KTH Matematik

Lecture sf2935: Curse of Dimensionality: High Dimensional Feature Spaces and Nearest Neighbour and Nearest Neighbour Regression

Timo Koski

1 Introduction

Due to the development of technology in various fields it is now a commonplace to find massive complex data with almost unlimited number of features [8]. In [7, pp 238−239] high-dimensional means data where the number of features, say d, exceeds the number of samples, say n.

Curse of dimensionality describes here the problems caused by the exponential increase in volume associated with adding extra dimensions to a (mathematical) space, or, 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.There are phenomena that we do not encounter in settings such as the three-dimensional physical space of our everyday experience.

This lecture dels with the curse of dimensionality in statistical learning using some of the statements [1]. The main topic is to examine the effect of high dimensions on nearest neighbor and nearest neighbor regression. This will be done by letting $d \to +\infty$, while keeping n fixed.

The phrase 'curse of dimensionality' was originated by the American mathematician and engineer Richard E. Bellman, when dealing with optimization (dynamic programming).

2 Preliminaries

2.1 Notations

We are dealing with $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. The **distance** (or metric) between $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^d$ is

$$\| \mathbf{x} - \mathbf{y} \| = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}. \tag{2.1}$$

We set

$$D(\mathbf{x}) = \| \mathbf{x} - \mathbf{0} \|^2 = \sum_{i=1}^{d} x_i^2 = \mathbf{x}^T \mathbf{x}, \tag{2.2}$$

which is the squared distance from $\mathbf{x}$ to the origin.

Hyperball

A hyperball $\mathrm{sp}^d(\mathbf{x}, r)$ in $\mathbb{R}^d$, $r > 0$, is

$$\mathrm{sp}^d(\mathbf{x}, r) = \{\mathbf{y} | \parallel \mathbf{y} - \mathbf{x} \parallel \leq r\}. \tag{2.3}$$

We can visualize this as a sphere centered at $\mathbf{x}$ with the radius r.

2.2 Probability

We recall/summarize briefly some tools of probability calculus, see [6].

i) Convergence in probability

A sequence of random vectors $\mathbf{X}_1, \mathbf{X}_1, \ldots, \mathbf{X}_n, \ldots,$ **converges in probability** to $\mathbf{X} \in \mathbb{R}^d$ if and only if

$$\lim_{n \to +\infty} \mathcal{P}\left(\parallel \mathbf{X}_n - \mathbf{X} \parallel > \varepsilon\right) \to 0 \tag{2.4}$$

for any $\varepsilon > 0$, as $n \to +\infty$.

If $\mathbf{X}$ is a constant $\mathbf{a}$ (i.e. not a random variable), we write

$$\mathbf{X}_n \xrightarrow{\mathcal{P}} \mathbf{a} \quad \text{as } n \to +\infty. \tag{2.5}$$

ii) Cramér-Slutzky Theorem part 1)

If $\mathbf{X}_n \xrightarrow{\mathcal{P}} \mathbf{a}$ as $n \to +\infty$ and $g(\mathbf{x})$ is a continuous real valued function at $\mathbf{a} \in \mathbb{R}^d$, then

$$g(\mathbf{X}_n) \xrightarrow{\mathcal{P}} g(\mathbf{a}), \quad \text{as } n \to +\infty. \tag{2.6}$$

iii) Cramér-Slutzky Theorem, part 2)

Here $\mathbf{X}_n$ and $\mathbf{Y}_n$ are real valued, $\in \mathbb{R}$, random variables.

If $\mathbf{X}_n \xrightarrow{\mathcal{P}} a \in \mathbb{R}$, as $n \to +\infty$, and and $\mathbf{Y}_n \xrightarrow{\mathcal{P}} b \in \mathbb{R}$, $b \neq 0$, as $n \to +\infty$, then

$$\frac{\mathbf{X}_n}{\mathbf{Y}_n} \xrightarrow{\mathcal{P}} \frac{a}{b}, \quad \text{as } n \to +\infty. \tag{2.7}$$

iv) Chebysjev's Inequality

$\mathbf{X}$ is a real valued, $\mathbf{X} \in \mathbb{R}$, random variable that has a mean $E[\mathbf{X}] = \mu$ and a finite variance , $\mathrm{Var}[\mathbf{X}] = \sigma^2$. Then

$$\mathcal{P}(|\mathbf{X} - \mu| > \varepsilon) \leq \frac{1}{\varepsilon}\sigma^2. \tag{2.8}$$

3 When is nearest neighbour meaningful?

Suppose that $\mathbf{X}_l$ are n I.I.D. vector valued $\mathbf{X}_l = (X_{1l}, X_{2l}, \ldots, X_{dl})$ random variables and

$$D_l \overset{\text{def}}{=} D(\mathbf{X}_l) = \sum_{i=1}^{d} X_{il}^2, l = 1, 2, \ldots, n. \tag{3.1}$$

We assume that

$$E[D_l] = d, \mathrm{Var}[D_l] = 2d. \tag{3.2}$$

Example 3.1 $X_{1l}, X_{2l}, \ldots, X_{dl}$ are I.I.D. $\sim N(0,1)$. Then, see [6, p. 283], $D_l = \sum_{i=1}^{d} X_{il}^2 \sim \chi^2(d)$ (chi-squared distribution) and thus $E[D_l] = d$, $\mathrm{Var}[D_l] = 2d$.

The formulas in (3.2) hold also, e.g., if every $\mathbf{X}_l$ has a multivariate skew normal distribution.

■

We set also

$$D_{\min} = \min_{1 \leq l \leq n} D_l, \quad D_{\max} = \min_{1 \leq l \leq n} D_l.$$

Then we have the following theorem due to [1]. Note that d is the dimension of $\mathbb{R}^d$.

Theorem 3.2 *If (3.2) holds, then for any* $\varepsilon > 0$

$$\lim_{d\to+\infty} \mathcal{P}\left(D_{\max} \leq (1+\varepsilon) D_{\min}\right) = 1. \tag{3.3}$$

Proof: Set

$$\zeta_l = \frac{D_l}{E\left[D_l\right]} = \frac{D_l}{d}$$

Then

$$E\left[\zeta_l\right] = 1, \quad \mathrm{Var}\left[\zeta_l\right] = \frac{1}{d^2}\mathrm{Var}\left[D_l\right] = \frac{2}{d},$$

where we used (3.2). Then Chebysjev$'$s inequality (2.8) tells that for every l

$$\mathcal{P}\left(|\zeta_l - 1| > \varepsilon\right) \leq \frac{1}{\varepsilon}\mathrm{Var}\left[\zeta_l\right] = \frac{2}{\varepsilon \cdot d}. \tag{3.4}$$

Hence we have that for every l, $\zeta_l \xrightarrow{\mathcal{P}} 1$, as $d \to +\infty$.

Convergence in probability implies convergence in distribution. Since ζ_l are independent, there is joint convergence in distribution,

$$(\zeta_1, \ldots, \zeta_n) \xrightarrow{d} \underbrace{(1, 1, \ldots, 1)}_{\substack{n \quad \text{components=1}}} .$$

Convergence in distribution to a constant implies convergence in probability to the same constant vector, and therefore, in the sense of (2.5),

$$(\zeta_1, \ldots, \zeta_n) \xrightarrow{\mathcal{P}} \underbrace{(1, 1, \ldots, 1)}_{\substack{n \quad \text{components=1}}} . \tag{3.5}$$

Here we understand

$$\mathbf{X}_d = \left(\zeta_1(d), \ldots, \zeta_n(d)\right),$$

so d and n switch roles vis-a- vis Cramér-Slutzky Theorem part 1) above.

Since max and min are continuous functions, the Cramér-Slutzky Theorem (2.6) entails by (3.5) that

$$\max\left(\zeta_1, \zeta_2 \ldots, \zeta_n\right) \xrightarrow{\mathcal{P}} 1 \quad \text{as } d \to +\infty \tag{3.6}$$

and

$$\min\left(\zeta_1, \zeta_2 \ldots, \zeta_n\right) \xrightarrow{\mathcal{P}} 1, \quad \text{as } d \to +\infty \tag{3.7}$$

Note that in this situation the function g is a map $g\left(\zeta_1, \zeta_2 \ldots, \zeta_n\right)$ from $\mathbb{R}^n \mapsto \mathbb{R}$ and $\mathbf{X}_d = (\zeta_1, \zeta_2 \ldots, \zeta_n)$, so d and n switch roles vis-a- vis Cramér-Slutzky Theorem part 1) above. Then

$$\frac{D_{\max}}{D_{\min}} = \frac{\frac{1}{d} D_{\max}}{\frac{1}{d} D_{\min}} = \frac{\frac{1}{d} \max_{1 \le l \le n} D_l}{\frac{1}{d} \min_{1 \le l \le n} D_l}$$

$$= \frac{\max_{1 \le l \le n} \frac{D_l}{d}}{\min_{1 \le l \le n} \frac{D_l}{d}} = \frac{\max_{1 \le l \le n} \zeta_l}{\min_{1 \le l \le n} \zeta_l} \xrightarrow{\mathcal{P}} 1, \quad \text{as } d \to +\infty.$$

by the Cramér-Slutzky Theorem (2.7), where one again notes the interchange of roles of d and n. We have now shown that

$$\frac{D_{\max}}{D_{\min}} \xrightarrow{\mathcal{P}} 1 \quad \text{,as } d \to +\infty \tag{3.8}$$

Therefore

$$\mathcal{P}\left(D_{\max} \le (1+\varepsilon) D_{\min}\right) = \mathcal{P}\left(\frac{D_{\max}}{D_{\min}} \le (1+\varepsilon)\right) = \mathcal{P}\left(\frac{D_{\max}}{D_{\min}} - 1 \le \varepsilon)\right)$$

$$= \mathcal{P}\left(\left|\frac{D_{\max}}{D_{\min}} - 1\right| \le \varepsilon)\right),$$

since $D_{\max} > D_{\min}$ so that $\frac{D_{\max}}{D_{\min}} > 1$ and $\left|\frac{D_{\max}}{D_{\min}} - 1\right| = \frac{D_{\max}}{D_{\min}} - 1$.
We have

$$\mathcal{P}\left(\left|\frac{D_{\max}}{D_{\min}} - 1\right| \le \varepsilon)\right) = 1 - \mathcal{P}\left(\left|\frac{D_{\max}}{D_{\min}} - 1\right| > \varepsilon)\right)$$

and therefore

$$\lim_{d \to +\infty} \mathcal{P}\left(D_{\max} \le (1+\varepsilon) D_{\min}\right) = \lim_{d \to +\infty} \mathcal{P}\left(\left|\frac{D_{\max}}{D_{\min}} - 1\right| \le \varepsilon)\right)$$

$$= 1 - \lim_{d \to +\infty} \mathcal{P}\left(\left| \frac{D_{\max}}{D_{\min}} - 1 \right| > \varepsilon) \right)$$
$$= 1 - 0 = 1,$$

since we have established (3.8) in the preceding. This is (3.3), as claimed ∎

Hence: nearest neighbour is unstable in high dimensions.

4 General Comments: Ultrametricity

In [4] and [3] the search dimensionality or the intrinsic dimension ρ of a metric space (like $\mathbb{R}^d$ equipped with $\| \mathbf{x} - \mathbf{y} \|$ as its metric) is defined as

INTRINSIC DIMENSION

$$\rho \stackrel{\text{def}}{=} \frac{(E[D])^2}{2\text{Var}[\text{D}]}. \tag{4.1}$$

Here the statistical distribution of the distances is signeld out as a key quantity. A large ρ implies exponential increase in nearest neighbour searching, which is typical for high dimensional spaces. For excample, with (3.2) we have

$$\rho = \frac{d}{2}.$$

In a high dimensional metric space, the difference between random distances is small compared to a random distance (as will be illustrated in section 5). Roughly equal distances is tantamount to equilateral triangles being formed between triplets of points. Thus high dimensional spaces become naturally trivially ultrametric, as triplets of points form equilateral triangles.

Thus high dimensional spaces in (the sparse) limit become naturally **ultrametric**.

ULTRAMETRIC SPACE

Let x,y and z be elements in $\mathcal{X}$. $d(x, y)$ is a function from $\mathcal{X} \times \mathcal{X}$ to non-negative real numbers. $d(x, y)$ is the distance between x and y, and must satisfy

- $d(x, y) = d(y, x)$ for all $x, y \in \mathcal{X} \times \mathcal{X}$.
- $d(x, y) \geq 0$ for all $x, y \in \mathcal{X} \times \mathcal{X}$.
- $d(x, y) = 0$ if and only if $x = y$.
- $d(x, y)$ satisfies the ultrametric inequality

$$d(x, z) \leq \max\{d(x, y), d(y, z)\}. \tag{4.2}$$

Then we say that (4.2) is the ultrametric property and that $\mathcal{X}$ equipped with such $d(x, y)$ is an **ultrametric space**.
An interpretation of (4.2): If the distance between x and z is big, and the distance between x and y is small, then the distance between y and z has to be big.

5 A Piece of Thinking in the Spirit of Sergey Brink

Let us consider the feature space $U = \{0, 1\}^d$, i.e. the binary hypercube consisting of binary d-tuples $\mathbf{x}$, $\mathbf{x} = (x_1, \ldots, x_d)$, $x_i \in \{0, 1\}$. The cardinality of U is $= 2^d$. Any $\mathbf{x}$ is also called a vertex of $U = \{0, 1\}^d$.

There is a metric on U, it is called the Hamming metric, and denoted by $d_H(\mathbf{x}, \mathbf{y})$. The Hamming metric is defined as the number of positions i, where $\mathbf{x}$ and $\mathbf{y}$ are differing, or

$$d_H(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{d} (x_i +_2 y_i) = \sum_{i=1}^{d} |x_i - y_i|,$$

where $1 +_2 1 = 0$, $0 +_2 0 = 0$, $1 +_2 0 = 1$, $0 +_2 1 = 1$. The alternative $|x_i - y_i|$

applies ordinary real arithmetic and its absolute value. It is easy to prove that d_H is in fact a metric using the representation $\sum_{i=1}^{d} |x_i - y_i|$.

An ultrametric is a special case of a metric. For a metric the triangle inequality, $d_H(\mathbf{x}, \mathbf{z}) \leq d_H(\mathbf{x}, \mathbf{y}) + d_H(\mathbf{y}, \mathbf{z})$ holds instead of (4.2) (note that (4.2) implies the triangle inequality).

A special property of $d_H(\mathbf{x}, \mathbf{y})$ is that the maximum value is equal to d. In fact

$$d_H(\mathbf{x}, \bar{\mathbf{x}}) = d, \tag{5.3}$$

where $\bar{\mathbf{x}}$ has the bits in $\mathbf{x}$ negated, i.e., $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_d)$, where for every i, $\bar{x}_i = 0$, if $x_i = 1$ and $\bar{x}_i = 1$, if $x_i = 0$.

Now we make a study of queries and their nearest neighbours in very large metric spaces in the spirit of [3] and [4].

A *query* is for our purposes simply a preassigned $\mathbf{q} \in U$. Then we draw N independent samples $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ from the uniform distribution $P(\mathbf{x})$ on $\{0,1\}^d$, or for each l

$$P(\mathbf{x}_l) = \frac{1}{2^d}.$$

This is equivalent to that the components x_i are independent Bernoulli variables $\sim \mathrm{Be}(1/2)$.

Let us define for $l = 1, \ldots, N$ the independent random variables, the distances to the query,

$$D_l \stackrel{\text{def}}{=} d_H(\mathbf{x}_l, \mathbf{q}). \tag{5.4}$$

It follows by definition of $d_H(\mathbf{x}_l, \mathbf{q})$ as a sum of zeros and ones that each D_l is a binomial random variable, $D_l \sim \mathrm{Bin}\left(\frac{d}{2}, \frac{d}{4}\right)$ and the D_l are independent. In all of this the query $\mathbf{q}$ is fixed. In addition, by the properties of the binomial distribution

$$E[D_l] = \frac{d}{2}, \mathrm{Var}[D_l] = \frac{d}{4}.$$

We set again

$$D_{\min} = \min_{1 \leq l \leq n} D_l, \quad D_{\max} = \min_{1 \leq l \leq n} D_l.$$

Then we standardize as before

$$\zeta_l = \frac{D_l}{E[D_l]} = \frac{2D_l}{d}$$

and

$$\mathrm{Var}[\zeta_l] = \frac{4}{d^2}\mathrm{Var}[D_l] = \frac{4}{d^2} \cdot \frac{d}{4} = \frac{1}{d}.$$

The condition (3.2) of theorem 3.2 is thus not true here. But a glance at the proof of theorem 3.2 shows, that (3.2) was needed in the step (3.4), when the Chebysjev inequality was applied. However, here we get

$$\mathcal{P}\left(|\zeta_l - 1| > \varepsilon\right) \leq \frac{1}{\varepsilon}\mathrm{Var}\left[\zeta_l\right] = \frac{1}{\varepsilon \cdot d}. \tag{5.5}$$

Hence we have again that for every l, $\zeta_l \xrightarrow{\mathcal{P}} 1$, as $d \to +\infty$.

Theorem 5.1 *$\mathbf{x}_1, \ldots, \mathbf{x}_N$ are I.I.D samples from the probability mass function*

$$P\left(\mathbf{x}_l\right) = \frac{1}{2^d}.$$

When for a fixed $\mathbf{q}$ $D_l = d_H\left(\mathbf{x}_l, \mathbf{q}\right)$ for all l, then for any $\varepsilon > 0$

$$\lim_{d \to +\infty} \mathcal{P}\left(D_{\max} \leq (1+\varepsilon) D_{\min}\right) = 1. \tag{5.6}$$

Proof: We checked above that $\zeta_l \xrightarrow{\mathcal{P}} 1$, as $d \to +\infty$. But then from this fact the rest of the proof required here follows ad verbatim as the proof of theorem 3.2. ∎

The conditions discussed above can be reconciled by saying that we require that the intrinsic dimension in (4.1) is

$$\rho \propto d.$$

Let now D denote a generic D_l or $D \overset{d}{=} D_l$. We want to compute the probability that D lies in the annulus defined by certain two hyperballs in the space $\{0,1\}^d$ equipped with the Hamming metric.

A hyperball in $\{0,1\}^d$ is (c.f. (2.3)

$$\mathrm{sp}_H^d(\mathbf{q}, r) = \{\mathbf{x} \in U | d_H\left(\mathbf{x}, \mathbf{q}\right) \leq r\} \tag{5.7}$$

If $d < r$, then $\mathrm{sp}_H^d(\mathbf{q}, r) = \{0,1\}^d$, i.e. the binary hypercube can be seen as a ball, too, with any vertex $\mathbf{q}$ as center. So we consider for a positive integer k the following set difference, or annulus of two shperes,

$$\mathrm{sp}_H^d\left(\mathbf{q}, \frac{d}{2} + k\right) \setminus \mathrm{sp}_H^d\left(\mathbf{q}, \frac{d}{2} - k\right) = \{\mathbf{x} \in U | \frac{d}{2} - k < d_H\left(\mathbf{x}, \mathbf{q}\right) \leq \frac{d}{2} + k\}.$$

If U is a ball with $\mathbf{q}$ chosen as **north pole** and $\bar{\mathbf{q}}$ as **south pole**, then in view of (5.3) the vertices $\mathbf{x}$ with $d_H\left(\mathbf{x}, \mathbf{q}\right) = \frac{d}{2}$ represent the **equator**.

We want now to compute $\mathcal{P}\left(D \in \mathrm{sp}_H^d\left(\mathbf{q}, \frac{d}{2}+k\right) \setminus \mathrm{sp}_H^d\left(\mathbf{q}, \frac{d}{2}-k\right)\right)$ or

$$\mathcal{P}\left(\frac{d}{2}-k < D \leq \frac{d}{2}+k\right).$$

One rewrites this probability as

$$= \mathcal{P}\left(\frac{-k}{\sqrt{\frac{d}{4}}} < \frac{D-\frac{d}{2}}{\sqrt{\frac{d}{4}}} \leq \frac{k}{\sqrt{\frac{d}{4}}}\right).$$

By first courses, see, e.g., [5, p. 162−163] we know that a random variable $\sim \mathrm{Bin}(Np, Np(1-p)$ can be approximated by a normal distribution with mean Np and variance $Np(1-p)$. If we now take advantage of this approximation, we obtain that

$$Z = \frac{D-\frac{d}{2}}{\sqrt{\frac{d}{4}}} \sim N(0,1),$$

since $E[D] = \frac{d}{2}$ and $\mathrm{Var}[D] = \frac{d}{4}$. and the binomially distributed D is approximately $N(\frac{d}{2}, \frac{d}{4})$. If $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution,

$$\mathcal{P}\left(\frac{-k}{\sqrt{\frac{d}{4}}} < Z \leq \frac{k}{\sqrt{\frac{d}{4}}}\right) = \Phi\left(\frac{k}{\sqrt{\frac{d}{4}}}\right) - \Phi\left(\frac{-k}{\sqrt{\frac{d}{4}}}\right)$$

$$= 2\Phi\left(\frac{k}{\sqrt{\frac{d}{4}}}\right) - 1.$$

In summary,

$$\mathcal{P}\left(\frac{d}{2}-k < D \leq \frac{d}{2}+k\right) \approx 2\Phi\left(\frac{k}{\sqrt{\frac{d}{4}}}\right) - 1. \tag{5.8}$$

This is the probality that D differs from the mean distance $\frac{d}{2}$ to $\mathbf{q}$ by at most $\pm k$ bits. We set

$$p_{k,d} \stackrel{\mathrm{def}}{=} 2\Phi\left(\frac{k}{\sqrt{\frac{d}{4}}}\right) - 1.$$

By Matlab or any other computing software we can find the numerical values of p_{kd} at will. Then we have, e.g.,

$$p_{30,1000} = 0.9422, p_{40,1000} = 0.9886, p_{50,1000} = 0.9984,$$

and

$$p_{200,10000} = 0.9999, p_{180,10000} = 0.9997, p_{160,10000} = 0.9986.$$

Hence we see, e.g., that the 99% of the sample distances D_l to any query $\mathbf{q}$ are for $k = 160$ and $d = 10000$ located around the equator of $\{0,1\}^d$ or at distances 5000 ± 160 bits from the query. This expresses part of the findings/ideas in [3]. In the binary hypercube with Hamming metric, the intrinsic dimension (4.1) becomes

$$\rho = \frac{d}{2}, \tag{5.9}$$

which is well illustrated by the computations above. In words, in high dimensional spaces a large ρ reflects that the difference between random distances D is small, which is quantitatively found by the numbers $p_{k,d}$.

Or, take now $\mathbf{x}$ and $\mathbf{y}$ both in the annulus of high probability

$$\{\mathbf{x} \in U | \frac{d}{2} - k < d_H(\mathbf{x}, \mathbf{q}) \leq \frac{d}{2} + k\}.$$

Then it follows that

$$d_H(\mathbf{x}, \mathbf{q}) - d_H(\mathbf{y}, \mathbf{q}) \leq \frac{d}{2} + k - (\frac{d}{2} - k) = 2k$$

and

$$d_H(\mathbf{x}, \mathbf{q}) - d_H(\mathbf{y}, \mathbf{q}) \geq \frac{d}{2} - k - (\frac{d}{2} + k) = -2k.$$

Hence

$$|d_H(\mathbf{x}, \mathbf{q}) - d_H(\mathbf{y}, \mathbf{q})| \leq 4k.$$

This seems to support the notion of a tendency to ultrametrical samples in high dimensions, that the difference between random distances is small compared to a random distance.

Appendix A: Facts about the Analytic Geometry of the Hypercube in $\mathbb{R}^d$

The *hypercube* Ω in $\mathbb{R}^d$ is the subset

$$\Omega \stackrel{\text{def}}{=} [0,1]^d.$$

The distance in (2.1), $\| \mathbf{x} - \mathbf{y} \|$, is used for $\mathbf{x} \in \Omega$ and $\mathbf{y} \in \Omega$.

> Assume that $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ are N I.I.D. samples of $U([0,1]^d) =$ the uniform distribution on $[0,1]^d$. Thus
>
> $$\mathbf{X}_l \sim U([0,1]^d).$$
>
> This assumption holds everywhere in Appendix A.

This means that

$$\mathbf{X} = (X_1, \ldots, X_d), X_i \sim U(0,1), \quad i = 1, 2, \ldots, d. \tag{A.1}$$

Thus for any (Borel) subset $A \subseteq \Omega$

$$\mathcal{P}(\mathbf{X} \in A) = \int \int \ldots \int_A dx_1 \ldots dx_d = \text{vol}(A)$$

(volume of A) and

$$\text{vol}(\Omega) = \mathcal{P}(\mathbf{X} \in \Omega) = 1.$$

A.1 The probability of a Subcube in Ω

Take $0 < s < 1$ and $s = b - a$ (=length of any side of the subcube).

$$\Omega_s = \{\mathbf{x} \in \Omega | a \leq x_i \leq b, \quad i = 1, \ldots, d\}$$

is a subcube (draw a picture in two dimensions). Then

$$\mathcal{P}\left(\mathbf{X} \in \Omega_s\right) = \mathcal{P}\left(a \leq X_1 \leq b, a \leq X_2 \leq b, \ldots, a \leq X_d \leq b\right)$$

and by I.I.D. this is

$$= \mathcal{P}\left(a \leq X_1 \leq b\right) \cdot \mathcal{P}\left(a \leq X_2 \leq b\right) \cdots \mathcal{P}\left(a \leq X_d \leq b\right).$$

When $U \overset{\mathrm{d}}{=} X_l$ for all l (I.I.D.) and by $U \sim U(0,1)$, the above becomes by (A.1)

$$= \left(\mathcal{P}\left(a \leq U \leq b\right)\right)^d = (b-a)^d = s^d.$$

For example, $d = 100$, $s = 0.95$, $s^d \approx 0.0059 = 0.59\%$ (little more tha a half percent).

Fact 1.: The subcube Ω_s is in high dimensions very sparsely populated.

A.2 The largest hyperball that fits entirely in Ω

The largest hyperball that fits entirely in Ω should be

$$\mathrm{sp}^d(\mathbf{q}, 1/2).$$

We compute the probability that one data point lies in $\mathrm{sp}^d(\mathbf{q}, 1/2)$, or,

$$\mathcal{P}(\mathbf{X} \in \mathrm{sp}^d(\mathbf{q}, 1/2)) = \int\int \ldots \int_{\mathrm{sp}^d(\mathbf{q},1/2)} dx_1 \ldots dx_d = K_d \cdot \left(\frac{1}{2}\right)^d, \quad \text{(A.2)}$$

where

$$K_d = \mathrm{vol}\left(\mathrm{sp}^d(\mathbf{q}, 1)\right).$$

The evaluation of the integral in (A.2) is a useful exercise in multivariate calculus. One solution is recapitulated in, e.g., [2, chapter 2.2-2.4]. There or elsewhere, [9, p. 74], one can find that

$$K_d = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)} = \frac{\pi^{d/2}}{\frac{d}{2}\Gamma(d/2)},$$

where $\Gamma(z)$ is the Euler Gamma function. **Stirling's formula** [9, p. 288] says

$$\Gamma(x) \approx \sqrt{2\pi} e^{-x} x^{x-1/2}$$

so that

$$\Gamma(d/2) \approx \sqrt{2\pi} \cdot e^{-d/2} \cdot (d/2)^{d/2-1/2}.$$

Thus

$$\mathcal{P}\left(\mathbf{X} \in \mathrm{sp}^d(\mathbf{q}, 1/2)\right) \approx \frac{1}{\sqrt{2\pi}} e^{\frac{d}{2}\ln(\pi) - (\frac{d}{2}-1)\ln(\frac{d}{2})} e^{-d(\ln(\ 2)-\frac{1}{2})}. \tag{A.3}$$

Here $\ln(\ 2) - \frac{1}{2} = 0.19$, hence $e^{-d(\ln(\ 2)-\frac{1}{2})} \to 0$, as $d \to +\infty$. We see by plotting, or checking that $\frac{d}{dx} f(x) < 0$ for all large enough x, that

$$f(x) = \frac{x}{2}\ln(\pi) - (\frac{x}{2} - 1)\ln(\frac{x}{2}) \to -\infty, \quad \text{as } x \to +\infty.$$

Hence $e^{\frac{d}{2}\ln(\pi) - (\frac{d}{2}-1)\ln(\frac{d}{2})} \to 0$, as $d \to +\infty$

Thus

$$\mathcal{P}\left(\mathbf{X} \in \mathrm{sp}^d(\mathbf{q}, 1/2)\right) \to 0,$$

as $d \to +\infty$.

Remark 5.1 By the above

$$\mathrm{vol}\,(\Omega) = 1 \quad \text{for all } d.$$

But the diameter of Ω is

$$\| \mathbf{1} - \mathbf{0} \| = \sqrt{d},$$

where $\mathbf{1}$ is the (cube corner) vector of ones in $\mathbb{R}^d$ and $\mathbf{0}$ is the origin in $\mathbb{R}^d$.

■

Fact 2.: The volume of $\mathrm{sp}^d(\mathbf{q}, 1/2)$ in high dimensions shrinks sharply as d grows and it is increasingly impopular that any point will be found in this hyper ball at all. Yet this hyperball is anchored on the sides of a hypercube with increasing diameter.

A.3 Expected Number of Points in $\mathrm{sp}^d(\mathbf{q}, 1/2)$

$\mathbf{X}_1, \ldots, \mathbf{X}_N$ are N I.I.D. random variables $\sim U([0,1]^d)$. Consider a *trial*:

check if $\mathbf{X}_l$ is in $\mathrm{sp}^d(\mathbf{q}, 1/2)$ (a success) or not (failure). These trials are independent and have only two possible outcomes.

$$p \stackrel{\text{def}}{=} \mathcal{P}(X_l \in \mathrm{sp}^d(\mathbf{q}, 1/2)) = \frac{\pi^{d/2}}{\frac{d}{2}\Gamma(d/2)} \cdot \left(\frac{1}{2}\right)^d, \tag{A.4}$$

does not depend on l. Let $Y =$ the number of successes in $\mathbf{X}_1, \ldots, \mathbf{X}_N$. Then we find that $Y \sim \mathrm{Bin}\,(N, p)$, see [5, pp. 113-117]. Thus

$$E\,[Y] = Np.$$

Therefore, in order that expected number successes or number of points in $\mathrm{sp}^d(\mathbf{q}, 1/2)$ to be at least one, we need in view of (A.4)

$$E\,[Y] \geq 1 \Leftrightarrow N \geq \frac{1}{p} = \frac{\Gamma\left(\frac{d}{2}\right)}{\pi^{d/2}} \cdot \left(\frac{d}{2}\right) 2^d.$$

For $d = 10$, $N \approx 401.5$.

Fact 3.: If, e.g., $d = 20$, then you may expect one of 40631627 points to be inside $\mathrm{sp}^d(\mathbf{q}, 1/2)$.

A.4 What is the probability that the nearest neighbour lies in the hyperball $\mathrm{sp}^d(\mathbf{q}, r)$?

Nearest Neighbour

$\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is a set of points $\mathbf{x}_l \in \Omega$. The nearest neighbour, denoted by $nn(\mathbf{q}) \in \mathcal{S}$, to $\mathbf{q} \notin \mathcal{S}$ is

$$\mathrm{nn}(\mathbf{q}) = \{\mathbf{x}_l \in \mathcal{S} | \parallel \mathbf{x}_l - \mathbf{q} \parallel \leq \parallel \mathbf{x}_j - \mathbf{q} \parallel, l \neq j\}. \tag{A.5}$$

Then we can define the minimum distance between $\mathbf{q}$ and $\mathcal{S}$,

$$\mathrm{nn}^{\mathrm{dist}}(\mathbf{q}) \stackrel{\text{def}}{=} \parallel \mathrm{nn}(\mathbf{q}) - \mathbf{q} \parallel . \tag{A.6}$$

We want to know the probability that the nearest neighbour to $\mathbf{q} \in \Omega$ in a data set $\mathcal{S}$ lies in the hyperball $\mathrm{sp}^d(\mathbf{q}, r)$ of radius r or

$$\mathcal{P}(\mathbf{q}, r) = \mathcal{P}(nn(\mathbf{q}) \in \mathrm{sp}^d(\mathbf{q}, r))$$

By our definitions, see (A.5) and (A.6),

$$\mathcal{P}(\mathbf{q}, r) = \mathcal{P}(\mathrm{nn}^{\mathrm{dist}}(\mathbf{q}) \leq r).$$

Next ensues a calculation using the rules of any first course in probability. One starts with

$$\mathcal{P}(\mathbf{q}, r) = 1 - \mathcal{P}(\mathrm{nn}^{\mathrm{dist}}(\mathbf{q}) > r).$$

But $\mathrm{nn}^{\mathrm{dist}}(\mathbf{q}) > r$ if and only if it holds for all l that $\parallel \mathbf{x}_l - \mathbf{q} \parallel > r$. Thus

$$\mathcal{P}(\mathrm{nn}^{\mathrm{dist}}(\mathbf{q}) > r) = \mathcal{P}(\parallel \mathbf{x}_1 - \mathbf{q} \parallel > r, \parallel \mathbf{x}_2 - \mathbf{q} \parallel > r, \ldots \parallel \mathbf{x}_N - \mathbf{q} \parallel > r)$$

$$= \mathcal{P}(\parallel \mathbf{x}_1 - \mathbf{q} \parallel > r) \cdot \mathcal{P}(\parallel \mathbf{x}_2 - \mathbf{q} \parallel > r) \cdots \mathcal{P}(\parallel \mathbf{x}_N - \mathbf{q} \parallel > r) = \left(\mathcal{P}\left(\parallel \mathbf{x}_1 - \mathbf{q} \parallel > r\right)\right)^N,$$

since the N samples are I.I.D.. Then

$$\left(\mathcal{P}\left(\parallel \mathbf{x}_1 - \mathbf{q} \parallel > r\right)\right)^N = \left(1 - \mathcal{P}(\parallel \mathbf{x}_1 - \mathbf{q} \parallel \leq r)\right)^N.$$

But

$$\mathcal{P}(\parallel \mathbf{x}_1 - \mathbf{q} \parallel \leq r) = \mathcal{P}\left(\{\mathbf{x} \in \Omega | \parallel \mathbf{x}_1 - \mathbf{q} \parallel \leq r\}\right) = \mathrm{vol}\left(\mathrm{sp}^d(\mathbf{q}, r) \cap \Omega\right).$$

Hence

Fact4.:

$$\mathcal{P}\left(\mathbf{q}, r\right) = 1 - \left(1 - \mathrm{vol}\left(\mathrm{sp}^d(\mathbf{q}, r) \cap \Omega\right)\right)^N.$$

With $r = 1/2$, $\mathrm{sp}^d(\mathbf{q}, 1/2) \cap \Omega) = \mathrm{sp}^d(\mathbf{q}, 1/2)$. We may be interested in three issues **i**) $-$ **iii**) for $r = 1/2$: The first two are immediate consequences of Fact 4.

i) When the data set becomes large but d is fixed,

$$\mathcal{P}(\mathbf{q}, 1/2) \to 1, \quad \text{as } N \to +\infty.$$

ii) When the dimension becomes high for fixed a data size, and for $r = 1/2$

$$\mathcal{P}(\mathbf{q}, 1/2) \to 0, \quad \text{as } d \to +\infty.$$

This follows by Appendix **A.2** .

iii) What will happen, if both d grows and the data set becomes large at the same time ? Let us assume that $N = e^{\lambda d - 1/2}$, or data volume grows exponentially in high dimension. In view of (A.3) we take that

$$\text{vol}\left(\text{sp}^d(\mathbf{q}, 1/2)\right) \approx e^{-d(\ln(\ 2) - \frac{1}{2})} \tag{A.7}$$

where we approximate with the factor as function of d that turns slowest to 0. We set

$$\lambda = \ln(\ 2) - \frac{1}{2}, \quad \ln N + 1/2 = d\lambda.$$

Then

$$\mathcal{P}\left(\mathbf{q}, 1/2\right) = 1 - \left(1 - \frac{e^{-1/2}}{N}\right)^N.$$

As $N \to +\infty$ we get

$$\mathcal{P}\left(\mathbf{q}, 1/2\right) \to 1 - e^{-e^{-1/2}}.$$

This is now approximately the probability that the nearest neighbour to $\mathbf{q}$ in an N I.I.D. sample lies in $\text{sp}^d(\mathbf{q}, 1/2)$.

We recall that the cumulative distribution function of the **standard Gumbel distribution** (with mean equal to the Euler's constant γ), see [6, ch. 7.5, p. 200], is

$$F(x) = e^{-e^{-x}}.$$

Hence we have found that in high dimensions d and for large data sets N with $N = e^{\lambda d - 1/2}$

$$\mathcal{P}(nn(\mathbf{q}) \in \text{sp}^d(\mathbf{q}, 1/2)) \approx 1 - F(1/2) = \mathbb{P}(X > 1/2),$$

when X has the standard Gumbel distribution. Gumbel distribution is an extreme value distribution [6, ch. 7.5].

Appendix B: k-Nearest Neighbour Regression

We consider the data generating model with unknown f

$$Y = f(\mathbf{x}) + \epsilon,$$

where $E[\epsilon] = 0$, $E[\epsilon^2] = \sigma_\epsilon^2$. The function $\widehat{f}_k(\mathbf{x})$ estimating $f(\mathbf{x})$ depends on the training data $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ as a **k-nearest neighbour regression**. This means for any $\mathbf{x} \in \mathcal{X}$

$$\widehat{f}_k(\mathbf{x}) = \frac{1}{k} \sum_{i:\mathbf{x}_i \in \mathrm{nn}_k(\mathbf{x})} y_i. \tag{B.1}$$

Here

$$\mathrm{nn}_k(\mathbf{x}) = \text{ the set of } k\text{th nearest neighbours to } \mathbf{x}.$$

i.e.,

$$= \{\mathbf{x}_{(i)} \in \mathcal{S} \mid\mid \| \mathbf{x}_{(1)} - \mathbf{x} \| \leq \| \mathbf{x}_{(2)} - \mathbf{x} \| \leq \ldots \| \mathbf{x}_{(k)} - \mathbf{x} \|\}.$$

For example with $k = 1$

$$\widehat{f}_1(\mathbf{x}) = y_l,$$

if $\mathbf{x}_l = nn_1(\mathbf{x})$. Assume that a training set feature vectors $\mathbf{x}_l$ are sampled as I.I.D. of

$$\mathbf{X}_l \sim U([0,1]^d)$$

as in Appendix A above, and that at the sampled points

$$Y_l = f(\mathbf{x}_l) + \epsilon_l$$

where f is unknown. We want to find the expected prediction error $EPE(\mathbf{x})$ for $\widehat{f}_k(\mathbf{x})$, when we choose a new point $\mathbf{x}$ and receive corresponding Y. We give this in the form of the bias -variance trade-off.

The general bias -variance trade-off is

$$EPE(\mathbf{x}) = \mathrm{E}\Big[\big(Y - \hat{f}(\mathbf{x})\big)^2\Big] = \mathrm{Bias}\big[\hat{f}(\mathbf{x})\big]^2 + \mathrm{Var}\big[\hat{f}(\mathbf{x})\big] + \sigma^2, \tag{B.2}$$

where

$$\mathrm{Bias}\big[\hat{f}(\mathbf{x})\big] = \mathrm{E}\big[\hat{f}(\mathbf{x})\big] - f(\mathbf{x}) \tag{B.3}$$

and

$$\mathrm{Var}\big[\hat{f}(\mathbf{x})\big] = \mathrm{E}\Big[\big(\hat{f}(\mathbf{x}) - \mathrm{E}[\hat{f}(\mathbf{x})]\big)^2\Big]. \tag{B.4}$$

We need to evaluate all of these terms with $\widehat{f}_k(\mathbf{x})$. First,

$$E\left[\widehat{f}_k(\mathbf{x})\right] = \frac{1}{k} \sum_{\sum_{i:\mathbf{x}_i \in \mathrm{nn}_k(\mathbf{x})}} E\left[Y_i\right]$$

$$= \frac{1}{k} \sum_{\mathrm{nn}_k(\mathbf{x})} E\left[f(\mathbf{x}_l) + \epsilon_l\right]$$

$$= \frac{1}{k} \sum_{\mathrm{nn}_k(\mathbf{x})} E\left[f(\mathbf{x}_l)\right] + \frac{1}{k} \sum_{\mathrm{nn}_k(\mathbf{x})} \underbrace{E\left[\epsilon_l\right]}_{=0}$$

$$= \frac{1}{k} \sum_{\mathrm{nn}_k(\mathbf{x})} E\left[f\left(\mathbf{x}_{(l)}\right)\right] = \frac{1}{k} \sum_{\mathrm{nn}_k(\mathbf{x})} f\left(\mathbf{x}_{(l)}\right) = \frac{1}{k} \sum_{l=1}^{k} f\left(\mathbf{x}_{(l)}\right).$$

Next,

$$\mathrm{Var}\left[\widehat{f}_k(\mathbf{x})\right] = \frac{1}{k^2} \sum_{i:\mathbf{x}_i \in \mathrm{nn}_k} \mathrm{Var}\left[Y_i\right]$$

$$= \frac{1}{k^2} \sum_{\mathrm{nn}_k(\mathbf{x})} \sigma_\epsilon^2 = \frac{1}{k^2} k \sigma_\epsilon^2$$

$$= \frac{1}{k} \sigma_\epsilon^2.$$

$$EPE(\mathbf{x}) = \mathrm{E}\Big[\big(Y - \hat{f}_k(\mathbf{x})\big)^2\Big] = \frac{1}{k}\sigma_\epsilon^2 + \left(f(\mathbf{x}) - \frac{1}{k}\sum_{l=1}^{k} f\left(\mathbf{x}_{(l)}\right)\right)^2 + \sigma_\epsilon^2.$$

Some comments:

1. In small dimension d, the closest neighbours will have their function values $f\left(\mathbf{x}_{(l)}\right)$ close to $f(\mathbf{x})$, whence their average $\frac{1}{k}\sum_{l=1}^{k} f(\mathbf{x}_{(l)})$ should be close to $f(\mathbf{x})$.

2. But in large dimension, Fact 2. above tells that the samples $\mathbf{x}_l \sim U([0,1]^d)$ will be close to the boundaries of the hypercube $[0,1]^d$. Hence for most $\mathbf{x} \in [0,1]^d$, the sum $\frac{1}{k}\sum_{l=1}^{k} f(\mathbf{x}_{(l)})$ will be an extrapolation from neighbouring samples rather than an interpolation between them. Hence

$$\left(f(\mathbf{x}) - \frac{1}{k}\sum_{l=1}^{k} f(\mathbf{x}_{(l)}) \right)^2$$

can be very large for large d and with large k.

3. It has been shown that for any distribution $\mathcal{P}$

$$\max_{\mathbf{x}\in[0,1]^d} |E\left[Y|X=\mathbf{x}\right] - \hat{f}_k(\mathbf{x})| \to 0,$$

as N, the number of samples in the training set, goes to $+\infty$. In high dimensions the convergence rate is slow.

$$\max_{\mathbf{x}\in[0,1]^d} \mid E\left[Y|X=\mathbf{X}\right] - \hat{f}_k(\mathbf{x}) \mid \to 0$$

The k-nearest neighbor regression is a **universal estimator**.

References

[1] Beyer, Kevin and Goldstein, Jonathan and Ramakrishnan, Raghu and Shaft, Uri: When is "nearest neighbor" meaningful? International conference on database theory ICDT 1999.*Lecture Notes in Computer Science*, vol 1540. Springer, pp. 217–235, Berlin, Heidelberg, 1999

[2] Blum, Avrim, Hopcroft, John and Kannan, Ravi *Foundations of data science*, 2016. `https://www.math.kth.se/matstat/gru/sf2935/`

[3] Brin, Sergey: *Near neighbor search in large metric spaces*, International Conference on Very Large Data Bases VLDB, 1995.

[4] Chávez, Edgar and Navarro, Gonzalo and Baeza-Yates, Ricardo and Marroquín, José Luis: Searching in metric spaces, *ACM computing surveys (CSUR)*, 33, 3, pp. 273−321, 2001.

[5] Devore, Jay L: *Probability and Statistics for Engineering and the Sciences, Fourth Edition*, 1995.

[6] Gut, Allan: *An Intermediate Course in Probability Second Edition*, 2009.

[7] James, Gareth and Witten, Daniela and Hastie, Trevor and Tibshirani, Robert: *An introduction to statistical learning*, 2013.

[8] National Research Council Committee on the analysis of massive data: *Frontiers in massive data analysis*, National Academies Press, 2013.

[9] Råde, L and Westergren, B: *Beta, Mathematics Handbook for Science and Engineering*, 1995.