Data in 0d : a point

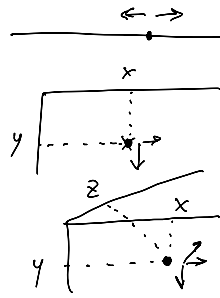
1d : a line , $x \in R$: stock, temperature

2d : a plane , $(x, y) \in R^2$: Image keypoints

3d : a cube , $(x, y, z) \in R^3$: points cloud, color

4d : ? , $(x, y, z, h) \in R^4$. RGBD camera

Nd : ? , $(x_1, x_2, \cdots x_n) \in R^n$ n-d features



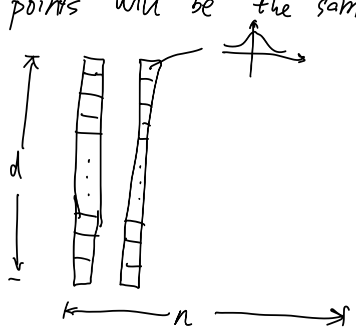
Data as vectors : $\bar{x} \in R^d$

$[x_1, x_2, x_3, \cdots, x_d]$

Properties of vectors : Geometry { length, dot product, orthogonality, cross-product }

Distances in high dimension :

If we generate n points at random in d-dimensions where each coordinate is a zero-mean unit-variance Gaussian, then for sufficiently large d [law of large numbers], with high prob the distances between all pairs of points will be the same.



two points

$$|\bar{y} - \bar{z}|^2 = \sum_{i=1}^{d} (y_i - z_i)^2$$

distance

This is equivalent to :

the sum of d independent samples of a new random variable x that is the squared difference of two Gaussians.

$\Rightarrow$ We are summing independent samples

$$x_i = (y_i - z_i)^2$$

from a random variable with bounded variance.

Law of large numbers $\Rightarrow$ average of samples from x $\longrightarrow E(x)$

Therefore, all pairs distance will approach the expectation of $\sum x$, which is the same value.

$\times$ If we have two points $\bar{y}$ and $\bar{z}$ from a d-dimensional Gaussian w/ unit variance in each dimension, then $|\bar{y}|^2 \approx d$ and $|\bar{z}|^2 \approx d$.

$$\sum^{d} (y_i)^2$$

Also,

$$E(y_i - z_i)^2 = E(y_i^2) + E(z_i)^2 - 2E(y_i z_i)$$

$$= Var(y_i) + Var(z_i) - 2E(y_i)E(z_i)$$

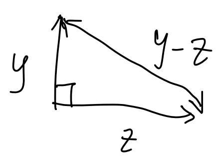
$$= 1 + 1 - 0 = 2$$

$$|\bar{y} - \bar{z}|^2 = \sum_{i=1}^{d} (y_i - z_i)^2 \approx 2d$$

d large

Then,

$$|\bar{y}|^2 \approx d, \quad |\bar{z}|^2 \approx d, \quad |\bar{y} - \bar{z}|^2 \approx 2d$$

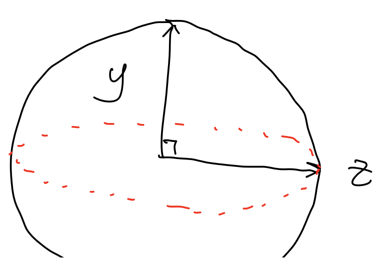


$$|\bar{y}|^2 + |\bar{z}|^2 = |\bar{y} - \bar{z}|^2$$

So $\bar{y}$ and $\bar{z}$ are orthogonal.

What does this mean?

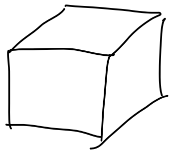
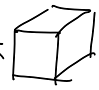
If we scale them to be unit length,



$\bar{y}$ and $\bar{z}$ are orthogonal for any pair, that means almost all the surface area of

their shap is in the equator.

* Volume:

K  $\xrightarrow[\text{by } f]{\text{shrink}}$ $f \cdot K$  , $Vol: K^3 \to (fK)^3$

$f^3 K^3$

So vol shrink by f^d.

If there is a set of points in R^d, they make an object A, we shrink it by a small amount ε so that a new object $(1-\varepsilon)A = \{(1-\varepsilon)x \mid x \in A\}$,

then $vol((1-\varepsilon)A) = (1-\varepsilon)^d vol(A)$

Since $1-\varepsilon \le e^{-\varepsilon}$, then

$$\frac{vol((1-\varepsilon)A)}{vol(A)} = (1-\varepsilon)^d \le e^{-\varepsilon d}$$

As $\underset{\text{high dim}}{d \to \infty}$, $\varepsilon^{-\varepsilon d} \to 0$, This means nearly all vol of A does not belong to $(1-\varepsilon)A$.

$\Rightarrow$ most volume of A is near its surface.