SF 2535 LECTURE 18

ON THE NEAREST NEIGHBOR IN VERY LARGE METRIC SPACES

T. Kuśki

2017

1 NOTATIONS

$x = (x_1, \dots, x_d)^T \quad x \in \mathbb{R}^d$

$D(x) = \sum_{i=1}^{d} x_i^2 = x^T x$ = SQUARED DISTANCE TO THE ORIGIN $0 \in \mathbb{R}^d$

2 SKEWNORMAL DISTRIBUTION

$X \sim 2\, \phi_d(x)\, \Pi(x)$

$\phi_d(x) = \frac{1}{(2\pi)^{d/2}} e^{-x^T x/2} \quad (= N(0, I_d))$

$\Pi(x)$ = A SKEWING FUNCTION

i) $0 \le \Pi(x) \le 1$

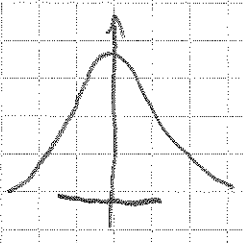
ii) $\Pi(-x) = 1 - \Pi(x)$

EXAMPLE $d = 1$

$\Pi(x) = \Phi(\lambda x) \qquad \Phi(x) = \int_{-\infty}^{x} \phi_1(u)\, du$

$-\infty < x < +\infty$

$2\, \Phi(\lambda x)\, \phi_0(x)$



$\lambda = 0 \qquad 2 \cdot \Phi(0, x)\, \phi_0(x) = 2 \cdot \frac{1}{2}\, \phi_0(x) = \phi_0(x)$

□

THM 1

$$x \sim \pi(x)\,\varphi_d(x) \Rightarrow D = x^T x \sim \chi^2_d$$

Proof: [1] ∎

χ^2_d = CHISQUARED WITH d DEGREES OF FREEDOM

$$f_D(d) = \frac{1}{\Gamma\left(\frac{d}{2}\right)}\, d^{\frac{d}{2}-1} \left(\frac{1}{2}\right)^{d/2} e^{-d/2}, \quad 0 \le d$$

FACTS

i) $E[D] = d$ ii) $\mathrm{Var}(D) = 2d$

See [2] ∎

3) CONVERGENCE IN PROBABILITY [3]

a) $X_1, \ldots, X_n, \ldots$ A SEQUENCE OF RANDOM VECTORS CONVERGES IN PROBABILITY TO $a \in \mathbb{R}^d$ IFF

$$\lim_{n \to \infty} P(\|X_n - a\| > \varepsilon) = 0$$

FOR ANY $\varepsilon > 0$.

WE WRITE

$$X_n \xrightarrow{P} a, \text{ cuando } n \to \infty$$

b) CRAMÉR - SLUTSKY'S THEOREM

i) IF $X_n \xrightarrow{P} a$ CUANDO $n \to \infty$ AND $g(x)$ IS A CONTINUOUS FUNCTION

$g(x): \mathbb{R}^d \to \mathbb{R}$

AT a, THEN

$g(X_n) \to g(a)$

CUANDO $n \to \infty$.

ü) $X_n \xrightarrow{P} a \ (\in \mathbb{R}) \quad X_n \in \mathbb{R}$,

$Y_n \xrightarrow{P} b \ (\in \mathbb{R}) \quad b \neq 0 \quad Y_n \in \mathbb{R}$

THEN $\dfrac{X_n}{Y_n} \xrightarrow{P} \dfrac{a}{b}$, as $n \to \infty$.

3) CHEBYSHEV'S INEQUALITY

$E[X] = \mu \qquad Var(X) = \sigma^2$

(X REAL VALUED R.V.)

$$P(|X - \mu| > \varepsilon) \leq \frac{1}{\varepsilon^2} \sigma^2$$

4) THE THEOREM [★]

$\underline{X}_1, \dots, \underline{X}_n$ $\quad$ $\underline{X}_i$ i.i.d. $\sim \pi(\underline{x}) g_d(\underline{x})$

$D_i = D(\underline{X}_i) \quad i = 1, \dots, n$

$D_{min} = \min_{1 \le i \le n} D_i \qquad D_{max} = \max_{1 \le i \le n} D_i$

THM 2. For any $\varepsilon > 0$

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

PROOF: SET $\mu = E[D_i]$ (AS $\underline{X}_i$ ARE i.i.d.)

AND $\boxed{y_i \triangleq \frac{D_i}{\mu}}$

THEN $E[y_i] = \frac{1}{\mu} E[D_i] = 1.$

$$Var[y_i] = \frac{1}{\mu^2} Var[D_i] = \frac{2s}{\mu^2} = \frac{2}{d}$$

THEN CHEBYSHEV'S INEQUALITY TELLS THAT FOR EVERY ε

$$\mathbb{P}(|\gamma_i - 1| > \varepsilon) \leq \frac{1}{\varepsilon^2} \operatorname{Var}(\gamma_i) = \frac{2}{d\varepsilon}.$$

$\therefore$ $\mathbb{P}(|\gamma_i - 1| > \varepsilon) \leq \frac{2}{d\varepsilon} \to 0$, as $d \to \infty$

i.e. $\gamma_i \xrightarrow{\mathbb{P}} 1$, as $d \to \infty$

d = DIMENSION OF THE EUCLIDEAN SPACE

HENCE, BY CRAMER'S THEOREM (i),

$\max(\gamma_1, \ldots, \gamma_n) \xrightarrow{\mathbb{P}} 1$, as $d \to \infty$

$\min(\gamma_1, \ldots, \gamma_n) \xrightarrow{\mathbb{P}} 1$, as $d \to \infty$

(SINCE MAX AND MIN ARE CONTINUOUS FUNCTIONS.

THEN

$$\frac{D_{max}}{D_{min}} = \frac{\max(D_1, \dots, D_n)}{\min(D_1, \dots, D_n)}$$

$$= \frac{\frac{1}{\mu}\max(D_1, \dots, D_n)}{\frac{1}{\mu}\min(D_1, \dots, D_n)}$$

$$= \frac{\max\left(\frac{D_1}{\mu}, \dots, \frac{D_n}{\mu}\right)}{\min\left(\frac{D_1}{\mu}, \dots, \frac{D_n}{\mu}\right)}$$

$$= \frac{\max(Y_1, \dots, Y_n)}{\min(Y_1, \dots, Y_n)}$$

$$\xrightarrow{P} 1, \text{ as } d \to \infty$$

(BY CRAMÉR - SLUTSKY ii)

THUS:

$$P\left(D_{max} \leq (1+\varepsilon) D_{min} \right)$$

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

$$= P\left(\left| \frac{D_{max}}{D_{min}} - 1 \right| \leq \varepsilon \right) \quad , \text{ as } D_{max} > D_{min}$$

BUT $\frac{D_{max}}{D_{min}} \xrightarrow{P} 1$, as $d \to \infty$

AND HENCE

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

WE GET

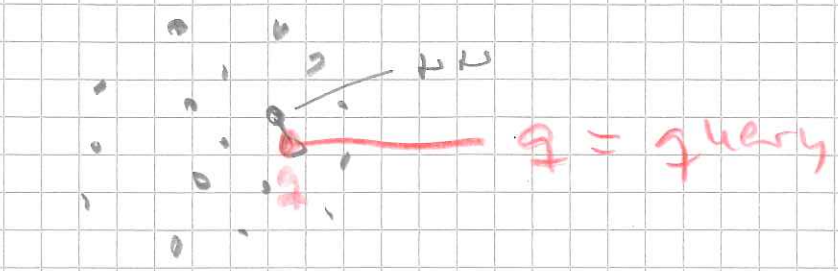
$$\lim_{d \to \infty} P\left(\left| \frac{D_{max}}{D_{min}} - 1 \right| \leq \varepsilon \right) = 1 - \lim_{d \to \infty} P\left(\left| \frac{D_{max}}{D_{min}} - 1 \right| > \varepsilon \right)$$

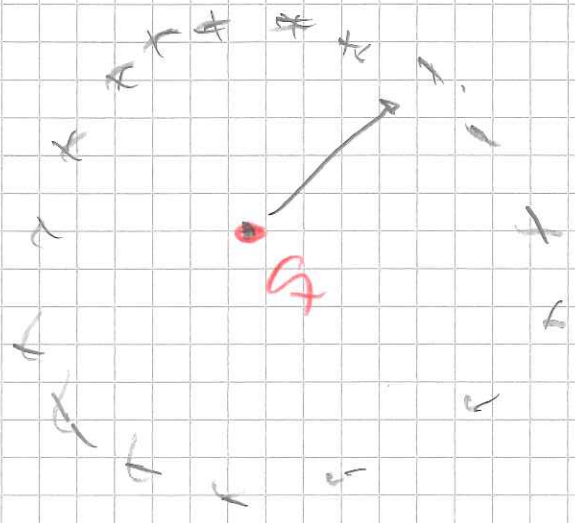
$$= 1 - 0 = 1.$$

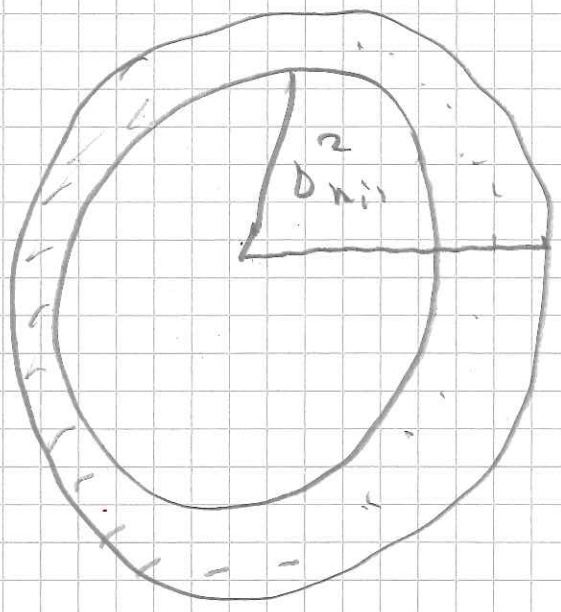
∎

NN

HENCE : Nearest Neighbor IS UNSTABLE IN HIGH DIMENSION





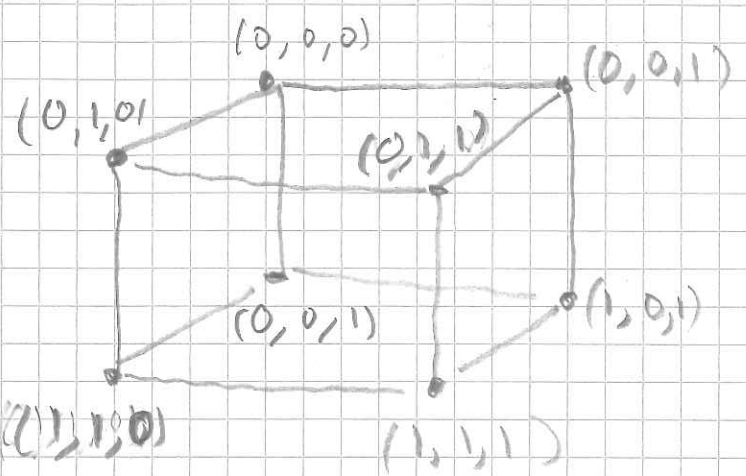


10/23

5) BINARY HYPERCUBE

$\mathcal{U}$ = THE BINARY HYPERCUBE

$= \{ x \mid x = (x_1, \dots, x_d),\ x_i \in \{0,1\} \}$



$d = 3$

$2^d = 8$

$|\mathcal{U}| = 2^d$ = CARDINALITY OF $\mathcal{U}$

I.E. THERE ARE 2^d VERTICES $\underline{x}$ IN $\mathcal{U}$

TAKE $X_i \sim Be(1/2)$ i.e. $\frac{1}{2} P(X_i = 1) = P(X_i = 0)$

FOR ALL $i = 1, \dots, d$ INDEPENDENTLY

$\therefore \underline{x} \in \mathcal{U} : \mathbb{P}(\underline{x}) = \frac{1}{2^d}$

HAMMING DISTANCE

x	y	$x \oplus y$
1	1	0
1	0	1
0	1	1
0	0	0

$\underline{x} \in U, \ \underline{y} \in U$

$$d(\underline{x}, \underline{y}) = \sum_{i=1}^{d} x_i \oplus y_i \ \left(= \sum_{i=1}^{d} |x_i - y_i| \right)$$

(real numbers)

= NUMBER OF POSITIONS WHERE x_i AND y_i DISAGREE OR WHERE $\underline{x} \neq \underline{y}$.

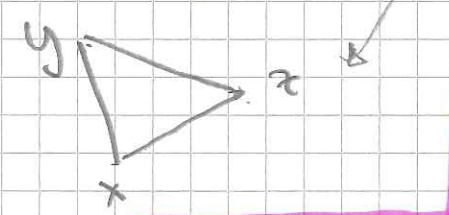
$d(\underline{x}, \underline{y})$ IS A METRIC ON $U \times U$

1) $d(\underline{x}, \underline{y}) \geq 0$, $d(\underline{x}, \underline{y}) = 0 \iff \underline{x} = \underline{y}$

2) $d(\underline{x}, \underline{y}) = d(\underline{y}, \underline{x})$

3) $d(\underline{x}, \underline{y}) \leq d(\underline{x}, \underline{z}) + d(\underline{z}, \underline{y})$

$d(\underline{x}, \underline{y})$ IS CALLED THE HAMMING DISTANCE.



TRIANGLE INEQUALITY IS READILY PROVED.

TAKE A FIXED $y \in U$

AND DRAW (INDEPENDENTLY OF y)

$$x \sim P(x) = \frac{1}{2^d}$$

THEN IT FOLLOWS

LEMMA

$$\Delta(x, y) \sim \text{Bin}\left(d, \frac{1}{2}\right)$$

PROOF: 1) d TRIALS WITH TWO OUTCOMES

$x_i = 0$ OR $x_i = 1$

SUCCESS = $\{x_i \neq y_i\}$ FOR TRIAL i

2) $P(x_i \neq y_i) = \frac{1}{2} = p$

SINCE y IS GIVEN, IF $y_i = 1$, $P(x_i = 0) = \frac{1}{2}$

$y_i = 0$ $P(x_i = 1) = \frac{1}{2}$

3) p IS THE SAME FOR ALL TRIALS

4) THE NUMBER OF TRIALS, d,

IS FIXED IN ADVANCE.

5) TRIALS ARE INDEPENDENT

1) ~~5)~~ IMPLY THAT , d,

THE CLAIM IN THE

HENCE LEMMA FOLLOWS; ■

$d(x,y)$ = THE NUMBER OF SUCCESSES IN d TRIALS $\sim Bin(d, \frac{1}{2})$.

■

BUT

$$X \sim Bin(n,p) \Rightarrow E[X] = np$$
$$Var[X] = np(1-p)$$

HENCE

$$E_x[d(x,y)] = \frac{d}{2} \qquad Var_x[d(x,y)] = \frac{d}{4}$$

FIX $y \in U$ AND TAKE AN I.I.D SAMPLE $x_1, \dots, x_n$ FROM $P(x)$

$$D_i := d(x_i, y) \qquad i = 1, \dots, n$$

$$D_{max} = \max_{1 \le i \le n} (D_1, \dots, D_n)$$

$$D_{min} = \min_{1 \le i \le n} (D_1, \dots, D_n)$$

BUT THM 2 ABOVE DEPENDS CRITICALLY ON THE FACT THAT

$$\frac{\mathrm{Var}(D_i)}{(E(D_i))^2} \to 0, \text{ as } d \to \infty$$

SHOULD HOLD.

HERE

$$\frac{\mathrm{Var}(D_i)}{(E(D_i))^2} = \frac{d/4}{d^2/4} = \frac{1}{d} \to 0$$

AS $d \to \infty$. BUT THEN WE SEE BY THE PROOF OF THM 2 THAT

THM 3,
IF $D_1, \ldots, D_n$ ARE INDEPENDENT HAMMING DISTANCE TO AN $y \in U$,
THEN

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

WE SEE CLEARLY THAT THIS HOLDS FOR <u>ANY</u> $y \in U$.

SOME CONCEPTS OF $\mathcal{U}$.

A VERTEX $x \in \mathcal{U}$ IS ORTHOGONAL TO A VERTEX $y \in \mathcal{U}$ IF AND ONLY IF

$$d(x,y) = \frac{d}{2}$$

- ~~x S.T. $d(x,y) \to \frac{d}{2}$ FOR A FIXED y~~ ARE SAID TO LIE IN THE EQUATOR.
- $0 \leq d(x,y) \leq d$
- $d(x,y) = d$ IFF $x = (1,1,\dots,1)$, $y = (0,0,\dots,0)$

BALL WITH RADIUS r AND CENTER IN y

$$B_r(y) = \{ x \in \mathcal{U} \mid d(x,y) \leq r \}$$

WE CAN ROTATE U SO THAT ANY OF ITS POINT IS ORIGIN (A NORTH POLE) $y = 0$

THE POINTS AT THE DISTANCE

$$d/2$$

FROM THE POLE y ARE CALLED THE EQUATOR.

WE RECALL THAT

$$\text{Bin}(n, p) \simeq N(np, \underbrace{np(1-p)}_{\text{NB VARIANCE}})$$

IF n IS LARGE.

HENCE WE GET THAT

$d(x, y)$ IS APPROXIMATELY $N\left(\frac{d}{2}, \frac{d}{4}\right)$

WHEN d IS LARGE.

SET $D = d(x, y)$

17/23

$$D \sim N\left(\frac{d}{2}, \frac{d}{4}\right)$$

WE COMPUTE THE PROBABILITY THAT D LIES IN

$$B_{\frac{d}{2}+k}(y) \setminus B_{\frac{d}{2}-k}(y)$$

$$= \{x \mid \tfrac{d}{2}-k \le d(x,y) \le \tfrac{d}{2}+k\}$$

OR DIFFERS FROM MEAN DISTANCE TO y BY AT MOST $\pm k$ BITS

$$P\left(\tfrac{d}{2}-k < D < \tfrac{d}{2}+k\right) =$$

$$= P\left(-k < D - \tfrac{d}{2} < +k\right)$$

$$= P\left(\frac{-k}{\sqrt{\frac{d}{4}}} < \underbrace{\frac{D-\frac{d}{2}}{\sqrt{\frac{d}{4}}}}_{\sim N(0,1) \text{ since } D \sim N(d/2, d/4)} < \frac{+k}{\sqrt{d/4}}\right)$$

$$\Phi(-x) = 1 - \Phi(x)$$

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

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

$$\rho_{ud} \cong 2\Phi\left(\frac{u}{\sqrt{d}}\right) - 1$$

$d = 100$

$u = 3$ $\rho_{ud} = 0.4515$

$u = 12$ $\rho_{ud} = 0.9836$

$u = 15$ $\rho_{ud} = 0.9973$

$d = 1000$

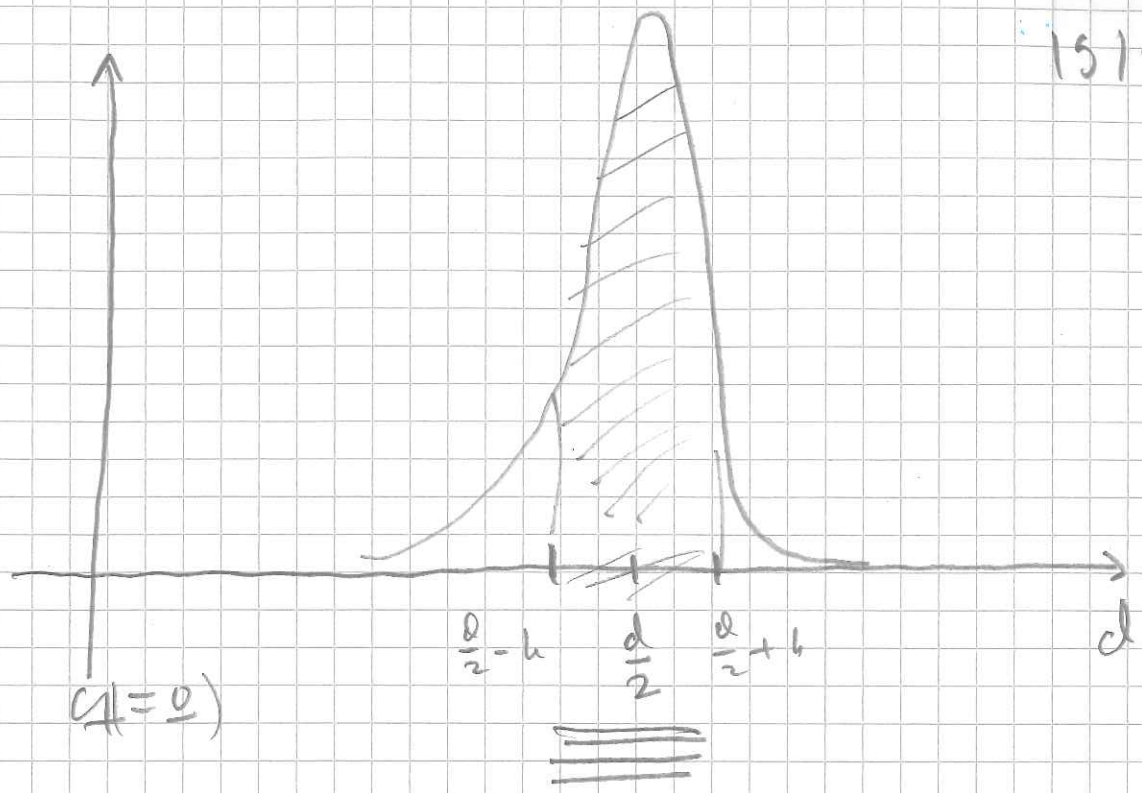
$u = 11$ $\rho_{ud} = 0.51$

$u = 12$ $\rho_{ud} = 0.55$

$u = 59$ $\rho_{ud} = 0.9998$

$d = 10000$

$u = 116$ $\rho_{ud} = 0.9998$



MOST OF THE PROBABILITY MASS IS LOCATED AROUND THE EQUATION FOR LARGE d.

ρ = INTRINSIC DIMENSIONALITY (S. BOZIN et. al.)

$$\rho = \frac{(E[d(x,y)])^2}{2\,Var[d(x,y)]}$$

ONE MORE OBSERVATION

IF $\alpha < 1/2$ AND

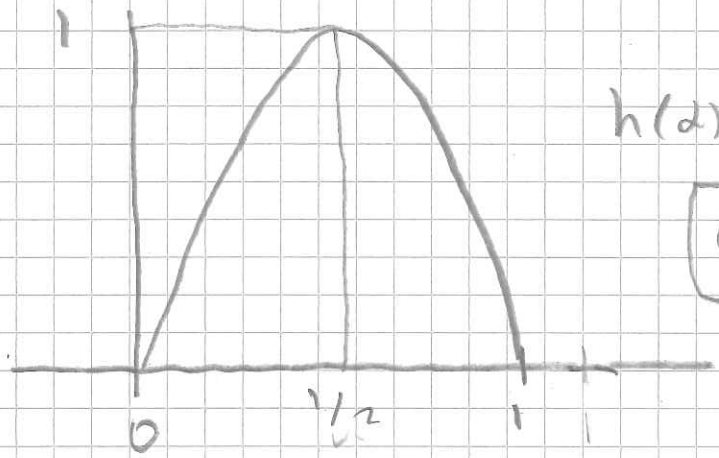
$$\sum_{k=0}^{\alpha d} \binom{d}{k} < 2^{d h(\alpha)}$$

$k = \alpha d$

$\alpha = k/d$

WHERE $h(\alpha) \triangleq -\alpha \log_2(\alpha) - (1-\alpha)\log_2(1-\alpha)$

FOR $0 \le \alpha \le 1$, IS THE ISING ENTROPY FUNCTION



$h(\alpha)$, $0 \le \alpha \le 1$

$0 \log_2 0 = 0$

$$\therefore P(B_k(y)) = P_x(d(x,y) \le k)$$

$$= \sum_{k=0}^{K} \binom{d}{k} \frac{1}{2}^k \left(\frac{1}{2}\right)^{d-k} = \frac{1}{2^d} \sum_{k=0}^{k} \binom{d}{k}$$

$$\le \frac{1}{2^d} 2^{d h(K/d)} \qquad (\alpha = k/d)$$

$$P(B_k(y)) = \frac{|B_k(y)|}{2^d} \leftarrow \text{NUMBER OF } x \text{ IN}$$

$$= \frac{\text{VOLUME OF } B_k(y)}{\text{TOTAL VOLUME}}$$

OR

$$\textcircled{II}\quad |B_k(y)| \le \frac{1}{2^{d(1-h(k/d))}} = \left(\frac{1}{2}\right)^{d(1-h(k/d))}$$

BECOMES SMALL WHEN d INCREASES FOR GIVEN k, SINCE

$$d(1-h(k/d)) \to \infty \text{ AS } d \to \infty$$

APPENDIX

$$h\left(\frac{k}{d}\right) = -\frac{k}{d}\log_2\frac{k}{d} - \left(1-\frac{k}{d}\right)\log_2\left(1-\frac{k}{d}\right)$$

$$d(1-h(k/d))$$

$$= d + d\left[\left(\frac{k}{d}\log_2\left(\frac{k}{d}\right) + \left(1-\frac{k}{d}\right)\log_2\left(1-\frac{k}{d}\right)\right)\right]$$

$$= d + k\log_2(k/d) - (d-k)\log_2\left(1-\frac{k}{d}\right)$$

$$= d - k\log_2(d) - \underbrace{\left(d\log_2\left(1-\frac{k}{d}\right) + k\log_2\left(1-\frac{k}{d}\right)\right)}_{=C(d)}$$

$$= \log_2\left(1-\frac{k}{d}\right)^d \to 0, \text{ AS } d\to\infty$$

$$\to \log_2 e^{-k} \text{ AS } d\to\infty$$

I.E.

$$d(1-h(k/d)) = d - k\log_2(d) + \underbrace{C(d)}_{\text{BOUNDED in } d}$$

WHEN d INCREASES, $d - k\log_2(d) \to +\infty$

□

HENCE WE SEE THAT FOR ARBITRARY RADIUS $k_1 < d/2$

$\mathbb{P}(\mathcal{B}_k(y))$

WILL BE VERY SMALL IF

d IS SUFFICIENTLY LARGE,

$=$

ANOTHER ANALYSIS OF THESE [7] PHENOMENA IN $\mathbb{R}^d$ IS FOUND IN [6]

REFERENCES

[1] WANG, BOYER, LEWTON:
STATISTICS AND PROBABILITY
LETTERS 66, 2004, 385-398

[2] WEIKENGREN: RADE: BETA

[3] A. GUT: AN INTERMEDIATE COURSE IN
PROBABILITY, 2ND ED.
2005

[4] BEYER, GOLDSTEIN,
RAMAKRISHNAN, SHAFT:
"WHEN IS 'NEAREST NEIGHBOR'
MEANINGFUL?"
LECTURE NOTES IN COMPUTER SCIENCE
1540; PP. 217-235, 1998

[5] S. BRIN: NEAR NEIGHBOR SEARCH
IN LARGE METRIC
SPACES
INT'L CONFERENCE ON VERY
LARGE DATA BASES VLDB
1995

[6] BLUM, HOPCROFT, KANNAN:
CHAPTER 2 in
FOUNDATIONS OF DATA SCIENCE