $n \to \infty$.

The analysis techniques used in such cases are generally not useful for:

- Finite number of data points n;
- and $t \gg n$,
- and sometimes $t > n$.

- This shortcoming of classical techniques of analysing data sets is called the "curse of dimensionality."
- We shall also see that high-dimensional data is also a "blessing."
- The phrase "curse of dimensionality" was coined by the inventor of dynamic programming, Richard Bellman (1920-1984).
- He used this phrase while describing the computational complexity of dynamic programming algorithms which have an exponential dependence upon the dimension of the state of the underlying dynamical system.

3. Applications

Some of the areas in which the curse of dimensionality occurs are:

- Optimization
- Function approximation
- Numerical integration
- Machine learning
- Image processing
- Statistical estimation
- Data mining.

- Consider a Cartesian grid of spacing ϵ on a unit hypercube in dimension t, where $0 < \epsilon \ll 1$.
- A hypercube is simply a generalization of a cube to higher dimensions.
- The number of points in this hypercube is in $O\left(\epsilon^{-t}\right)$.
- This number can be extremely high. Therefore, in optimization, function approximation, and numerical integration problems, the number of function evaluations or searches required is in $O\left(\epsilon^{-t}\right)$.

4. Empty Space Property

– The number of high-dimensional data points required to estimate the parameters of the associated model to a specified accuracy is usually very high.

– As the number of data points in such cases is typically very sparse, we have the so-called *empty space phenomenon.*

– This is another manifestation of the curse of high dimensionality.

– The empty space phenomenon, or property, can also be studied by considering the volumes of t-dimensional hypersphere and hypercube, where $t \in \mathbb{P}$.

– A hypersphere is a generalization of a sphere to higher-dimensional spaces.

– The hypersphere of radius r in a t-dimensional space is the set of points $\mathcal{H}_r$, where

$$\mathcal{H}_r = \begin{array}{c} \{(x_1, x_2, \ldots, x_t) \mid \\ \sum_{i=1}^{t} x_i^2 \leq r^2,\ x_i \in \mathbb{R},\ 1 \leq i \leq t,\ r \in \mathbb{R}^+ \end{array} \Big\}$$

– Let the volume of the t-dimensional hypersphere of radius $r \in \mathbb{R}^+$ be $V_t(r)$.

– Also let the volume of the corresponding circumscripted hypercube be $C_t(r)$.

That is, the length of the side of this hypercube is equal to $2r$.

– Thus

$$V_t(r) = \frac{\pi^{t/2}}{\Gamma(t/2+1)} r^t, \quad \text{and} \quad C_t(r) = (2r)^t$$

where $\Gamma(\cdot)$ is the gamma function. Observe that

$$\lim_{t \to \infty} \frac{V_t(r)}{C_t(r)} \to 0$$

The gamma function is a generalization of the factorial of an integer for nonintegral values. The gamma function $\Gamma(a), a \in \mathbb{C}$ is defined as

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt, \qquad \mathrm{Re}(a) > 0$$

Note the recursion $\Gamma(a+1) = a\Gamma(a)$. It can also be shown that $\Gamma(1/2) = \sqrt{\pi}$. The gamma function reduces to the factorial function for integer values of its argument. That is, $\Gamma(n+1) = n!$, $\forall\ n \in \mathbb{N}$.

– The above result implies that as the dimension t increases, the volume of the hypersphere becomes insignificant when compared to the volume of the corresponding circumscripting hypercube.

– Therefore if $r = 1/2$, the volume of the hypercube is unity; and the volume of the hypersphere of radius $r = 1/2$ tends to 0, as t tends to infinity.

– We next demonstrate that most of the volume of the hypersphere is near its surface in a thin shell (crust). Consider a shell of thickness ϵr, where $0 < \epsilon \ll 1$.

– The ratio of the volume of the shell and the hypersphere is

$$\frac{V_t(r) - V_t(r(1-\epsilon))}{V_t(r)} = \frac{\left\{1^t - (1-\epsilon)^t\right\}}{1^t}$$

– For fixed value of ϵ, and as $t \to \infty$, the above ratio tends to unity.

– Thus the shell of the hypersphere contains most of the volume of the hypersphere for large values of t.

5. **Sensitivity to the Distance Metric**

– For a fixed set of data points, let $d_{\max}$ and $d_{\min}$ be the maximum and minimum distances respectively between the given set of high-dimensional data points. Then

$$\lim_{t\to\infty} \frac{d_{\max} - d_{\min}}{d_{\min}} \to 0$$

– This result implies that in a set of high-dimensional data points, distance function looses its sensitivity, as the difference between the maximum and minimum distances becomes relatively negligible.

– For example, this property makes nearest-neighbor data mining techniques difficult to address in high-dimensional spaces.

6. Isotropic Gaussian Distribution

– Consider a set of t-dimensional data points. In these data points, values of each attribute have a Gaussian distribution.

– Let X_i be the Gaussian random variable associated with the ith attribute, where $1 \leq i \leq t$.

– Further $X_i \sim \mathcal{N}\left(0, \sigma^2\right)$ for $1 \leq i \leq t$; and these t random variables are mutually independent of each other.

– Let $x = (x_1, x_2, \ldots, x_t)$ be a data point, where $x_i \in \mathbb{R}$, $1 \leq i \leq t$.

– Also let the corresponding multivariate isotropic probability density function be $p(x)$. Then

$$\begin{aligned} p(x) &= \frac{1}{(2\pi)^{t/2}\,\sigma^t} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{t} x_i^2\right\}, \\ x_i &\in \mathbb{R},\ 1 \leq i \leq t \end{aligned}$$

This density function is called the t-dimensional spherical normal distribution.

– Let $Z = \sum_{i=1}^{t} X_i^2$, and $R = \sqrt{Z}$. The expected value of the sum of squared attribute values is

$$\mathcal{E}(Z) = t\sigma^2$$

For large values of t, the sum of squared values of x_i's is concentrated about its mean.

– However, observe that the density function has a maximum value at the origin.

– Also the probability density function of the random variable R is

$$f_R(r) = \begin{cases} 0, & r \in (-\infty, 0] \\ \dfrac{2r^{t-1}}{\Gamma\left(\frac{t}{2}\right)(2\sigma^2)^{t/2}} \exp\left\{-\frac{r^2}{2\sigma^2}\right\}, & r \in (0, \infty) \end{cases}$$

– This density function peaks at $r = \sigma\sqrt{t-1} \triangleq r_0$.

– For large values of t, that is for high-dimensional data, $f_R(r)$ is negligible for $r \in (0, r_0)$.

– Further, a relatively large part of the area under the function $f_R(\cdot)$ is concentrated in the interval $[r_0, r_0 + \alpha)$, where $\alpha \in O(\sigma)$.

– That is, the function $f_R(\cdot)$ is concentrated farther away from the origin.

7. **Blessings**

– High-dimensional data points also come with their share of blessings.

– It is asserted that, as $t \to \infty$, the concentration of the measure phenomenon can be conveniently described by certain asymptotic methods.