Curse of dimensionality

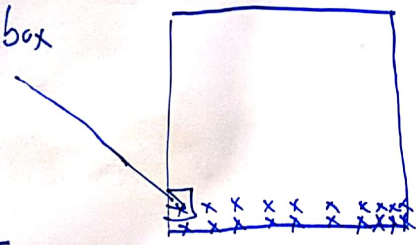
exponential growth with number of dimensions needed

l^d

how big does the box has to be

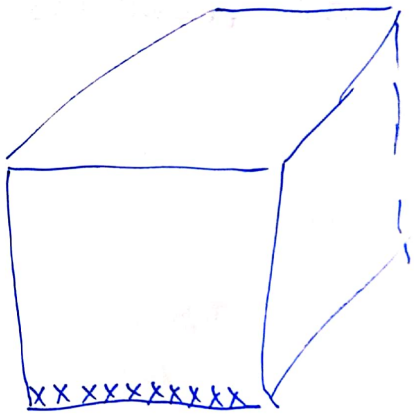
10 points 1 cm

- average distance to neighbour 0.05 cm



100 points $1\ cm^2$

- In order to cover the same average distance



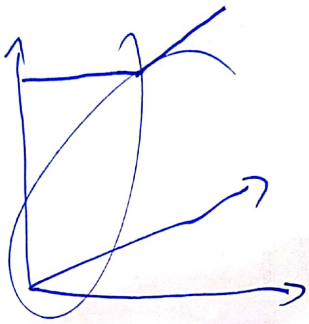
1000 $1\ cm^3$

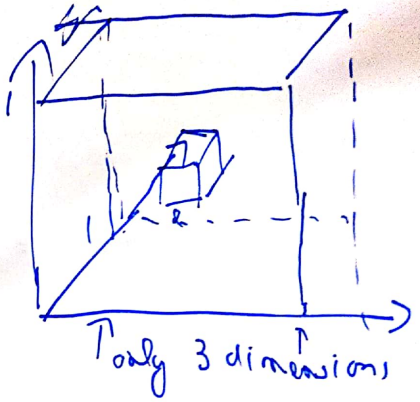
- In order to have the same average distance

→

d dimensionality

Data drawn from hypercube





$\mathbb{R}^d$

edges length 1

- all with distance 1

- Draw data uniformly at random from this hypercube
- What is the smallest little cube length ℓ that encapsulates the K nearest neighbours of this point (to find K nearest neighbours draw tiny box around it)
- How big does the little box have to be

volume of the little box $\ell^d \approx \frac{K}{n}$ $\ell \approx \left(\frac{K}{n}\right)^{\frac{1}{d}}$ ← contains K points out of n if since they are uniformly distributed

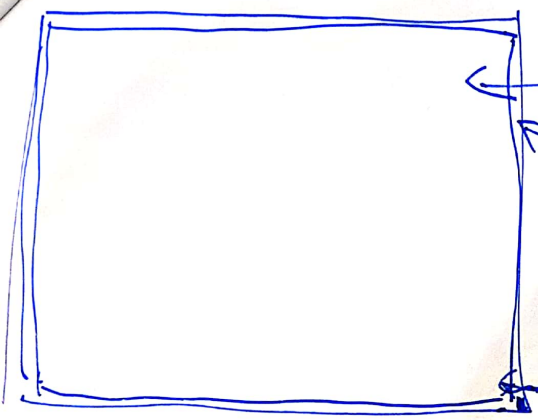
solve for this

K ≅ K nearest neighbours

n ≅ total number of points

- volume containing K out of n points
- volume same as ratio

does it mean?



- entire space empty (interior only 10 points
- remaining 990 points squeezed on the edge between the two boxes

K nearest neighbours here but many other points next to it
↳ difference in distance nothing *

- Fact that little box is so big can only mean that K nearest neighbours also on the edges
  ⇒ otherwise could draw smaller box
- Points not K nearest neighbours between the two boxes (right next to K nearest neighbours)
- K nearest neighbours not close at all - no notion of closeness, really far away on the edges
- !! All the points are really far away at the edges !
- All points roughly same distance from each other
- Counterintuitive ⇒ brain made for 3 dimensional spaces

* unreasonable to say one point should have same label as another
- not significantly further far apart

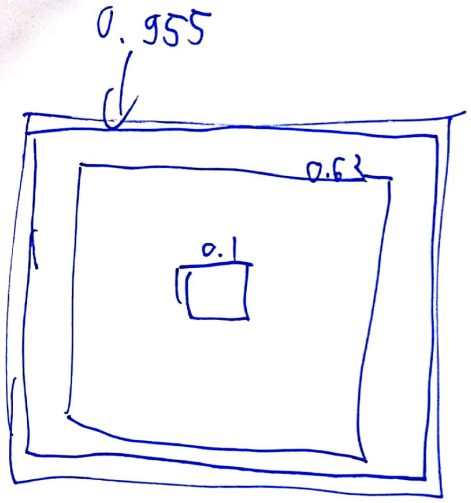
- for $\ell \approx \left(\frac{K}{n}\right)^{\frac{1}{d}}$, $K = 10$ $n = 1000$

$\approx \left(\frac{10}{1000}\right)^{\frac{1}{d}}$

$\approx \left(\frac{1}{100}\right)^{\frac{1}{d}}$

⇒ Solve for several values of d

d	ℓ
2	0.1
10	0.63
100	0.955
1000	0.9956



1000 dimensions: - have the box from which the data is sampled
- For any given point if we just look at K nearest neighbours ~~how much space~~
- How much space does this box contain, containing only my K nearest neighbours ⇒ is essentially same size as total box

dimensional spaces if you draw points randomly would be in its interior

⇒ in high dimensional spaces never the case → interior is empty, never anything there

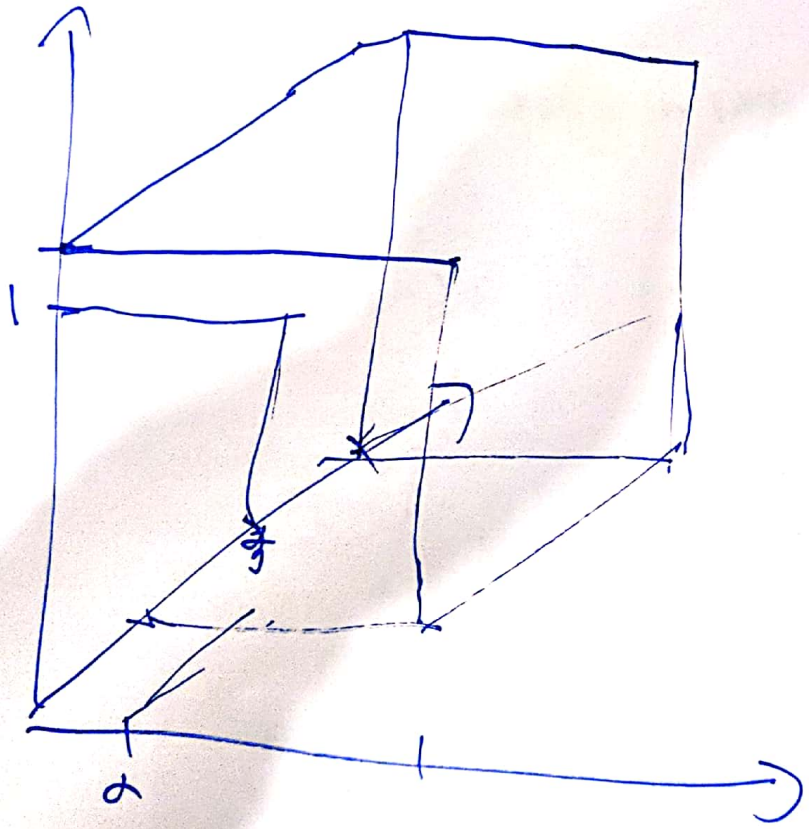
⇒ everything far away from each other

⇒ Assumption of K nearest neighbours that nearby points have the same label names

→ nothing nearby, everything the about same distance from each other

...er intuition cube everything is at the edges

- Drawing a point at random in the cube could mean drawing every single coordinate independently

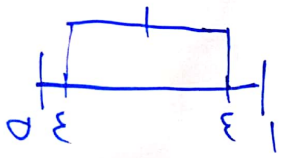


- Length of 1, some number between 0 and 1 apparently

- In order for it to be in the interior it cannot be at the edge

Probability of not being at the edge

Interior $1-2\epsilon$ ← probability of not being on the edge



$< \epsilon$ away from the edge = at the edge

~~ϵ pre~~

- ~~Intuition in~~ Probability of interior in every single dimension

$(1-2\epsilon)^d$ - in every single dimension I cannot be at the edge.

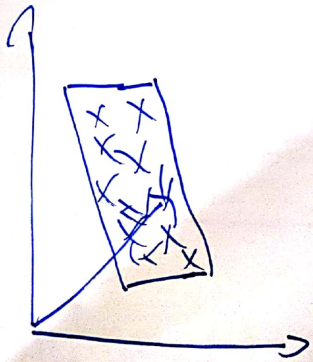
< 1 - If there is a single dimension at the edge makes you an edge point

goes to 0 quickly with large power

⇒ probability of hitting interior in high dimensional space is basically 0

How can this work e.g. on images?

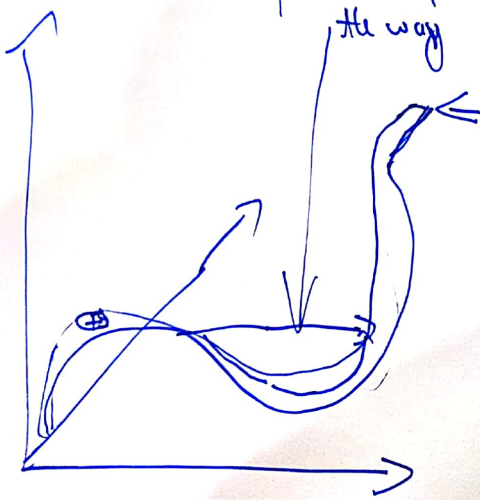
- pictures are not uniformly distributed
- no assumptions about the space high dimensional data algorithms e.g. K nearest neighbours do not work
- BUT - High dimensional data, but within this space is a subspace that is much smaller
  - Data only lies on that subspace ↑ not uniformly distributed



$\mathbb{R}^d$

- never draw data out of that subspace
- Assumption - low intrinsic dimensionality
  - ↪ Low dimensional sub space
  - ↳ Low dimensional sub-manifold

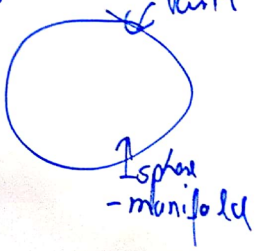
dimensional sub manifold.



points close, but if you would want to go on the manifold would have to traverse all the way

surface - eg. 10 dimensional in 1000 dimensional space
- can be curled up, so really need 1000 dimensional space to represent data
- Surface itself explores 1000 dimensions
- But data itself never leaves underlying manifold which is low dimensional

Manifold: property 1: locally euclidean, if you live in a tiny area of the subspace and move around. would have no idea space is curved - looks flat
- globally not euclidean



- K nearest neighbours only works locally