Sunday, June 2, 2024 11:39 AM

Curse of Dimensionality

1) binary features $\Rightarrow f_1^{(3-d)} f_2 f_3$ = # datapoints = 2^3

10-d $\Rightarrow$ # datapoints $\Rightarrow 2^{10}$

dimensions ↑ $\propto$ overfitting ↑ $\rightarrow$ model performance↓

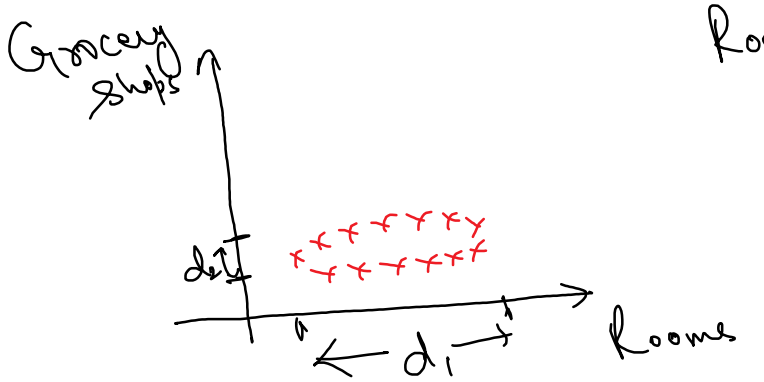
2) Distance f'ns → Euclidean distance

NLP $\Rightarrow$ Hamming | Cosine similarity

Feature Extraction

(PCA)

Feature Selection



Rooms Grocery shops Price of flat

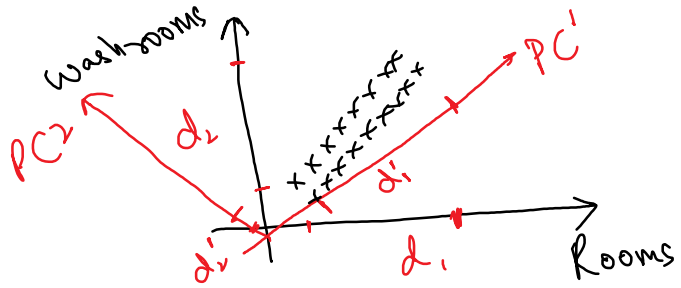
$d_1 > d_2$

Rooms will choosen

higher variance → higher info

Grocery Shops will be dropped

Feature Extraction



$d_1 \gg d_2$ $\quad d_1' > d_2'$

Rooms & washrooms → Size

Projection of $\vec{x}$ on $\hat{u}$ (unit) $= \frac{\vec{u} \cdot \vec{x}}{\|\vec{u}\| \to 1} = u^T x \to [1A]$

$$MOF = \frac{\sum_{i=1}^{n} (x - \mu)^2}{n} \Rightarrow \frac{\sum_{i=1}^{n} (u^T x_i - u^T \mu)^2}{n}$$

(mean; unit vector; np.hard)

PCA

1 → Mean Centering

2 → Co-variance Matrix $f_1 \quad f_2 \quad f_3$

$$\begin{array}{c} \\ f_1 \\ f_2 \\ f_3 \end{array} \begin{bmatrix} \text{var} f_1 & \text{cov}(f_1 f_2) & \text{cov}(f_1 f_3) \\ \text{cov}(f_1 f_2) & \text{var} f_1 & \text{cov}(f_2 f_3) \\ \text{cov}(f_1 f_3) & \text{cov}(f_2 f_3) & \text{var} f_3 \end{bmatrix}$$

3 → Eigen decomposition

Info comparison

PC1 > PC2 > PC3