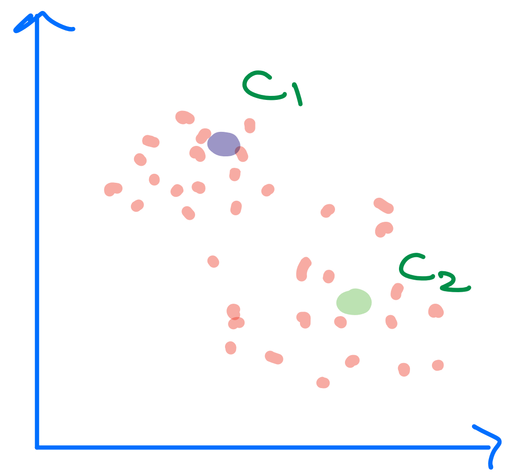
Agenda

- Matrix Factorization in Clustering
- How do we find d
- Non-Negative MF

$x_i \rightarrow 1 - n \qquad c_j \rightarrow 1 - 1$



$$\min_{c_j} \sum_{j=1}^{K} \sum_{x_i \in c_j} \| x_i - c_j \|^2$$

$$\begin{array}{c|ccccccc}
 & 1 & 2 & 3 & 4 & - & - & \cdots k \\
\hline
1 & & & & & & & \\
2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
3 & & & & & & & \\
4 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
\vdots & & & & & & & \\
n & & & & & & &
\end{array}$$

Data point belongs to cluster 4

Cluster Assignment Matrix

→ Represents the cluster x_i's belongs to

Constraint 1: $Z_{ij} \in \{0, 1\}$

Constraint 2: $\sum_{j=1}^{k} Z_{ij} = 1$

} (Hard Clustering

$$\begin{bmatrix} \updownarrow & \updownarrow & \updownarrow & & \updownarrow \\ x_1 & x_2 & x_3 & \cdots & x_n \\ \updownarrow & \updownarrow & \updownarrow & & \updownarrow \end{bmatrix}_{d \times n} \qquad \begin{bmatrix} \updownarrow & \updownarrow & \updownarrow & & \updownarrow \\ c_1 & c_2 & c_3 & \cdots & c_k \\ \updownarrow & \updownarrow & \updownarrow & & \updownarrow \end{bmatrix}_{d \times k}$$

X^T (data Matrix) $\qquad$ C (cluster Co-ordinates)

$$\min_{c_j} \sum_{j=1}^{k} \sum_{x_i \in c_j} \| x_i - c_j \|^2$$

$\downarrow$

matrix Form

$$\min_{c_j, z_{ij}} \sum_{i=1}^{n} z_{ij} \| X_i - c_j \|^2$$

$\downarrow$

$$\boxed{\min_{C, Z} \| \underset{(n \times d)}{X} - \underbrace{\underset{(n \times k)}{Z} \underset{(k \times d)}{C^T}}_{n \times d} \|^2}$$

Forbenious Norm

$$\|\vec{x}\| \Rightarrow \sqrt{x_1^2 + x_2^2 --- x_d^2}$$ Square of L2 Norm

$$\|\vec{x}\|^2 \Rightarrow x_1^2 + x_2^2 --- x_d^2$$

$$\|\vec{x}\|^2 \Rightarrow \sum_{i=1}^{d} x_i^2$$

We can extend this to Matrix as well

$$\|X\|^2 = \sum_{rows} \sum_{cols} (x_{row,col})^2$$

Forbenious Norm

$$\boxed{\min_{C, Z} \|\underset{(n \times d)}{X} - \underbrace{\underset{(n \times k)}{Z} \underset{(k \times d)}{C^T}}_{n \times d}\|_F^2}$$

X

Z and C

S.t. ① $Z_{ij} \in \{0, 1\}$

② $\sum_{j=1}^{k} Z_{ij} = 1$

RecSys

Optimize
$$\boxed{\min_{U_i, I_j} \sum_{i,j} \left(A_{ij} - U_i I_j\right)^2 \quad \text{s.t. } A_{ij} \neq \text{Null}}$$

So in Clustering, we can reformul-ate the objective as decomposition of matrix X into Z and C where Z has two constraints

Clustering can be solved with Constrained Matrix Factorization

Soft-Clustering

$$\boxed{\min_{C, Z} \Big\| \underset{(n\times d)}{X} - \underbrace{\underset{(n\times k)}{Z}\,\underset{(k\times d)}{C^T}}_{n\times d} \Big\|_F^2}$$

X Z and C

s.t. ① $Z_{ij} \in \{0,1\}$ ~~✗~~ $\rightarrow Z_{ij} \in [0,1]$

② $\sum_{j=1}^{k} Z_{ij} = 1$

$$
Z \quad
\begin{array}{c|cccccc|l}
 & 1 & 2 & 3 & 4 & \cdots & k & \\
\hline
1 & & & & & & & \\
2 & & & & & & & \\
3 & & & & & & & \\
4 & 0.1 & 0.2 & 0 & \cdots & \cdots & 0.3 & \Sigma = 1 \\
\vdots & & & & & & & \\
n & & & & & & & \\
\end{array}
$$

Takeaway

$\boxed{\text{MF}}$
- → RecSy
- → PCA(EVD), SVD
- → K-means (Hard-Clu)
- → GMM (Soft-Clu)

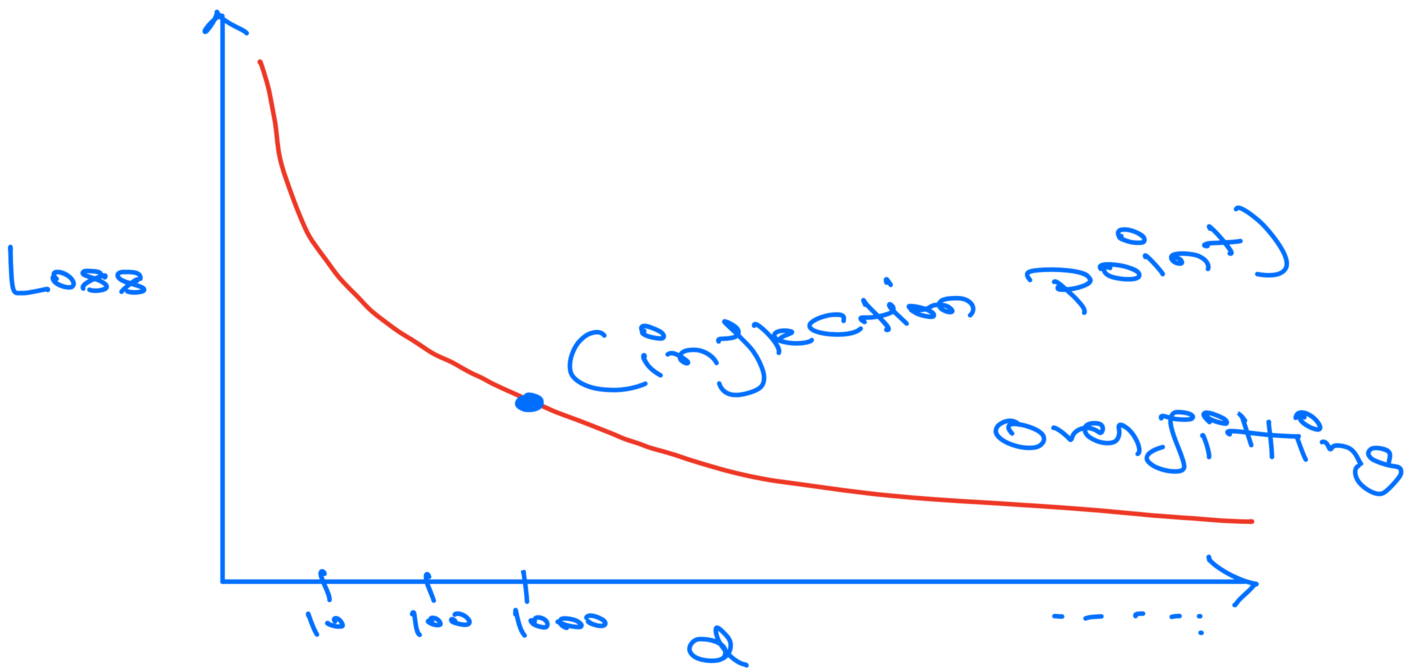
$\boxed{\text{How do we find 'd'?}}$

RecSys

Optimize
$$
\boxed{\begin{gathered}\min_{U_i,\, I_j} \sum_{i,j} \left(A_{ij} - U_i I_j \right)^2 \\ \text{s.t. } A_{ij} \neq \text{Null}\end{gathered}}
$$

$$\min_{U, I} \left\| \underbrace{A}_{n\times m} - \underbrace{U}_{n\times d} \cdot \underbrace{I^T}_{d\times m} \right\|_F^2 \Rightarrow \text{Loss}$$

Q?: How can we find best possible value of d



* The above Method can risk Overfitting

Option 2:

loss $(A_{ij} \neq \text{Not Null})$

80% (random Sample)

Training the model

20% (random Sample)

(Unseen) + (labels)

Tuning parameter

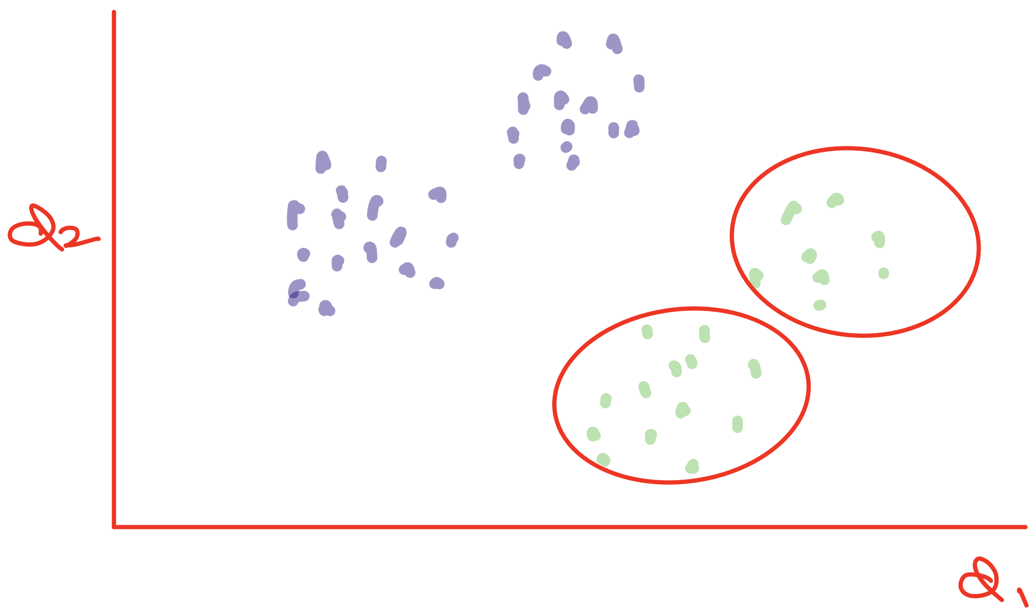
Can we interpret d?

=> No, there is no meaning or specific interpretation

=> It's hidden dimension i.e. no of dimension used for representing any item or user

=> Very high value of d means you learn every pattern and can risk overfitting / slow compute

$d=2$: Can we interpret



$$\boxed{\text{Non Negative MF (NMF)}}$$

$$A_{n\times m} = B_{n\times d} \times C_{d\times m}$$

s.t.

$$B_{ij} \geq 0 \quad \forall\, i,j$$

$$C_{ij} \geq 0 \quad \forall\, i,j$$

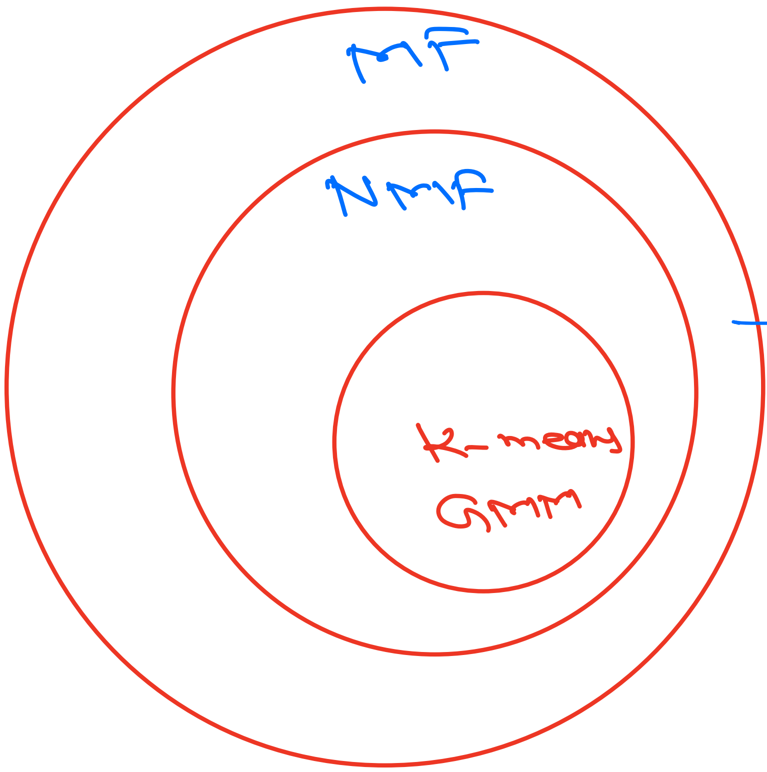
Clustering:

$$X_{n\times d} = Z^T_{n\times k} \cdot C_{k\times d} \rightarrow \text{Non Negative?}$$

$Z \rightarrow$ Non-Negative?

×

$$Z < \begin{array}{l}\{0,1\} \\ [0,1]\end{array} \Big\} \text{ Non-Negative}$$



→ MF ⊆ NMF ⊆ Clustering

Images

(0-255)

↓

normalize

(0-1)

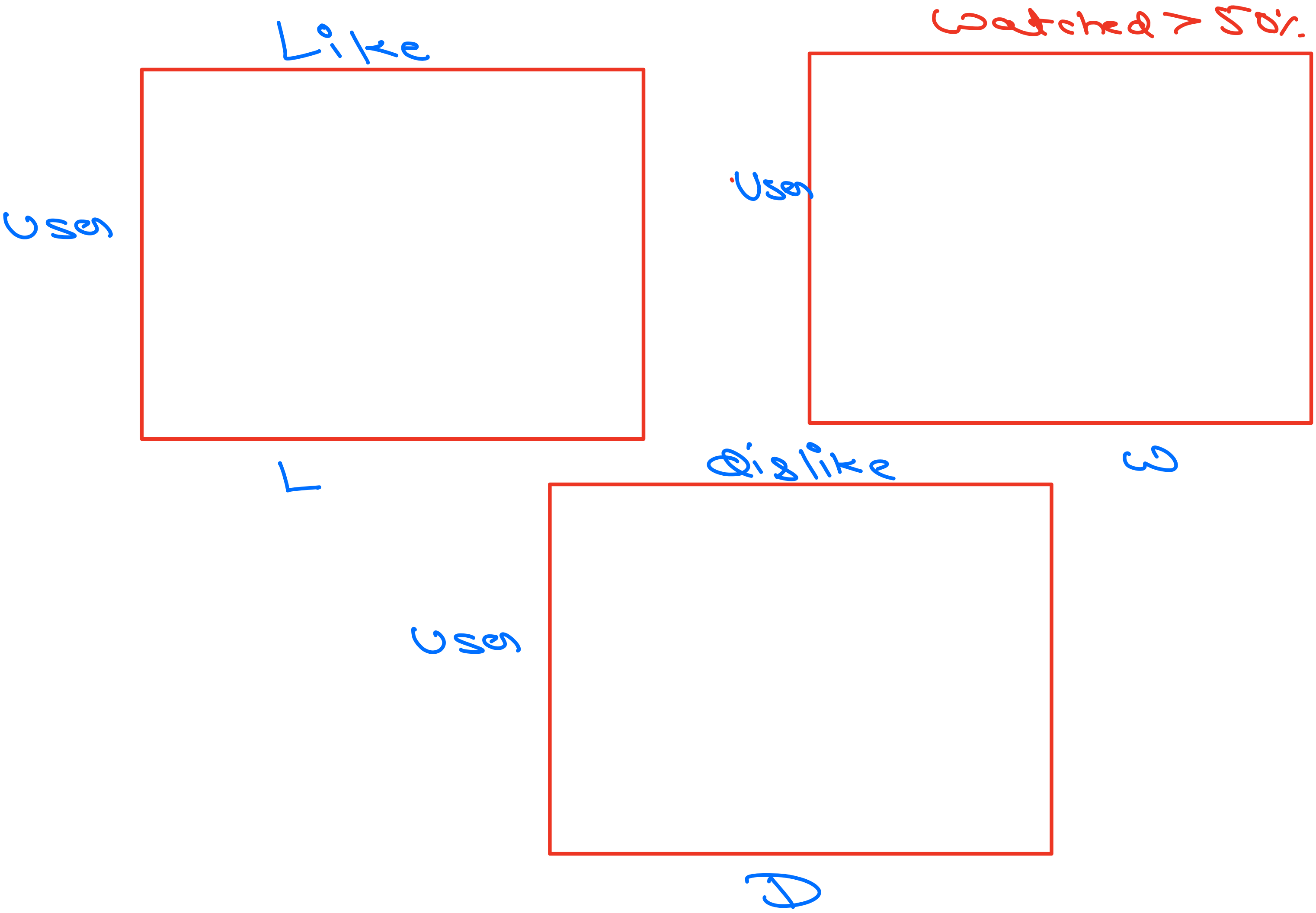
BW: $L \times B$

Color: $L \times B \times C$

$64 \times 64 \Rightarrow 256$ pixels

$\xrightarrow{\text{NMF}}$ Extract features

Eigen Faces



Option1 :

$$A_{n\times m} = \alpha_1 L + \alpha_2 W + \alpha_3 D$$

$\alpha_1 \rightarrow +1$

$\alpha_2 \rightarrow +0.5$

$\alpha_3 \rightarrow \left.\begin{matrix}-1\\-2\end{matrix}\right\}$ Negative

Option2: Independent Decomposition

$$\left.\begin{matrix} L \rightarrow G_i^L,\ H_i^L \\ W \rightarrow G_i^W,\ H_i^W \\ D \rightarrow G_i^d,\ H_i^d \end{matrix}\right\} \text{Cosine Sim}$$

Review : Final Winner Solution