Curse of Dimensionality

↳ d is high 'dimension'

① ML Binary features

$$\begin{cases} f_1, f_2, f_3 \longrightarrow \text{Total \# datapoints} = 2^3 = 8 \\ f_1, f_2, f_3, \ldots, f_{10} \longrightarrow \quad = 2^{10} = 1024 \end{cases}$$

as dim ↑; the # datapoints to perform good model increase exponentially

D

Hughes phenomenon

"n" → size of dataset is fixed

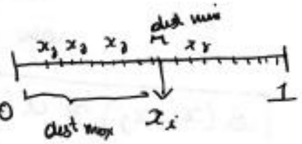
performance ↓ as dim ↑.

② Distance functions (Euclidean distance)

Curse of Dim :- intuition of distance in 3D is not valid in high dimensional spaces.

1-D world → n random pt

$$\text{distance minimum}(x_i) = \min_{x_j \neq x_i} \{ \text{dist}(x_i, x_j) \}$$



↓ distance to nearest point from x_i where $x_i \neq x_j$

$$\text{distance maximum}(x_j) = \max_{x_j \neq x_i} \left\{ \text{euc dist}(x_i, x_j) \right\}$$

(3D) $$\frac{\text{dist max}(x_i) - \text{dist\_min}(x_i)}{\text{dist\_min}(x_i)} > 0$$

when $d = 1, 2, 3$

as dim $\uparrow$ $$\left[\lim_{d \to \infty} \frac{\text{distmax}(x_i) - \text{distmin}(x_i)}{\text{distmin}(x_i)} \to 0 \right]$$

A dimensionality increases $\uparrow$ ratio approaches 0.

$\Downarrow$

$(\text{distmax}(x_i) \approx \text{distmin}(x_i))$

high dimensional space if you take

- n random pts

x_i $\text{distmax}(x_i) \approx \text{distmin}(x_i)$

$\Downarrow$

every pair of pts
are equally dist. from each other

$$\boxed{d(x_i, x_j) \approx d(x_i, x_k)}$$

3.) KNN → euc- distance

↓

high dimensional space

eucledian distance → logically not make sense.

(doesnot make sense in high dimension space.)

↳ solution : Cosine similarity → high dimensionality

↳ but less effected

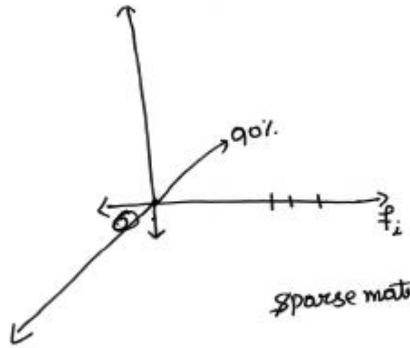
NOTE :-

Twist :- i) If data is high dimensional & dense → impact of dim. is high

ii) versus high dimensional & sparse → impact of dimensionality is lower.

↓

not uniform random /spread of data in the sparse dimension. (non zero less)



sparse matrix

③ Overfitting & underfitting

dimensionality ↑ , overfitting ↑ ⟶ linear regression.

Classification oriented → forward feature selection : pick most useful subset of features

→ Dimension - reduction :- PCA , tsne → donot use class label and not classification oriented

→ KNN on text data

↳ cosine similarity instead of Euclidean dist.

↳ sparse representation instead of dense representation

↓

bag of words