Note: An alternative proof of the vulnerability of $k$-NN classifiers in high intrinsic dimensionality regions

Teddy Furon  
Univ Rennes, Inria, CNRS, IRISA - France

Abstract

This document proposes an alternative proof of the result contained in article [1]. The proof is simpler to understand (I believe) and leads to a more precise statement about the asymptotical distribution of the relative amount of perturbation.

I. CONTEXT

Let us consider a collection of $n$ points $\{x_i\}_{i=1}^n \in \mathbb{R}^d$, plus a query point $q$ everywhere in this space. Suppose that an artificial intelligent program bases its decision on the collection points neighbouring the query. Especially, the decision will take collection point $x_i$ into account if it is one of the $k_t$ nearest neighbours of $q$.

Suppose that this is not the case for that query $q$ and this collection point $x$. In other words $x$ is only the $k_x$ nearest neighbour of $q$ with $1 \leq k_t < k_x \leq n$. Let us rename by $t$ the collection point which is the $k_t$ nearest neighbour of $q$, we have that $t < x$, with $t := \|t - q\|$ and $x := \|x - q\|$.

We are interested in the amount of perturbation to be applied to collection point $x$ so that the program takes it into account. This perturbation pushes $x$ to a new point $y$ so that its neighbour rank is $k_y \leq k_t$, at a distance $y \leq t$ away from $q$.

The amount of perturbation is evaluated by the ratio $\delta := \|x - y\|/\|x - q\|$, with $\delta \in (0, 1)$. Obviously, this ratio is set to the minimum if $x$ is pushed onto $y$ in a direct line towards $q$: $y \in [x, q]$. This implies that:

$$k_y < k_t \iff y < t \iff \delta > 1 - \frac{t}{x}.$$  

(1)

Note that this quantity depends on the configuration of the collection points locally around $q$. From now on, we will consider that this collection of points is indeed random. This means that distances $t$ and $x$ are occurrences of absolutely continuous random variables $T_n$ and $X_n$ (note that $0 < T_n < X_n$), and so is the relative amount of perturbation $\delta$ w.r.t. r.v. $\Delta_n := 1 - T_n/X_n$. The subscript $n$ stresses that the size of the collection is a major factor: as the size increases, we expect that the $k_t$ (resp. $k_x$) neighbour comes closer to $q$.

We do not impose a specific distribution of the collection points in $\mathbb{R}^d$, but only of their distances from $q$.

Assumption 1. The distances of collection point w.r.t. the query $q$ are independent and identically distributed, whose c.d.f. is denoted by $F: \mathbb{R}^+ \to [0, 1]$ and p.d.f. $f: \mathbb{R}^+ \to \mathbb{R}^+.$
Then

\int_{0}^{1} g(\xi) e^{h(\xi)} d\xi < +\infty

\Rightarrow 0^+ \implies \frac{A\xi}{\alpha} \sim \alpha > -1 \quad (11)

\frac{A\beta}{\beta} \sim \beta > 0 \quad (12)

Then

\int_{0}^{1} g(\xi) e^{h(\xi)} d\xi \sim \frac{A\alpha}{\beta} \frac{\alpha + 1}{\beta} e^{\alpha n(\alpha n)^{-\frac{\alpha + 1}{\beta}}} \quad as \ n \to +\infty \quad (13)
Before applying this lemma, we need the following assumption:

**Assumption 2.** The c.d.f. $F(\cdot)$ is a regularly varying function around $0^+$ (as defined by J. Karamata).

This assumption holds from the theory of extreme values. As $n$ increases, the distances of the $k_x$ and $k_t$ nearest neighbour tap into the lower tail of the distribution, which ought to be regularly varying because it is lower bounded by 0. This implies that, there exists a parameter $\ell > 0$ (so-called index of regular variation, or intrinsic dimensionality in \cite{1}) s.t.

$$
\lim_{x \to 0^+} \frac{F((1 - \delta)x)}{F(x)} = \lim_{\xi \to 0^+} \frac{F((1 - \delta)F^{-1}(\xi))}{\xi} = (1 - \delta)^\ell.
$$

(14)

We then rewrite \cite{9} in the form

$$
\int_0^1 g(\xi) e^{nh(\xi)} \, dx / B(k_x, n - k_x + 1)
$$

(9) with

$$
h(\xi) = \log(1 - \xi)
$$

(15)

$$
g(\xi) = \left(1 - I_{(1 - \delta)\ell}^{-1}(\xi)(k_t, k_x - k_t)\right) \xi^{k_x - 1}(1 - \xi)^{-k_t},
$$

(16)

in order to instantiate the constants of Lemma 1 as

$$
a = 0, \quad c = 1, \quad \beta = 1, \quad A = 1 - I_{(1 - \delta)\ell}^{-1}(k_t, k_x - k_t), \quad \alpha = k_x - 1.
$$

(17)

This leads to the following asymptotical expression and limit: $\forall \delta \in (0,1)$

$$
F_{\Delta_n}(\delta) \sim I_{1 - (1 - \delta)\ell}(k_x - k_t, k_t)\frac{n!}{n^{k_x}(n - k_x)!} n^{-\frac{1}{2}} I_{1 - (1 - \delta)\ell}(k_x - k_t, k_t).
$$

(18)

This can be restated as follows:

**Proposition 1.** As the size $n$ of the collection increases, the relative amount of perturbation $\Delta_n$ converges in distribution to $\Delta := 1 - (1 - B)^{1/\ell}$ with $B \sim \text{Beta}(k_x - k_t, k_t)$.

I did not find any close-form expression for $E[\Delta]$. Knowing that:

$$
E[B] = \frac{k_x - k_t}{k_x},
$$

(19)

$$
\forall[B] = \frac{E[B](1 - E[B])}{(k_x + 1)},
$$

(20)

we see that $B$ concentrates around its expectation as $k_x$ becomes large. Thanks to a second order Taylor series, we obtain

$$
E[\Delta] \approx 1 - \left(\frac{k_t}{k_x}\right)^{1/\ell} \left(1 - \frac{\ell - 1}{2\ell^2} \frac{k_x - k_t}{k_t(1 + k_x)}\right).
$$

The term $1 - \left(\frac{k_t}{k_x}\right)^{1/\ell}$ is the main result contained in article \cite{1}. It outlines that “the amount of perturbation required to subvert neighborhood rankings diminishes” with the local intrinsic dimensionality $\ell$ of this neighborhood. For instance, for large $\ell$, this further simplifies into $E[\Delta] \approx \log(k_t/k_x)/\ell$, which shows that $\ell$ has a bigger impact than the ratio $k_t/k_x$.

On the contrary, translating quantiles of $B$ to quantiles of $\Delta$ is easier as $x \to 1 - (1 - x)^{1/\ell}$ is a monotonic function. For instance, the median of $\Delta$ is approximately, for $k_t \geq 2$ and $k_x \geq k_t + 2$:

$$
\Delta_m \approx 1 - \left(\frac{3k_t - 1}{3k_x - 2}\right)^{1/\ell}.
$$

(21)
REFERENCES

[1] L. Amsaleg, J. Bailey, A. Barbe, S. M. Erfani, T. Furon, M. E. Houle, M. Radovanović, and N. X. Vinh, “High intrinsic dimensionality facilitates adversarial attack: Theoretical evidence,” *IEEE Transactions on Information Forensics and Security*, pp. 1–1, 2020.

[2] V. Bonnaillie-Noël, “Méthode de Laplace et de la phase stationnaire,” ENS de Cachan, Tech. Rep., 2004.