ST 2535 LECTURE 20

# k-NN REGRESSION IN HIGH DIMENSIONS

2017 - 12 - 14

T. Kypraios

# 0. INTRODUCTION

$$Y = f(X) + \varepsilon$$

$$\hat{f}_{opt} = \underset{f}{\operatorname{argmin}} \; E\left[(Y - f(X))^2\right]$$

$$\Rightarrow \hat{f}_{opt}(x) = E[Y \mid X] = \int_{-\infty}^{\infty} y \, p_{Y|X}(y|x) \, dy$$

We discuss learning of $\hat{f}_{opt}$ IF WE DO NOT KNOW THE DISTRIBUTION $p_{Y|X}(y|x)$.

WE STUDY NEAREST-NEIGHBOR REGRESSION IN HIGH DIMENSIONAL SPACES

$$\hat{f}_{opt}(x) \approx \hat{f}_k(x) = \text{Average}(y_i \mid x_i \in nn_k(x))$$

$nn_k(x)$ = k nearest neighbors

$=$

HERE $x$ LIES IN THE UNIT HYPERCUBE

C.F. EXERCISES 4.7

EXERCISE 4 pp. 164 - 165

IN JAMES - WITTEN - HASTIE TIBSHIRANI

AN INTRODUCTION TO STATISTICAL LEARNING

(Project 1)

# 1. BASIC ASSUMPTIONS AND NOTATIONS

$\mathcal{X}$ = DATA SPACE = $[0,1]^d$

= A HYPERCUBE IN $\mathbb{R}^d$

$x \in \mathcal{X}$, $y \in \mathcal{X}$ $x, y$ points

$$d(x,y) = \|x-y\|_2 = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$

= $L_2$- distance

$\mathcal{D} = \{x_1, \dots, x_N\}$ A DATA SET $x_i \in \mathcal{X}$

**ASSUMPTION**

THE DATA POINTS $x_1, \dots, x_N$ ARE I.I.D

& SAMPLED OF $\underline{X}_1, \dots, \underline{X}_N$ respectively

$\underline{X}_i \in \mathcal{U}([0,1]^d)$ = UNIFORM DIST'N

$$\text{Vol}(\mathcal{X}) = 1$$

$\underline{X} = (\underline{X}_1, \dots, \underline{X}_d)$ $\quad \underline{X}_i \sim \mathcal{U}(0,1)$

$A \subseteq [0,1]^d$

$$P(\underline{X} \in A) = \int\int \dots \int_A dx_1 \dots dx_d$$

$$= \text{Vol}(A).$$

## HYPERBALL (in $\mathbb{R}^d$)

$$Sp^d(x,r) = \{ y \in \mathbb{R}^d \mid \|y-x\|_2 \leq r \}$$

$q \in \mathbb{R}^d$ $nn(q)$ = NEAREST NEIGHBOR TO $q$ in $\mathcal{D}$.

IS DEFINED AS

$$nn(q) = \{ x_i \in \mathcal{D} \mid \forall x' \in \mathcal{D} : \|x_i - q\|_2 \leq \|x' - q\|_2 \}$$

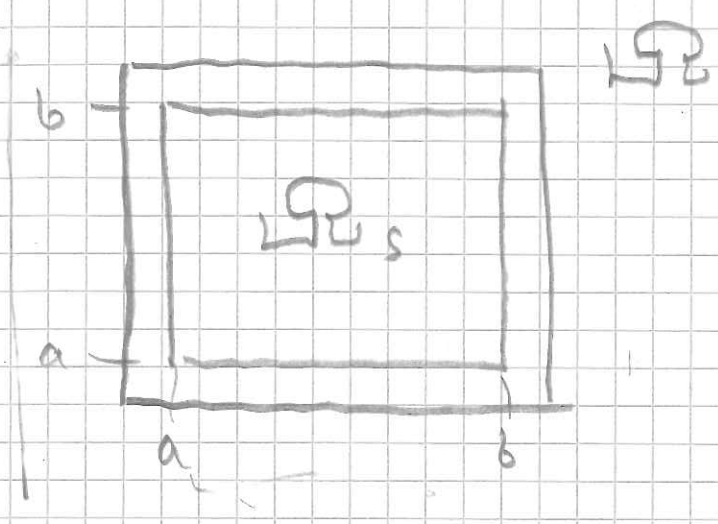
$$nn^{dist}(q) = \| nn(q) - q \|_2$$

## 2. FACTS

FACT 1. PROBABILITY OF A SUBCUBE

TAKE $s < 1$ AND $b - a = s$

$$\square_s = \{ x \in \square \mid a \le x_i \le b \quad i = 1, \dots, d \}$$




$$P((X_1, \dots, X_d) \in \square_s)$$

$$= P(a \le X_1 \le b, a \le X_2 \le b, \dots, a \le X_d \le b)$$

ASSUMPTION

$$= P(a \le X_1 \le b) \cdot P(a \le X_2 \le b) \cdot \dots \cdot P(a \le X_d \le b)$$

$$= \left( P(a \le U \le b) \right)^d \qquad U \in U[0,1]$$

$$= (b-a)^d = s^d$$

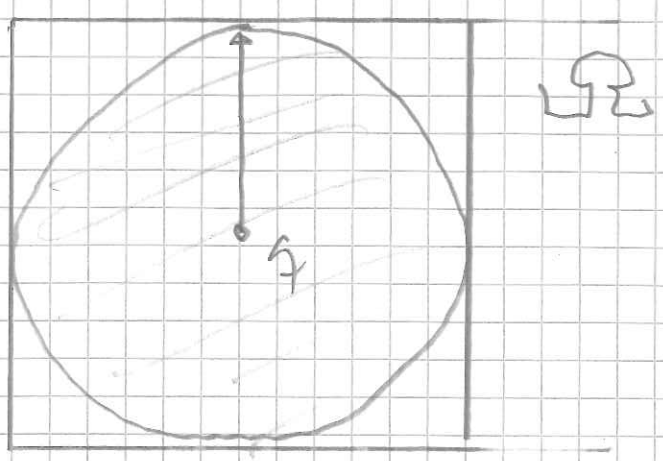
EX. $d = 100$, $s = 0.95$

$$s^d = 0.95^{100} \simeq 0.005 = 0.5\%$$

FACT 2

THE DATA SPACE IS VERY SPARSELY POPULATED.

FACT 2 THE LARGEST HYPERBALL THAT FITS ENTIRELY IN $\Omega$




$$sp^d\left(q, \tfrac{1}{2}\right) = \left\{ x \in \Omega \mid \|x - q\|_2 \leq \tfrac{1}{2} \right\}$$

SOUGHT: THE PROBABILITY THAT ONE DATA POINT LIES IN $sp^d(q, 1/2)$.

$$P\left(\bar{X} \in sp^d\left(q, 1/2\right)\right) = \int_{sp^d(q,1/2)} \cdots \int dx_1 \cdots dx_d$$

$$= K_d \left(\tfrac{1}{2}\right)^d$$

$K_d$: volume of the unit ball

$$sp^d(q, 1) = \left\{ x \in \mathbb{R}^d \mid \|x - q\|_2 \leq 1 \right\}$$

CHAPTER 2 in FOUNDATIONS OF DATA SCIENCE SHOWS THAT

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

$\Gamma(x+1) = x\,\Gamma(x)$ Euler's Gamma Function

STIRLING'S FORMULA

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

$$\Gamma\left(\frac{d}{2}\right) \sim \sqrt{2\pi}\, e^{-\frac{d}{2}} \cdot \left(\frac{d}{2}\right)^{\frac{d}{2}-\frac{1}{2}}$$

$$\therefore \Gamma\left(\overline{X} \in Sp^d\left(1, 1/2\right)\right)$$

$$\simeq \frac{\pi^{d/2}}{\sqrt{2\pi}} \frac{e^{d/2}}{\left(\frac{d}{2}\right)^{d/2-1/2}} \cdot \left(\frac{1}{2}\right)^{\frac{d}{2}}$$

$$\pi^{d/2} = e^{\frac{d}{2}\ln\pi} \qquad d/2^{\,d/2+1} = e^{\left(\frac{d}{2}+1\right)\ln d/2}$$

$$\simeq \frac{1}{\sqrt{2\pi}}\, e^{\frac{d}{2}(\ln\pi - 1) - \left(\frac{d}{2}+1\right)\ln d/2 \cdot \left(\frac{1}{2}d/2\right)}$$

$$\to 0 \quad \text{as} \quad d \to \infty$$

| $d$ | $P(\xi \in S_\gamma^d(0, 1/2))$ |
|---|---|
| 2 | 0.787 |
| 4 | 0.308 |
| 10 | 0.002 |
| 20 | $2.46 \cdot 10^{-8}$ |
| 40 | $3.278 \cdot 10^{-21}$ |
| 100 | $1.868 \cdot 10^{-70}$ |

FACT 2:

THE VOLUME OF $S_\gamma^d(0, 1/2)$

SHRINKS STRONGLY AS $d$ GROWS

AND IT IS INCREASINGLY

UNPROBABLE THAT ANY POINT

WILL BE FOUND IN THIS BALL

AT ALL!

## REMARKS:

$$\mathrm{vol}(\Omega) = 1$$

$$\mathrm{diameter}(\Omega) = \|\underline{1} - \underline{0}\|_2 = \sqrt{d}$$

$\underline{1} = (\underbrace{1, \dots, 1}_{d \text{ one}})$

$\underline{0} = (\underbrace{0, \dots, 0}_{d \text{ zeros}})$

$$\mathrm{vol}(S^d(\tfrac{1}{2})) \to 0 \text{ as } d \to +\infty,$$

$$\mathrm{diameter}(\Omega) \to +\infty \text{ as } d \to +\infty$$

=

FACT 3

$$p \equiv P(X \in S_p^d(q, 1/2))$$

$X_1, \dots, X_N$ I.I.D., N GIVEN IN ADVANCE

$$\text{SUCCESS} = P(X_i \in S_p^d(q, 1/2))$$

$Y$ = NUMBER OF SUCCESSES IN $X_1, \dots, X_N$

$$Y \sim \text{Bin}(N, p)$$

$$E[Y] = Np$$

$$E[Y] \geq 1 \iff \boxed{N \geq \frac{1}{p} = \frac{\Gamma(\frac{d}{2}+1)}{\pi^{d/2} \cdot (\frac{1}{2})^d}}$$

$\therefore$ $d = 10$, $N = 401,5$ SO THAT WE MAY EXPECT IN AVERAGE AT LEAST ONE POINT IN $S_p^d(q, 1/2)$

$d = 20$ $N = 4063162$

etc

**FACT 4 : WE COMPUTE**

$$P(Q,r) = P\left( nn(q) \leq sp^d(q,r) \right)$$

= PROBABILITY THAT $nn^{dist}(q) \leq r$.

$$P(Q,r) = 1 - P( nn^{dist}(q) > r )$$

$$= 1 - P( \|x_1 - q\|_2 > r, \dots, \|x_N - q\| > r )$$

$$= 1 - P(\|x_1 - q\|_2 > r) \cdot \dots \cdot P(\|x_n - q\| > r)$$

$$= 1 - \left( P(\|x - q\|_2 > r) \right)^N$$

$$= 1 - \left( 1 - P(\|x - q\|_2 \leq r) \right)^N$$

$$= 1 - \left( 1 - P( \{ x \in \mathbb{R}^d \mid \|x - q\|_2 \leq r \} ) \right)$$

$$= 1 - \left( 1 - Vol( sp^d(q,r) \cap [0,1]^d ) \right)^N$$

$r = 1/2$ $\quad sp^d(q, 1/2) \cap [0,1]^d = sp^d(q, 1/2)$

$d \to \infty$ $\quad P(Q,r) \simeq 0$

$\therefore$ $nn^{dist}(q)$ GROWS WITH $d$!

$$P(nn(q) \in S_p^d(q,r)) =$$

I.E. PROBABILITY THAT $nndist(q) \leq r$ =

$$= 1 - (1 - Vol(S_p^d(q,r) \cap \square))^N$$

1) $P(nn(q) \in S_p^d(q,r)) \to 1$,

as $N \to \infty$ (LARGE DATA SET)

FOR GIVEN $d$

2) $P(nn(q) \in S_p^d(q,r)) \to 0$,

as $d \to \infty$

3) BOTH $N \to \infty$, $d \to \infty$ $\Rightarrow$ ?

# 3. EXPECTED PREDICTION ERROR & BIAS-VARIANCE TRADE-OFF



## 3.1 IN GENERAL

$$Y = f(X) + \varepsilon \qquad E[\varepsilon] = 0,\ E[\varepsilon^2] = \sigma_\varepsilon^2$$

$\hat{f}$ = A MODEL OF $f$ ESTIMATED FROM TRAINING DATA

$$\mathcal{D} = (x_i, y_i)_{i=1}^{N} \quad \text{i.e.} \quad y_i = f(x_i) + \varepsilon_i$$

TAKE A FIXED $x$ AND CONSIDER

$$\text{EPE}(x) = E\left[(Y - \hat{f}(x))^2\right]$$

E = EXPECTATION W.R.T $P(x,y)$ GENERATING $\mathcal{D}$

"MEAN SQUARE ERROR FOR $\hat{f}$"

- $E[Y] = E[f(x) + \varepsilon]$

$$= f(x) + E[\varepsilon] = f(x)$$

- $\text{Var}[Y] = E[(Y - f(x))^2]$

$$= E[\varepsilon^2] = \sigma_\varepsilon^2.$$

EPE = EXPECTED PREDICTION ERROR

$$E\left[(Y - \hat{f}(x))^2\right]$$

$$= E\left[Y^2 - 2Y\hat{f}(x) + \hat{f}^2(x)\right]$$

$$= E[Y^2] - 2E[Y\hat{f}(x)] + E[\hat{f}^2(x)]$$

- $E[Y^2] = Var(Y) + (E(Y))^2$

$$= \sigma_\varepsilon^2 + f^2(x)$$

- $E[Y\hat{f}(x)] = E[(f(x) + \varepsilon)\hat{f}(x)]$

$$= E[f(x)\hat{f}(x)] + E[\varepsilon\hat{f}(x)]$$

$$= E[f(x)\hat{f}(x)] + E[\hat{f}(x)] \cdot \underbrace{E(\varepsilon)}_{=0}$$

$$= E[f(x)\hat{f}(x)] = f(x)E[\hat{f}(x)]$$

$$\therefore EPE(x) = \sigma_\varepsilon^2 + f^2(x) - 2f(x)E[\hat{f}(x)] + E[\hat{f}^2(x)]$$

- $E[\hat{f}^2(x)] = Var(\hat{f}(x)) + (E[\hat{f}(x)])^2$

∴ $EERR(x) =$

$$\sigma_\varepsilon^2 + \underbrace{Var(\hat{f}(x))}_{\text{Variance}} + \underbrace{\left(E[\hat{f}(x)] - f(x_i)\right)^2}_{\text{BIAS}^2}$$

Variance ↑ as $\hat{f}$ estimated from training data

BIAS²: model complexity error due to approximation

ETC.

## 3.2 KNN & NEAREST NEIGHBOR REGRESSION

$Y = f(X) + \varepsilon$ as above

$(x_i, y_i)_{i=1}^N$ training data

$$\hat{f}(x) = \hat{f}_k(x) = \frac{1}{k} \sum_{i:\, x_i \in nn_k(x)} y_i = \frac{1}{k} \sum_{nn_k(x)} y_i$$

$nn_k(x)$ = the set of k-th nearest neighbors

$$\|x_{(1)} - x\|_2 \leq \|x_{(2)} - x\|_2 \leq \dots \leq \|x_{(k)} - x\|_2$$

↑ nearest neighbor

$k=1$ $\hat{f}_k(x) = y_i$

Training data interpolation:

if $x_i = nn(x)$ $x_i \in D$

AS $N, k \to \infty$ SO THAT $\frac{k}{N} \to 0$

IT CAN BE SHOWN THAT

$$\hat{f}_k(x) \to E[Y|x]$$

RATE OF CONVERGENCE

$$O\left(\frac{1}{N^{1/d}}\right)$$

(Recall that

$$E\left[(Y - f(X))^2\right]$$

is minimized by $f(x) = E[Y|x]$)

UNIVERSAL ESTIMATOR! ☒

BUT: ASSUME $X_i \sim U([0,1]^d)$

$Y_i = f(x_i) + \varepsilon_i$ f UNKNOWN

- $P(X \in sp^{10}(0.8)) = 0.8^{10} \approx 0.1$

  80% OF RANGE IN EVERY DIMENSION TO COVER 10% OF DATA

- MOST DATA POINTS $X_1, \ldots, X_N$ ARE CLOSER TO BOUNDARIES THAN ANY OTHER DATA POINT

WE NEED

$$E\left[\hat{f}_k(x)\right] = \frac{1}{k} \sum_{n \in N_k(x)} E[Y_i]$$

$$= \frac{1}{k} \sum_{n \in N_k(x)} E[f(x_i) + \varepsilon_i] =$$

$$= \frac{1}{k} \sum_{n \in N_k(x)} f(x_i) + \underbrace{E[\varepsilon_i]}_{=0} = \frac{1}{k} \sum_{n \in N_k(x)} f(x_i)$$

$$= \frac{1}{k} \sum_{\ell=1}^{k} f(x_{(\ell)})$$

$x_1, \dots, x_N$ ORDERED BY $\|x_{(1)} - x\|_2 \leq \|x_{(2)} - x\|_2 \leq \dots \leq \|x_{(k)} - x\|_2 \leq \dots$

IND.

$$Var\left(\hat{f}_k(x)\right) \overset{\text{ind.}}{=} \frac{1}{k^2} \sum_{n \in N_k(x)} Var(Y_i)$$

$$= \frac{1}{k^2} \sum_{n \in N_k(x)} \sigma_\varepsilon^2 = \frac{1}{k^2} \cdot k \cdot \sigma_\varepsilon^2 = \frac{1}{k} \sigma_\varepsilon^2$$

$$1.^\circ \quad EPE(x) = \sigma_\varepsilon^2 + \left(f(x) - \frac{1}{k} \sum_{\ell=1}^{k} f(x_{(\ell)})\right)^2 + \frac{\sigma_\varepsilon^2}{k}$$

EXPECTED PREDICTION ERROR FOR $k$ NN - REGRESSION

IN SMALL DIMENSION, FOR SMALL $h$, THE FEW CLOSEST NEIGHBORS WILL HAVE THEIR FUNCTION VALUES $f(x_{(i)})$ CLOSE TO $f(x)$, SO THEIR AVERAGE SHOULD BE CLOSE $f(x)$

IN ADDITION

$$\max_x |E[Y|X=x] - \hat{f}_k(x)| \to 0$$

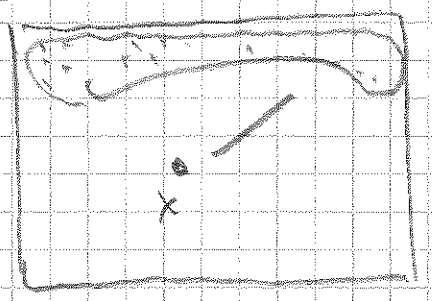
WITH PROBABILITY ONE AND WITH RATE $\approx O\left(\left(\frac{1}{n}\right)^{1/d}\right)$ ← SEE THAT DIFFERENT FROM 1

A UNIVERSAL APPROXIMATOR!

BUT : IN LARGE DIMENSION

$\mathcal{D} = \{x_1, \dots, x_N\}$, N SAMPLES $\mathcal{U}([0,1]^d)$

FOR A HIGH VALUE OF $d$ THE SAMPLES IN $\mathcal{D}$ ARE CLOSE TO THE BOUNDARY OF $[0,1]^d$ (as shown in 2.)



FOR ANY $k \in \mathbb{N}$, $\frac{1}{k} \sum_{\ell=1}^{k} f(x_{(\ell)})$

IS AN EXTRAPOLATION FROM NEIGHBORING SAMPLES RATHER THAN INTERPOLATION BETWEEN THEM

$$\left( f(x) - \frac{1}{k} \sum_{\ell=1}^{k} f(x_\ell) \right)^2$$

CAN BE VERY LARGE

CONVERGENCE RATE $\mathcal{O}\left(\frac{1}{N^{1/d}}\right)$

IS STILL VALID, BUT IS VERY SLOW.