



CENTRUM APLIKOVANÉ KYBERNETIKY

České vysoké učení technické v Praze - fakulta elektrotechnická

Features of Nearest Neighbors Distances in High-Dimensional Space

Technical report

Marcel Jiřina and Marcel Jiřina, jr.

www@c-a-k.cz

2004



Institute of Computer Science
Academy of Sciences of the Czech Republic

Features of Nearest Neighbors Distances in High-Dimensional Space

Marcel Jiřina and Marcel Jiřina, jr.

Technical Report No. 913

July 2004

Abstract. Methods of nearest neighbors are essential in wide range of applications where it is necessary to estimate probability density (e.g. Bayes's classifier, problems of searching in large databases). This paper contemplates on features of distribution of nearest neighbors' distances in high-dimensional spaces. It shows that for uniform distribution of points in n -dimensional Euclidean space the distribution of the distance of the i -th nearest neighbor to the n -power has Erlang distribution. A power approximation of the newly introduced probability distribution mapping function of distances of nearest neighbors in the form of suitable power of the distance is presented. An influence of the boundary effect is also discussed. Also presented is way to state distribution mapping exponent q for a probability density estimation including boundary effect in high-dimensional spaces.

Keywords: Multivariate data; Nearest neighbors; Probability distribution mapping function; Probability density mapping function; Power approximation

1 Introduction

In a probability density estimate by the method of k nearest neighbors [5], [8] or in problems of searching in large databases [1], [2], [13], [14], distances of the nearest neighbor or several nearest neighbors are essential.

There are at least two distinct problems dealing with the nearest neighbor or several nearest neighbors to a given point in n -dimensional Euclidean space E_n . One of them is a nonparametric technique of probability density estimation in multidimensional data space [11], [5], and the other is a problem of the nearest neighbor searching which is interesting and important in database applications [1], [2], [8]. The first class of problems looks for the highest quality of the probability density estimation, and efficiency and speed are secondary. For the other class, searching in large databases, the maximal performance, i.e. the speed of the nearest neighbor searching is a primary task.

The rather strange behavior of the nearest neighbors in high dimensional spaces was found. For the problem of searching the nearest neighbor in large databases, the boundary phenomenon (boundary effect) was studied in [1] using approximation by so called bucketing algorithm and l_{max} metrics. In [3] the problem of finding k nearest neighbors was studied in general metric spaces and in [10], the so-called concentration phenomenon was described. It was found in [2] and in [10] that, increasing dimensionality, the distance to the nearest data point approaches the distance to the farthest data point of the data set.

For probability density estimation by the k -nearest-neighbor method in E_n , the best value of k must be carefully tuned to find optimal results. Let there be a ball with its center in x and containing k points. Let the volume of the ball be V_k , and the total number of points m_T . Then for the probability density estimate in point x (a query point [13], [14]), it holds [5] that

$$p_k(x) = \frac{k / m_T}{V_k}. \quad (1)$$

It will be shown that starting from some k the value of $p_k(x)$ is not constant for larger k , as it should be, but lessens. It is caused by the boundary phenomenon.

The goal of this study is to analyze the distances of the nearest neighbors from the query point x and the distances between two of these neighbors, the i -th and $(i-1)$ -st of randomly distributed points without and with boundary effect consideration in Euclidean space E_n . We introduce the probability distribution mapping function, and the distribution density mapping function which maps probability density distribution of points in E_n to a similar distribution in the space of distances, which is one-dimensional, i.e. E_1 . Thus the dimensionality problem is significantly reduced. The power approximation of the probability distribution mapping function in the form of $(\text{distance})^q$ is introduced and a way is shown to choose distribution mapping exponent q for a probability density estimation including the boundary effect in high dimensions. It will be also shown that exponent q is something like local value of correlation dimension [7], [12].

2 Probability Density Estimate Based on Powers of Distances

The nearest-neighbor-based methods usually use (1) for a probability density estimate and are based on the distances of neighbors from a given (unknown) point, i.e. on a simple transformation $E_n \rightarrow E_1$.

Using the neighbor distances for the probability density estimation should copy the features of the probability density function based on real data. The idea of most nearest-neighbors-based methods as well as of kernel methods [8] does not reflect the boundary effects. That means that for any point x the statistical distribution of the data points x_i surrounding it is supposed to be independent of the location of the neighbor points and their distances x_i from point x . This assumption is not often met, especially for small data sets and higher dimensions.

To illustrate this, let us consider points uniformly distributed in a cube and a ball inserted tightly into the cube. The higher is space dimension, the smaller is amount of the cube occupied by the ball. In other words, the majority of points lie outside the ball somewhere "in the corner" of the cube (the boundary phenomenon [1]). It seems that in farther places from the origin, the space is less dense than near the origin. To illustrate this, let us consider uniformly distributed points in cube $(-0.5, +0.5)^n$ (individual dimensions are measured in centimeters, cm). Let there be a ball with its center in the origin

and the radius equal to 0.5 cm. This ball occupies $\frac{4}{3}\pi \cdot 0.5^3 = 0.524 \text{ cm}^3$, i.e. more than 52 % of that cube in a three-dimensional space, 0.080746 cm^6 , i.e. 8 % of the unit cube in 6-dimensional space, 0.0026 cm^{10} in 10-dimensional space, and $3.28\text{e-}21 \text{ cm}^{40}$ in the 40-dimensional space. It can then be seen that starting by some dimension n , say 5 or 6 and some index i , the i -th nearest neighbor does not lie in such a ball around point x but somewhere “in the corner” in that cube but outside the ball (the boundary effect [1]). It follows from it that this i -th neighbor lies farther from point x than it would follow from the uniformity of distribution.

Let us look at function $f(i) = r_i^n$, where r_i is the mean distance of the i -th neighbor from point x . The function grows linearly with index i in the case of uniform distribution without the boundary effect mentioned. In the other case this function grows slower than linearly and therefore we suggest choosing function $f(i) = r_i^q$, where $q \leq n$ is a suitable power discussed later.

3 Uniform Distributions without Boundary Effects

Let us assume random and uniform distribution of points in some subspace S of E_n . Further suppose that point x is inside S in the following sense: For each neighbor y of point x , the ball with its center at x and the radius equal to $\|x-y\|$ lies inside S . This is the case where the boundary effects do not take place.

In this chapter one-dimensional case is studied and it is shown that the distance of the i -th nearest neighbor is given by Erlang distribution; the multidimensional case is the subject of the next part of this paper.

Points spread on a line randomly and uniformly

In this case the distance Δ between two neighbor points is a random variable with exponential distribution function $P(\Delta) = 1 - e^{-\lambda\Delta}$ and probability density $p(\Delta) = e^{-\lambda\Delta}$ [4], [6]. For this distribution the mean is $E\{\Delta\} = 1/\lambda = d$ and it is the mean distance between two neighbor points.

We can look at the query point x from two points of view, x as randomly dropped on a line with randomly and uniformly spread points, or as point x being chosen from the points on this line. We show that there is no essential difference.

Random point dropped on a line with randomly and uniformly spread points

Now imagine a line with randomly and uniformly spread points, i.e. the distance between two neighbor points is such as discussed in the previous paragraph. Let there be a new point x dropped randomly and with uniform distribution function on this line. The question is what is the distribution of the distance of this point to the nearest point on the line.

Let the interval, to which the point x is dropped be (a, b) . Then the length of the interval between two (fixed) neighbor points on a line has exponential distribution with $\lambda = 1/d$ [13], [14].

A new point is dropped somewhere between some already existing pair of points. If there was mean distance between neighbor points d and if there were N points and the line was limited to interval (a, b) , the mean distance is $d \frac{N}{N+1}$ now. For $N \rightarrow \infty$, the mean distance goes to d . It is the same case as that

of one point selected from the already existing points, see the next paragraph.

Randomly chosen point on a line with randomly and uniformly spread points

Let there be a query point x selected from the points already existing. The mean distance between neighbor points is again d . The question now is what is the distribution of the distance of this point to the nearest point on a line. Let x_1 be the distance of point x to the nearest neighbor, i.e. the lesser of the distances of point x either to the point on the left or to the point on the right. The length of the interval between the neighbor points is thus broken to a half and x_1 has exponential distribution with $1/\lambda = d/2$, i.e. one half of the mean distance of the neighbor points on the line.

The second, third, etc. nearest neighbor

Let us sort consecutive nearest neighbors from point x in an ascending order. The mean distance between two successive points is $d/2$. It is the same situation as in the previous case. From it it follows that the distance between two successive points has the exponential distribution with $\lambda = 2/d$. $1/\lambda$ corresponds to one half of the mean distance of the neighbor points on the line. We have already the query point x and its nearest neighbor denoted "1" in Fig. 1. The question is what the distance of the second, third, ... nearest neighbor from point x is for the case above. The situation is illustrated in Fig. 1.

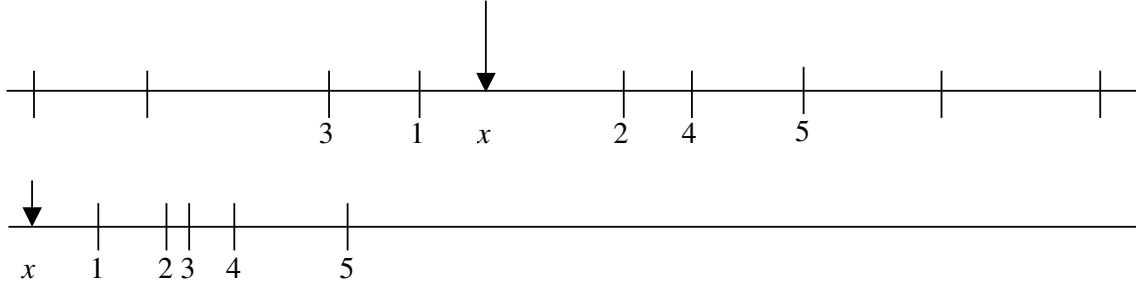


Fig. 1. Neighbors of point x on a line with randomly and uniformly spread points (atop) and sorted according to the distance from point x (below).

In Fig. 1 the upper line shows some random points and point x on the line. The nearest neighbors are denoted by numbers - the first, the second etc. The bottom line shows these neighbors in their order and the true distances from point x . Let the difference in the distances of two successive neighbors on the bottom line, the $(i-1)$ -st and the i -th be denoted $x_{i-1,i}$. With the exception of point x , the points are distributed independently and uniformly but there are twice more points on the same length. We can see that $x_{i-1,i}$, $i = 2, 3, \dots$ and x_1 are all independent random variables. Then the mean distance between them, i.e. between two successive points, is $d/2$. This shows that the distance between two successive points on the bottom line has exponential distribution with $\lambda = 2/d$, i.e. $1/\lambda$ corresponds to one half of the mean distance between the neighbor points on the line.

Composed distribution

Let the true distance of the i -th neighbor be denoted x_i . Let the difference in the distances of two successive neighbors, the $(k-1)$ -st and the k -th be denoted $x_{k-1,k}$. Then it holds that

$$x_i = x_1 + \sum_{k=2}^i x_{k-1,k}$$

because the distance is simply a sum of all successive differences. It is also a sum of independent exponentially distributed random variables. The probability density of the sum of independent random variables is given by convolution of the probability densities of these variables. Assuming exponential distribution of individual random variables, the composed distribution has Erlang distribution $\text{Erl}(i, \lambda)$.

Our problem was studied in connection with mass service (queuing) systems. The problem of i independent exponential servers working in series (cascade) is solved in [9]. The servers have the same exponential distribution of the service time with constant λ . Generally, the resulting total service time after the i -th server is given by Erlang distribution $\text{Erl}(i, \lambda)$ [4], [9] $p_i(x) = \frac{1}{i!} \lambda^i x^{i-1} e^{-\lambda x}$.

In our case the service time of one server is analogous to the difference of distances between two successive neighbors (also the distance between two neighbor points on the bottom line in Fig. 3) and the total time after the i -th server is analogous to the distance of the i -th point from point x . It holds that $\lambda = 2/d$ because there is one half of the mean distance of the neighbor points on the line mentioned above.

For statistical distribution of distances of the first, the second, the third, ..., the i -th nearest point from point x with $\lambda = 2/d$ it holds that $\text{Erl}(1, \lambda) = \text{Exp}(\lambda)$, $\text{Erl}(2, \lambda) = \lambda^2 x e^{-\lambda x}$, etc. It will be proved later.

4 Probability distribution mapping function

To study a probability distribution of points in the neighborhood of a query point x in n -dimensional Euclidean space E_n , let us construct individual balls around point x embedded one into another like peels of onion. The radii of individual balls can be expressed by formula $r_i = \text{const.}^{1/n} V_i$. A mapping between the mean density ρ_i in an i -th peel and its radius r_i is $\rho_i = p(r_i)$. $p(r_i)$ is the mean probability density in the i -th ball peel with radius r_i . The probability distribution of points in the neighborhood of a query point x is thus simplified to a function of a scalar variable. We call this function probability distribution mapping function $D(x, r)$, and its partial derivation according to r the distribution density mapping function $d(x, r)$.

Definition

In E_n the distribution mapping function $D(x, r)$ of the neighborhood of the query point x is function $D(x, r) = \int_{B(x, r)} p(z) dz$, where r is the distance from the query point and $B(x, r)$ is the ball with center x and radius r .

Definition

In E_n the distribution density mapping function $d(x, r)$ of the neighborhood of the query point x is function $d(x, r) = \frac{\partial}{\partial r} D(x, r)$, where $D(x, r)$ is a distribution mapping function of the query point x and radius r .

Note. It can be seen that for fixed x the function $D(x, r)$, $r > 0$ is monotonously growing from zero to one. Functions $D(x, r)$ and $d(x, r)$ for x fixed are one-dimensional analogs to the probability distribution function and the probability density, respectively.

4.1 Nearest Neighbors in E_n

A number of points in a spherical neighborhood with a center in the query point x and the probability distribution mapping function $D(x, r)$ grows with the n -th power of distance from the query point. Let us denote this n -th power of distance from the query point $d_{(n)}$. More generally let the q -th power of distance from the query point be $d_{(q)}$.

Definition

Let a, b be distances of two points from a query point x in E_n . Let $d_{(q)} = |a^q - b^q|$.

Note that we differentiate between the $d_{(q)}$ and distance $d = |a - b|$ sometimes with indexes d_k, d_{k-1} , and so on.

Uniform distribution

Let a query point x be surrounded by other points uniformly distributed to some distance d_0 from the query point x . It means that there is a ball $B(x, d_0)$ with points uniformly distributed inside it. Up to distance d_0 we deal with uniform distribution. In this case the number of points in a ball neighborhood with center in the query point x grows with the n -th power of distance from the query point (up to distance d_0). At the same time, the probability distribution mapping function $D(x, r)$ also grows with the n -th power of distance from the query point x .

Using $d_{(n)}$ instead of r , both the number of points and the $D(x, r)$ grows linearly with r^n . The distribution density mapping function $d(x, r^n)$ taken as $\frac{\partial}{\partial (r^n)} D(x, r^n)$ is constant, see Fig 2.

It can be seen that $d_{(n)}$ of successive neighbors is a random variable with exponential distribution function. The $d_{(n)}$ of the i -th nearest neighbor from the query point is given by the sum of $d_{(n)}$ between the successive neighbors. Then, it is a random variable with Erlang distribution $\text{Erl}(d_{(n)}, i, \lambda)$, $\lambda =$

$1/\bar{d}_{(n)}$, where $\bar{d}_{(n)}$ is a mean $d_{(n)}$ between the successive neighbors. The only difference is that instead of distance r the n -th power of distance is used in an n -dimensional Euclidean space and then $d_{(n)} = r_i^n - r_{i-1}^n$, $r_0^n = 0$. This is proved in the next paragraph. For illustration see Fig. 2.

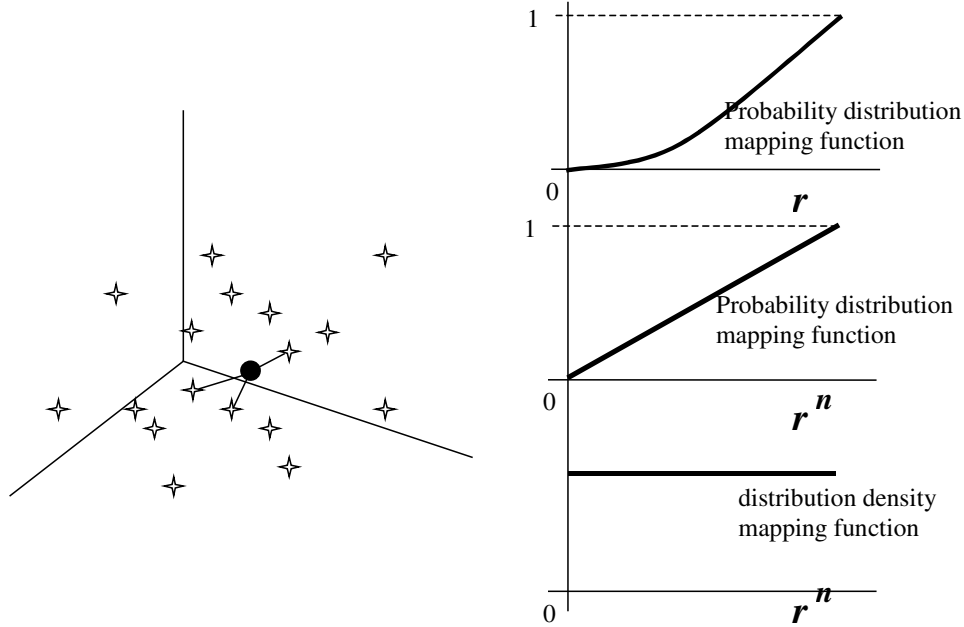


Fig. 2. The probability distribution mapping function $D(x, r)$ and $D(x, r^n)$, and the distribution density mapping function $d(x, r^n)$ for uniform distribution of points in E_n .

Distribution of $d_{(n)}$ of the k -th nearest neighbor

Let us suppose a ball in an n -dimensional space containing uniformly distributed points over its volume.

Theorem

For each query point $x \in E_n$ let a distance r exists, such that in ball $B(x, r)$ with center x and radius r the points will be spread uniformly with mean value of $d_{(n)}$ between two nearest neighbors equal to \bar{d} and let $\lambda = 1/\bar{d}$. Then the $d_{(n)}$ of the k -th nearest neighbor of point x is a random variable with Erlang distribution $\text{Erl}(d_{(n)}, k, \lambda)$, i.e.

$$F(d_{(n)}) = 1 - \exp(-\lambda d_{(n)}) \sum_{j=0}^{k-1} \frac{(\lambda d_{(n)})^j}{j!}$$

$$f(d_{(n)}) = \frac{\lambda^k}{k!} d_{(n)}^{k-1} \exp(-\lambda d_{(n)})$$

Proof

Let us denote the i -th nearest neighbor of point x by x_i , its distance from point x by d_i , its $d_{(n)}$ by $d_{(n)i}$. Let us introduce a mapping $Z: x_i \rightarrow R_1+$: $Z(x_i) = d_{(n)i}$, i.e. points x_i are mapped to points on the straight line, the query point x to point 0.

Let probability density of points in $B(x, r)$ be $p = \text{const}$ according to the assumption of uniformity. Let us build concentric balls with center in point x and radii $\rho < r$, i.e. inside ball $B(x, r)$. The number of points in ball $B(x, \rho)$ is $N_\rho = \int_{B(x, \rho)} p \, dv$, where dv is an element of the volume of ball $B(x, r)$, and

$N_\rho = p S_n \rho^n$ where $S_n \rho^n$ is the volume of ball $B(x, r)$; S_n is the constant dependent on dimension n only.

Let the total points x_i in ball $B(x, r)$ be N . Then $p = \frac{N}{S_n r^n}$ and from it $N_\rho = N \frac{\rho^n}{r^n}$. From it follows that the number of points in distance ρ from point x grows linearly with ρ^n and with proportionality constant

$\lambda = N/r^n$. Therefore let us measure the distance in $d_{(n)}$, it means not in centimeters (cm) but in cm^n . From it follows that in mapping Z points x_1, x_2, \dots are distributed randomly and uniformly, and mean $d_{(n)}$ of the neighbor pairs is equal to $r^n / N = 1/\lambda$. We have then uniform distribution of the points on $d_{(n)}$, then the $d_{(n)}$ between the neighbor points has exponential distribution [13], [14] with parameter λ , and then the $d_{(n)}$ of the k -th point x_k from point x is given by the sum of $d_{(n)}$ between successive neighbors. Then it is a random variable that is a sum of random variables with identical exponential distribution [4], [9] with parameter λ , then $d_{(n)}(x_k) = \text{Erl}(d_{(n)}, k, \lambda)$. \square

Example of a Uniform Ball

Let us suppose a ball in an n -dimensional space containing uniformly distributed points over its volume. Let us divide the ball on concentric “peels” of the same volume. Using the formula $r_i = S(n)/2^{1/n} \sqrt[n]{V_i}$ we obtain a quite interesting succession of radii corresponding to individual volumes. Symbol $S(n)$ denotes the volume of a ball with unit radius in E_n ; note $S(3) = 4/3\pi$. The higher space dimension leads to consideration that more similar values of the radii approaching the outer ball radius may be obtained. The inner part of the ball seems thus to be nearly empty, i.e. without points other than x ; it is shown in Table 1. Note that in E_3 , a ball of radius 4.642 cm, has the same volume as the last peel between radii 9.655 cm and 10 cm which is only 0.345 cm thick, and, at the same time, has a ten times smaller volume than a ball of radius 10 cm. In E_{15} these differences are much larger. From this example we can easily see that true distances give a rather strange picture on distribution of points in the neighborhood of the query point. It gives impression as if all neighbor points were concentrated in the outer part of the ball. It is really so [1], but it does not holds for density of points. The $d_{(n)}$ or radius to the n -th power corresponds much better to the probability distribution mapping function $D(x, r) = \text{const} \cdot r^n$. The radii of balls to the n -th power grow thus linearly. Differences are then constant, which corresponds well to uniformity of the distribution.

Table 1. Distances of the first till the tenth nearest neighbors in E_3 and in E_{15} .

For a three-dimensional ball:									
4.642	5.848	6.694	7.368	7.937	8.434	8.879	9.283	9.655	10
For a fifteen-dimensional ball:									
8.577	8.983	9.229	9.407	9.548	9.665	9.765	9.852	9.930	10

Unit cube with uniform distribution

Let a unit cube $(-0.5, 0.5)^n$ be given. Distribution of points is uniform inside the cube and there are no points outside it. Let us consider four special positions of the query point x :

- x lies in the origin, i.e. in the center of this cube
- x lies in the center of one wall of the cube
- x lies in distance l ($0 < l < 0.5$) from the center of the cube on the straight line connecting the center of the cube and the center of one of its walls
- x lies in a corner of the cube.

All these cases are illustrated in Fig. 3, where lines are drawn approximately for E_3 but captions are valid for general dimension n as shown in Table 2 below. Solid lines show parts where the course of functions is known exactly, dashed lines denote parts where the course is not known exactly because it varies from case to case and, moreover, it is difficult to derive formulas for it. These parts are either monotonously increasing or monotonously decreasing.

Case 2 (see Table 2) is very similar to “half-cube” $(-0.5, 0.5)^{n-1} \times (0, 0.5)$ with the query point in the origin; the values of b and c are slightly different: $a = 0.5^n$, $b = (0.5\sqrt[n]{n})^n$, $c = S(n)/2^n$.

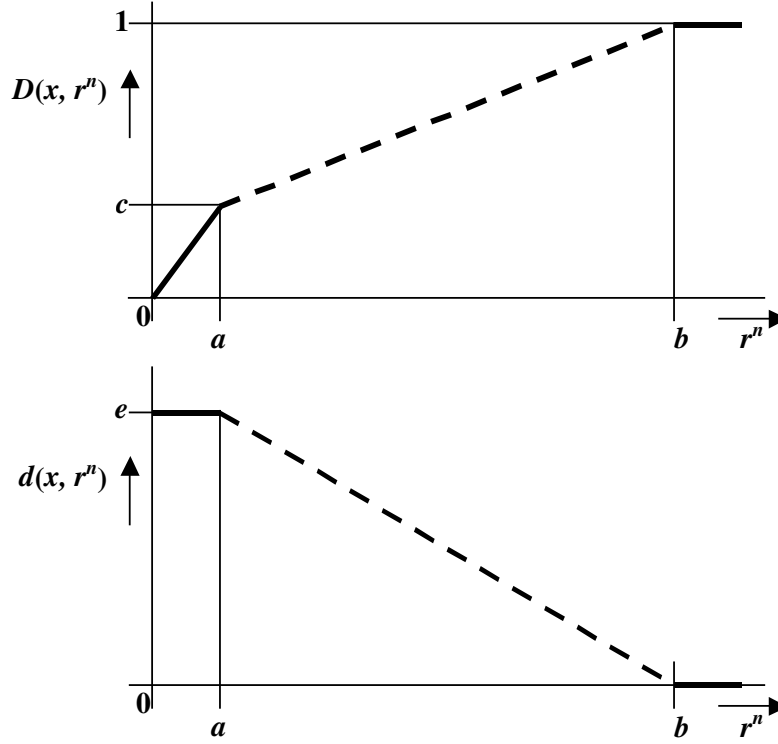


Fig. 3. Distribution mapping function and distribution density mapping function for four special positions of the query point x in the unit cube. Values of a , b , c are in Table 2 and it holds that $e = c/b$. The solid lines show the true course of the functions, the dash line shows that the true course is not known exact.

Table 2. Values of parameters a , b , c of the distribution mapping function and the distribution density mapping function for four special positions of the query point x in the unit cube. $S(n)$ is the volume of a ball with unit radius.

Case	Position of the query point x	a	b	c
1	in the center of the cube	0.5^n	$(0.5\sqrt{n})^n$	$S(n)/2^n$
2	in the center of one wall of the cube	0.5^n	$\sqrt{((n-1)0.5^2+1)}^n$	$\frac{1}{2}S(n)/2^n$
3	in the distance l ($0 < l < 0.5$) from the center of the cube on the straight line connecting the center of the cube and the center of one of its walls	$(0.5-l)^n$	$\sqrt{((0.5+l)^2+(n-1)0.5^2)}^n$	$\frac{(1-2l)}{n}S(n)/2^n$
4	in a corner of the cube	1	\sqrt{n}^n	$S(n)/2^n$

4.2 Influence of dimensionality and a total number of points

Distances, as well as $d_{(n)}$, of several nearest neighbors depend on probability distribution $p(z)$ of points in the neighborhood of a query point x and also on the true density of the points in this neighborhood. It is clear that for the same function $p(z)$ and a small number of samples the neighbors lie farther from the query point x and also one from the other than if there were lots of points distributed with the same $p(z)$.

Data considered

Let the data set of total m_T samples be given in the form of a matrix X_T with m_T rows and n columns. Each sample corresponds to one row of X_T and, at the same time, corresponds to a point in n -dimensional Euclidean space E_n , where n is the sample space dimension.

Example of uniform distribution in a ball of radius R in E_n

Let the query point x be in the origin. The distribution density mapping in E_n is given by the constant function $d(x, r^n) = 1/R^n$ for $r \in (0, R)$ and zero otherwise. Then $D(x, r^n) = r^n/R^n$ for $r \in (0, R)$ and $D(x, r^n) = 1$ for $r > R$. $D(x, r^n)$ can be considered as a probability distribution function. Let there be m_T samples with this distribution. Let r_i^n be mean r^n for the i -th nearest neighbor of the query point x . It holds that $D(x, r_i^n) = i/m_T$. Thus we have $r_i^n = iR^n / m_T$.

Example of normal distribution

Let us consider a distribution of m_T samples with distribution density mapping function $d(x, r^n) = 2N(0,1)|_{r^n \geq 0} = 2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}r^{2n}}$ in E_n . Coefficient 2 was introduced to get $D(x, r_n) \rightarrow 1$ for $r_n \rightarrow \infty$.

Thus $D(x, r_n)$ can be considered as a probability distribution function. For the mean r_i^n of r^n for the i -th nearest neighbor of the query point x it holds that $r_i^n = N^{-1}(0.5 + 0.5i / m_T)$.

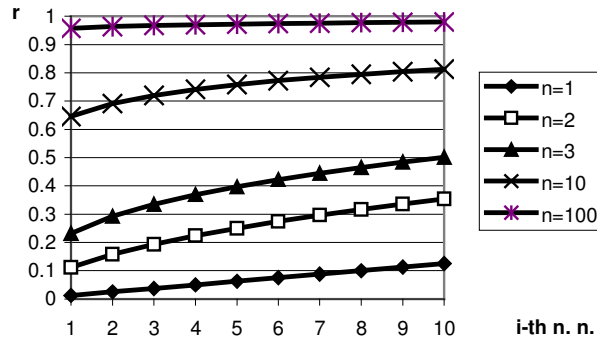


Fig. 4. Distances r_i^n of the first several nearest neighbors for different space dimensions n .

Fig. 4 shows distances r_i^n of the first several nearest neighbors for different space dimensions n . Fig. 5 shows distances r_i^n for different numbers of points (samples in the set) for different space dimensions n .

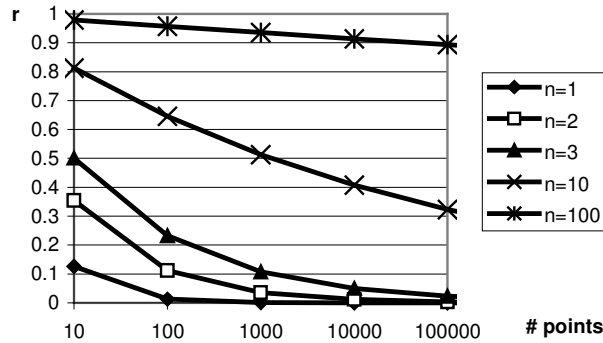


Fig. 5. Distances r_i^n for different numbers of points for different space dimensions n .

5 Influence of boundary effects

The problem of boundary effects was studied in [1] in l_{max} metric and for different problems of searching the nearest neighbor in large databases. Taking the boundary effects into account, the

estimation of the searching time was lesser than if no boundary effect was considered and was closer to reality.

5.1 Boundary effect phenomenon

Suppose that for any query point x in E_n and any distribution of points there is a spherical neighborhood with radius r (it corresponds to $d_{(n)} = r^n$) so that the probability density in this spherical neighborhood can be considered to be uniform.

Let the boundary effect be understood as a phenomenon that

the distribution density mapping function $d(x, r^n)$ is a (not strictly) decreasing function and decreases starting from some point a , see Fig. 4 for illustration;

this fact influences the mean distances of neighbors of the query point so that these distances do not correspond to uniform distribution, i.e. that the mean $d_{(n)}$ of two successive neighbors of the point x are not constant.

It can be seen that three facts influence the boundary effect, the character of the probability distribution of points, the position of the query point, and the total number of points m_T , i.e. the size of the data set. The boundary effect is demonstrated on the example of a uniform cube in Fig. 3.

The nearest neighbor influenced by boundary effect

In this part we deal with the finite volume around the query point x with uniformly distributed points. There are no points outside the volume considered. For some first $k-1$ nearest neighbors of point x , the ball $B(x, d_{k-1})$ with center x and the radius d_{k-1} equal to the distance of the $(k-1)$ -st neighbor lies whole in the volume considered. Then no boundary effect arises. If for the k -th nearest neighbor the ball $B(x, d_k)$ does not lie whole in the volume considered, the distance d_k is influenced because the ball $B(x, d_k)$ is not uniform any more.

Let us consider the example of a unit cube with uniform distribution with four special positions of the query point x already discussed. The task is now: There are given radius a and the number of points (samples) m_T , what is the index k of the k -th nearest neighbor to the query point x such that just for this k -th nearest neighbor the boundary effect arises, i.e. for some constant C it holds that $r_i^n = iC$ for $i = 1, 2, \dots, k-1$ and $r_k^n \neq kC$?

Let the distribution density mapping function on interval $(0, a)$ be constant $d(x, r^n) = e$. Then, on interval $(0, a)$, the distribution mapping function is $D(x, r^n) = e \cdot r^n$ and let $D(x, r^n) \rightarrow 1$ for $r^n \rightarrow \infty$. Thus $D(x, r^n)$ can be considered as a probability function. It holds that $D(x, r_i^n) = i/m_T$, i.e. $e \cdot r_i^n = i/m_T$. Setting $r_i^n = a$ we get $k = 1 + \lfloor e \cdot a \cdot m_T \rfloor$, where $\lfloor \cdot \rfloor$ denotes the integer part.

Boundary effect in a uniform unit cube

Example: For the unit cube with uniform distribution and the query point x in the origin, i.e. in the center of this cube, there is $ae = c = S(n)/2^n$, then $k = 1 + \lfloor S(n)m_T / 2^n \rfloor$, see Fig. 6.

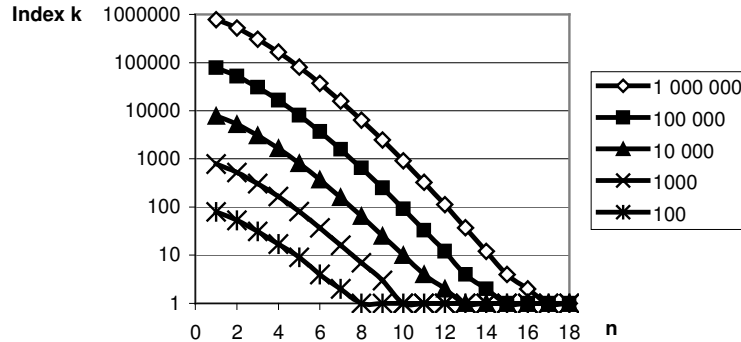


Fig. 6. Index k of the k -th nearest neighbor for which the boundary effect arises for different dimensionalities n and for $m_T = 100$ till 1,000,000 points (a unit cube with uniform distribution and the query point x in the origin)

5.2 Power approximation of the probability distribution mapping function

Here we consider situations where there are boundary effects or non-uniform distribution. A course of the probability distribution mapping function is often not known and it is not easy to derive it analytically. Therefore, we suggest using power approximation r^q of the probability distribution mapping function.

Definition

Power approximation of the distribution mapping function $D(x, r^n)$ is function r^q such that $\frac{D(x, r^q)}{r^q} \rightarrow \text{const}$ for $r \rightarrow 0+$. The exponent q is a distribution-mapping exponent. The variable $\alpha = q/n$ is called distribution mapping ratio.

Note. We often omit a multiplicative constant of the distribution mapping function.

Using approximation of the distribution mapping function by $D(x, r^n) = \text{const} \cdot (r^n)^\alpha$, the distribution mapping exponent is $q = n\alpha$. The distribution-mapping exponent reminds Grassberger-Procaccia's correlation dimension [7], [12]. There are two essential differences. First, the distribution-mapping exponent is local feature of the data space because it depends on position of query point; the correlation dimension is a feature of the whole data space. Second, the distribution mapping exponent is related to data only, not to any assumption about attractor behind them – of course, attractor influences its value and its distribution over the data space.

Power approximation for normal distribution

Chap. "Normal n -dimensional distribution" has shown on an example that true distances of the first nearest neighbor till the tenth one lie in the distance between 0.65 cm and 0.98 cm for $n = 10$ till $n = 100$. If the interval for r^n is chosen from zero to one, the approximation is rather good with $\alpha = 0.85$ and maximal difference 0.0164 between $D(x, r^n)$ and its power approximation. This approximation is shown in Fig. 7.

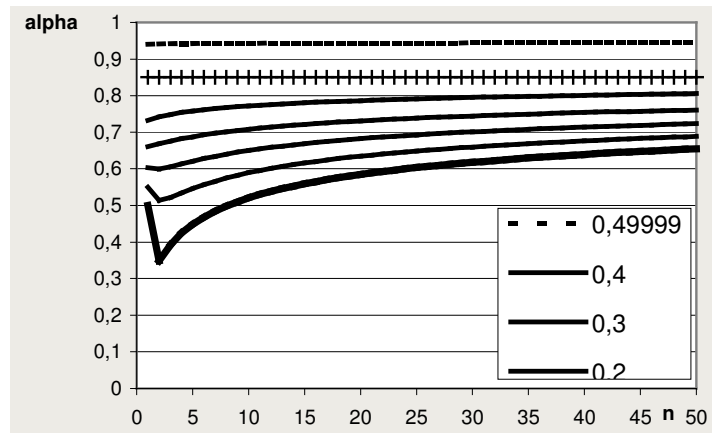


Fig. 7. Distribution mapping ratio α as a function of dimension n for normal distribution and for the uniform unit cube, with the query point in the distance l ($0 < l < 0.5$, see legend) from the center of the cube to the center of one of its walls.

The case of very small l is nearly identical to four other cases discussed in Section "Unit cube with uniform distribution".

5.3 Power approximation with the constant distribution density mapping function

In this part we assume that power approximation is applicable to the N -th nearest neighbor, i.e. the distribution density mapping function is constant as far as to the N -th nearest neighbor from point x .

Theorem

Let for each query point $x \in E_n$ exist a distribution mapping exponent q and a distance r such that in ball $B(x, r)$ with center x and radius r the points are spread with mean $d_{(q)}$ between two nearest neighbors equal to δ and let $\lambda = 1/\delta$. Then the $d_{(q)}$ of the k -th nearest neighbor of point x is a random variable with Erlang distribution $\text{Erl}(d_{(q)}, k, \lambda)$, i.e.

$$F(d_{(q)}) = 1 - \exp(-\lambda d_{(q)}) \sum_{j=0}^{k-1} \frac{(\lambda d_{(q)})^j}{j!}$$

$$f(d_{(q)}) = \frac{\lambda^k}{k!} d_{(q)}^{k-1} \exp(-\lambda d_{(q)}) \cdot$$

Proof

Let us denote the i -th nearest neighbor of point x by x_i , its distance from point x by d_i , its $d_{(q)}$ by $d_{(q)i}$. Let us introduce a mapping $Z: x_i \rightarrow R_1+$: $Z(x_i) = d_{(q)i}$, i.e. points x_i are mapped to points on the straight line, the query point x to point 0. Let the total points x_i in ball $B(x, r)$ be N . Then the distribution mapping function is $d(x, \rho^q) = \text{const} \frac{N}{r^q}$. From the assumption it follows that the number of points in distance ρ

from point x grows linearly with ρ^q . Then there is a proportionality constant $\lambda = N/r^q$. From it it follows that in mapping Z points x_1, x_2, \dots are distributed randomly and uniformly, and mean $d_{(q)}$ of the neighbor pairs is $r^q / N = 1/\lambda$. We have then uniform distribution of points on $d_{(q)}$ and then the $d_{(q)}$ between the neighbor points has exponential distribution [13], [14] with parameter λ , then $d_{(q)}$ of the k -th point x_k from point x is given by the sum of $d_{(q)}$ between successive neighbors. Then it is a random variable that is the sum of random variables with identical exponential distribution [4], [9] with parameter λ , then $d_{(q)}(x_k) = \text{Erl}(d_{(q)}, k, \lambda) \cdot \square$

Fig. 9 illustrates the theorem see the next paragraph.

5.4 Simulation analysis

Simulation analysis was used to find a distribution-mapping exponent valid globally for the whole data set considered. From this different point of view, the differences in the results were compared with special cases studied before.

For all simulations there were 32000 samples used with a uniform distribution in the unit n -dimensional cube. With some exception, the behavior of the first 10 nearest neighbors of the query point x was studied. After the samples were generated, each of them was taken as a query point x and for each sample 10 nearest neighbors were found and their distances r_i , $i = 1, 2, \dots, 10$ recorded. After that a distribution mapping exponent q was found as a value for which the mean values of r_i^q grow approximately linearly, i.e. successive differences of their means $d_{qi} = E(r_i^q) - E(r_{i-1}^q)$, $r_0^q = 0$ are approximately constant. The values of q are shown in the second column of Table 3. The values of q for $n = 1, 20$ and 40 were used for analytic approximation by function $q = u + vn + w\sqrt{n}$, in numbers $q = -1.114337 + 0.467465n + 1.646872\sqrt{n}$. As the value of this function is larger than n for $n = 2, 3$ and 4 , values $q = n$ are used in these three cases.

Table 3. Values of the distribution mapping exponent q for some uniform n -dimensional cubes found by simulation and approximated.

n	q found by simulation	q approximated
1	1	1
5	4.8	4.90
10	9	8.77
20	15.6	15.6
40	28	28

Fig. 8 shows differences $d_{qi} = E(r_i^q) - E(r_{i-1}^q)$, $r_0^q = 0$ for $n = 20$ and three different exponents. We can see that the value $q = 15.6$ is a proper value of the distribution mapping exponent.

Fig. 8 shows that the boundary effect influences the mean positions of neighbors also for the whole set of points. If there were no boundary effect, the line for power equal to 20 would be a straight

horizontal line. It clearly follows from the theory but it is very difficult to arrange simulation with exclusion of the boundary effect, especially for such dimension as $n = 20$.

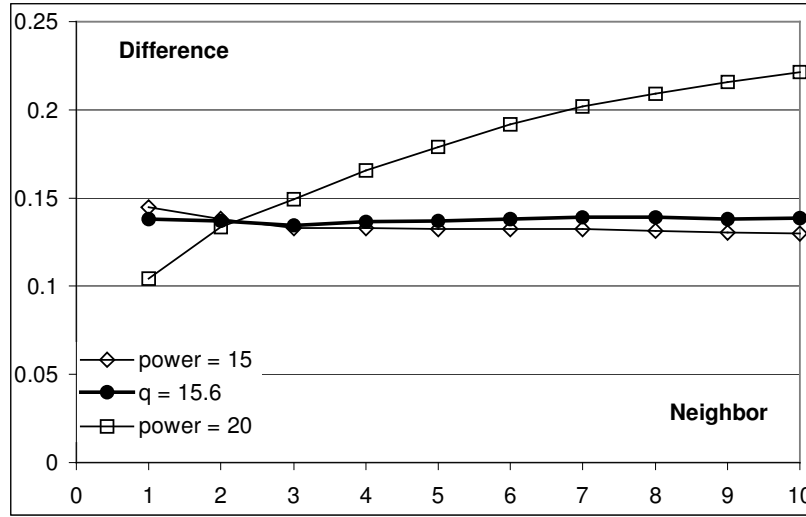


Fig. 8. Differences $d_{qi} = E(r_i^q) - E(r_{i-1}^q)$, $r_0^q = 0$ for $n = 20$ and three different exponents.

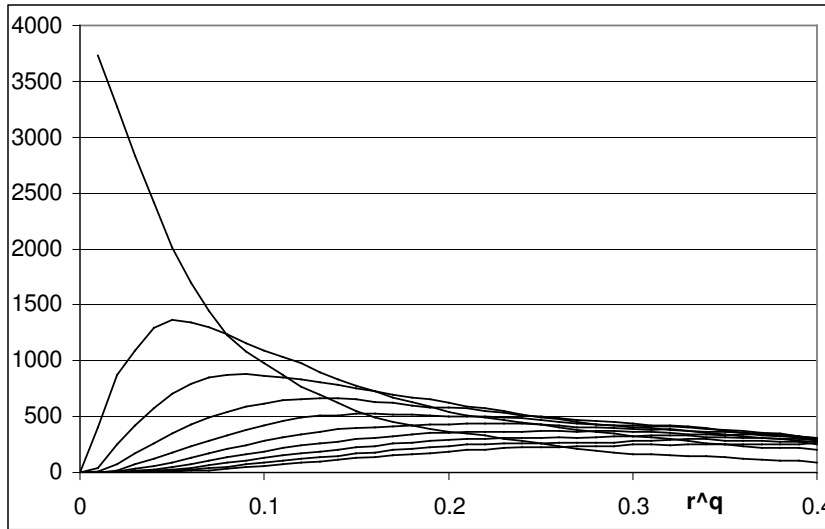


Fig. 9. Histograms of d_q of variable r_i^q , $q = 15.6$, $n = 20$. Lines from top to bottom $i = 1, 2, \dots, 10$.

In Fig. 9 histograms of d_q of variable r_i^q for $q = 15.6$, $n = 20$ and for the first 10 neighbors are shown. The histograms were smoothed using averaging over five values. The bin size is 0.01. It can be seen that the histograms show, in fact, probability density functions for Erlang distribution for indexes 1 (exponential) to 10. This is just what was expected according to Theorem 2 when the distribution mapping function with a proper distribution mapping exponent was used.

6 Conclusion

By using a notion of distance, i.e. a simple transformation $E_n \rightarrow E_1$, the problems with dimensionality are easily eliminated at a loss of information on the true distribution of points in the neighborhood of the query point. It is known [1], [13], [14] that for larger dimensions something like local approximation of real distribution by uniform distribution does not exist. But the assumption of at least local uniformity in the neighborhood of a query point is usually inherent in methods based on the distances of neighbors.

This problem is solved by introduction of a power approximation of the probability distribution mapping function here. An essential variable of this approximation is the distribution mapping exponent. By using this exponent the real distribution is transformed to be uniform. It is possible to do it either locally or globally. A local approach was shown for several special cases of positions of the query point in a uniform hypercube and on multinomial normal distribution. The global case was shown for uniform hypercube. The results in these two approaches differ, of course, but on the other hand, they correspond well to each other. In essence, there are two ways in estimating the distribution mapping exponent. One of them is to estimate this exponent globally for the whole data set and rely on not too large local differences. The other way is to estimate the distribution mapping exponent locally, i.e. for each query point anew. A disadvantage of this approach is a large possible error in the distribution mapping exponent when a small number of neighbor points is used. Processing of a larger number of points, on the other hand, makes estimation closer to global estimation, especially for small data sets.

Acknowledgement

This work was supported by the Ministry of Education of the Czech Republic under project No. LN00B096.

References

- [1] S. Arya, D.M. Mount and O. Narayan, Accounting for boundary effects in nearest neighbor searching, *Discrete and Computational Geometry*, Vol. 16 (1996) 155-176.
- [2] K. Beyer et al., When is "nearest neighbor" meaningful?, in: *Proc. of the 7th International Conference on Database Theory*, (Jerusalem, Israel, 1999) 217-235.
- [3] E. Chávez, K. Figueroa, G. Navarro, A fast algorithm for the all k nearest neighbors problem in general metric spaces, *Scientific Literature Digital Library CiteSeer*, <http://citeseer.nj.nec.com/correct/462760>.
- [4] J.C. Demaret and A. Gareet, Sum of exponential random variables. in: *AEÜ*, Vol. 31, No. 11 (1977) 445-448.
- [5] R.O. Duda, P.E. Hart and D.G. Stork, *Pattern classification*, Second Edition, John Wiley and Sons, Inc., (New York, 2000).
- [6] W.T. Eadie et al., *Statistical methods in experimental physics*, North-Holland (1982).
- [7] P. Grassberger and I. Procaccia, Measuring the strangeness of strange attractors, *Physica*, Vol. 9D, (1983) 189-208.
- [8] A. Hinneburg, C.C. Aggarwal and D.A. Keim, What is the nearest neighbor in high dimensional spaces?, in: *Proc. of the 26th VLDB Conf.*, (Cairo, Egypt, 2000) 506-515.
- [9] L. Kleinrock, *Queueing Systems*, Vol. I: Theory, (John Wiley & Sons, New York, 1975).
- [10] V. Pestov, On the geometry of similarity search: Dimensionality curse and concentration of measure, in: *Information Processing Letters*, Vol. 73, No. 1-2, (2000) 47-51.
- [11] B.W. Silverman, *Density estimation for statistics and data analysis*, (Chapman and Hall, London, 1986).
- [12] E.W. Weisstein, Correlation Dimension, *Interactive Mathematics Encyclopedia MathWorld*, <http://mathworld.wolfram.com/CorrelationDimension.html>.
- [13] E.W. Weisstein, Poisson Distribution, *Interactive Mathematics Encyclopedia MathWorld*, <http://mathworld.wolfram.com/PoissonDistribution.html>.
- [14] E.W. Weisstein, Poisson Process, *Interactive Mathematics Encyclopedia MathWorld*, <http://mathworld.wolfram.com/PoissonProcess.html>.