② k-NN

Let $N_k(\vec{x}) = \{$ indices of k-NN of $\vec{x}$ in $\underline{D}\}$

k-NN predictor

→ Regression $\hat{y} = g(x) = \frac{1}{K} \sum_{i \in N_k(x)} y_i$

Predict unweighted average of neighbours

→ Classification $\hat{y} = g(\vec{x}) = \underset{\text{class } c}{\operatorname{argmax}} \; \underset{i \in N_k(\vec{x})}{\#}(y_i = c)$

$$= \underbrace{\underset{c}{\operatorname{argmax}} \sum_{i \in N_k(\vec{x})} I(y_i = c)}_{\text{unweighted majority vote}}$$

③ Distances

→ Most common

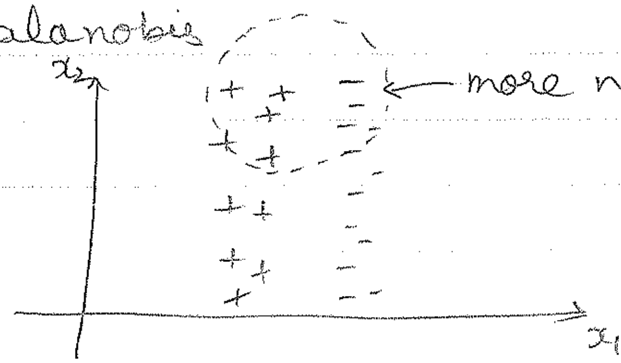
Euclidean Distance / L_2-norm of difference

$\vec{x}, \vec{z} \in \mathbb{R}^d$

$$d(\vec{x}, \vec{z}) = \left[\sum_{i=1}^{d} (x_i - z_i)^2 \right]^{1/2}$$

→ Let's generalize this in 2 ways

① Mahalanobis



New definition

$$d^2(\vec{x}, \vec{z}) = 10(x_1 - z_1)^2 + (x_2 - z_2)^2$$

↑ deviations in dim 1 should be penalized more

In general $$d^2(\vec{x}, \vec{z}) = \sum_{i=1}^{d} \sigma_i^2 (x_i - z_i)^2$$

$$= (\vec{x} - \vec{z})^T \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2 \end{bmatrix} (\vec{x} - \vec{z})$$

More generally,

$$d^2(\vec{x}, \vec{z}) = (\vec{x} - \vec{z})^T A (\vec{x} - \vec{z})$$

Set $A = I_{d \times d}$ ⇒ Euc. dist

Note $A \succeq 0$

positive semi-definite

Definition : $A = A^T$ symetric

& $\vec{x}^T A \vec{x} \geq 0 \quad \forall \vec{x} \in \mathbb{R}^d$

$\longleftrightarrow$

Other generalization

Minkowski - distance / L_p-norm of difference

$$d(\vec{x}, \vec{z}) = \left[\sum_{i=1}^{d} |x_i - z_i|^p \right]^{1/p}$$

$p = 2 =$ Euc. dist

$p = 1 \equiv$ Manhattan distance

$$= \sum_{i=1}^{d} |x_i - z_i|$$

$p \to \infty \equiv$ Max - distance

$$= \max_i |x_i - z_i| \qquad 1 \leq i \leq d$$

Why? Simple proof.

$$\lim_{p \to \infty} \left[\sum_{i=1}^{d} |x_i - d_i|^p \right]^{1/p}$$

Let j = index of max-difference $= \underset{i=1,\dots,d}{\operatorname{argmax}} |x_i - z_i|$

[For simplicity, assume unique argmax]

$$= \lim_{p \to \infty} \left[|x_j - z_j|^p + \sum_{i \neq j} |x_i - z_i|^p \right]^{1/p}$$

$$= \lim_{p \to \infty} |x_j - z_j|^{p/p} \left[1 + \underbrace{\sum_{i \neq j} \left(\frac{|x_i - z_i|}{|x_j - z_j|} \right)^p}_{(<1)^p \to 0 \text{ as } p \to \infty} \right]^{1/p}$$

$$\left[1 + \sum_{i \neq j} (\to 0) \right]$$

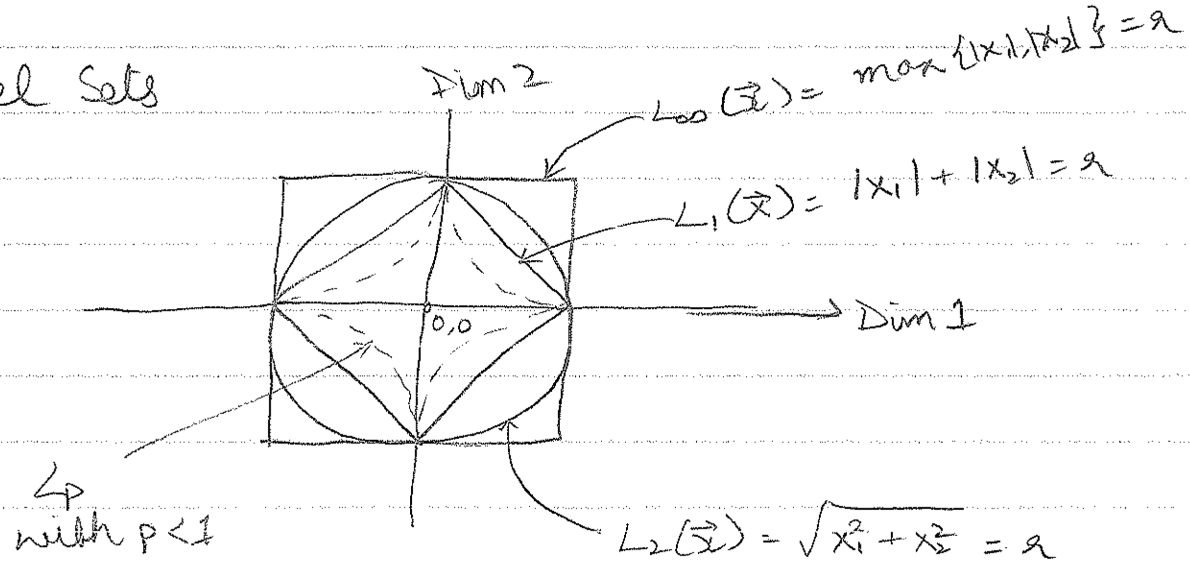
$$= |x_j - z_j|$$

∎

Similarly $p = 0$

$d(\vec{x}, \vec{z}) = \#$ dims where $x_i \neq z_i$

Level Sets

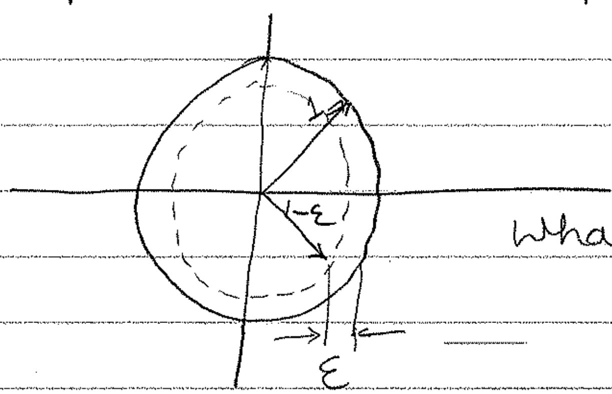


③ Curse of Dimensionality

Learning in high-dimensional space ≡ d-large is difficult.

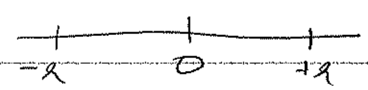
In particular, NN "shouldn't work". Why?
∴ distances/neighbours become meaningless.

→ Example #1: Consider sphere in $\mathbb{R}^d$ centred at $\vec{0}$
radius $r = 1$

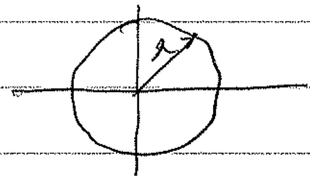


What is volume of outer ε-shell?

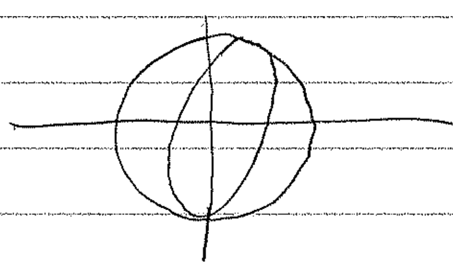
Well, what is volume of sphere?

1D 

$2r$

2D 

πr^2

3D 
$\frac{4}{3}\pi r^3$

$$\equiv k_d r^d$$

Now, $\frac{\text{Volume (Shell)}}{\text{Volume (Sphere)}} = \frac{k_d (1)^d - k_d (1-\varepsilon)^d}{k_d 1^d}$

$$= 1 - (1-\varepsilon)^d$$

$$\rightarrow 1 \quad \text{as } d \rightarrow \infty$$

$\Rightarrow$ Nearly all volume lies in outer ε-shell

Assume uniform density of data [Hint: Problem!]

$\Rightarrow$ Nearly all mass lies in shell

$\Rightarrow$ Nearly all data-points lie in shell

$\Rightarrow$ All neighbours are equally apart!

Example 2: $\vec{x} = (x_1, \ldots, x_d)$

Assume $x_1, \ldots x_d$ are I.I.D random vars

[Hint: Problem]

Consider Normalized distance^2 to origin:

$$D = \frac{1}{d} \|\vec{x} - \vec{0}\|_2^2 = \frac{1}{d} \sum_{i=1}^{d} x_i^2$$

Recall, Central Limit Theorem

If $z_1, \dots, z_n$ are I.I.D RVs with

$$E[z_i] = \mu \quad \forall i$$
$$Var(z_i) = \sigma^2 \quad \forall i$$

then $$\frac{1}{n} \sum z_i \longrightarrow N\left(\mu, \frac{\sigma^2}{n}\right)$$

as $n \to \infty$

So in our case $z_i = x_i^2$

$$D = \frac{1}{d} \sum x_i^2 \longrightarrow N\left(E[x_i^2], \frac{Var(x_i^2)}{d} \right)$$

$$Var(X_i^2) \equiv \text{constant}$$
$$\Rightarrow \quad \frac{Var(X_i^2)}{d} \longrightarrow 0 \quad \text{as } d \to \infty$$

$\Rightarrow$ D is nearly a constant.

Problem!