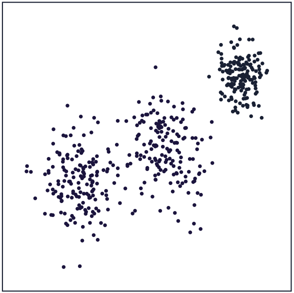
MF TUTORIAL PART 3A

Clustering -- k-means

Sibylle Hess and Michiel Hochstenbach

THE K-MEANS CLUSTERING MODEL



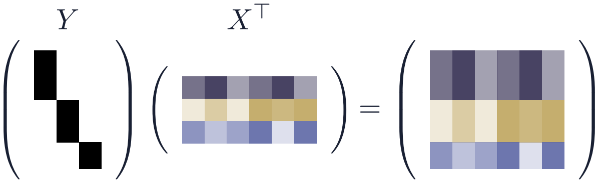
The general intuition which defines suitable clusterings is:

- points **WITHIN** one cluster are **SIMILAR**,
- points from **DISTINCT** clusters are **DISSIMILAR**.

Minimizing the within cluster scatter

$$\min \sum_{c=1}^{k} \frac{1}{|\mathcal{C}_c|} \sum_{j,l \in \mathcal{C}_c} \|D_{j\cdot} - D_{l\cdot}\|^2 \quad \text{s.t. } \{\mathcal{C}_1, \ldots, \mathcal{C}_k\} \in \mathcal{P}(1, \ldots, m)$$

K-MEANS IS MATRIX FACTORIZATION



$$\min_{Y,X} \; \|D - Y X^\top\|^2 \qquad \text{s.t. } X \in \mathbb{R}^{n \times k}, Y \in \mathbb{1}^{m \times k}$$

The set $\mathbb{1}^{m \times k}$ contains all **BINARY MATRICES** which **INDICATE A PARTITION** of m points into k sets:

$$\mathbb{1}^{m \times k} = \{ Y \in \{0,1\}^{m \times k} \, | \, |Y_{j\cdot}| = 1 \text{ for } j \in \{1, \ldots, m\} \}$$

The many objectives of k-means

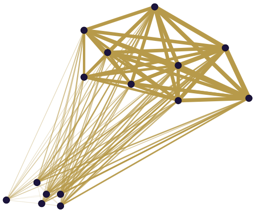
The following objectives are equivalent to (KM):

$$\min_{Y,X} \|D - YX^\top\|^2 \quad \text{s.t. } X \in \mathbb{R}^{n\times k}, Y \in \mathbb{1}^{m\times k}$$

$$\max_{Y} \operatorname{tr}(Z^\top DD^\top Z) \quad \text{s.t. } Z = Y(Y^\top Y)^{-1/2}, Y \in \mathbb{1}^{m\times k}$$

$$\max_{Y} \sum_{c=1}^{r} \frac{Y_{\cdot c}^\top DD^\top Y_{\cdot c}}{|Y_{\cdot c}|} \quad \text{s.t. } Y \in \mathbb{1}^{m\times k}$$

The matrix $W = DD^\top$ is a similarity matrix: $sim(j,l) = D_{j\cdot}D_{l\cdot}^\top$



K-MEANS AS GRAPH CLUSTERING

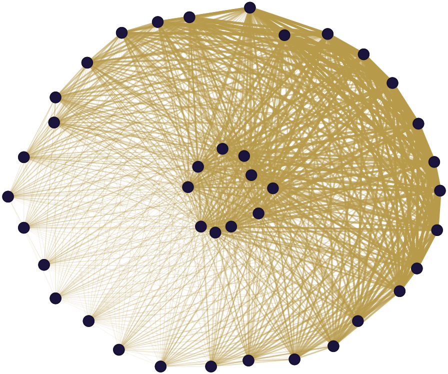
Interpreting the matrix $W = DD^\top$ as similarity matrix

$$W_{j,l} = sim(j,l) = D_{j\cdot}D_{l\cdot}^\top = \cos(\sphericalangle(D_{j\cdot}, D_{l\cdot}))\|D_{j\cdot}\|\|D_{l\cdot}\|$$

the **TRACE OBJECTIVE** maximizes the **AVERAGE SIMILARITIES** of points within one cluster:

$$\mathrm{tr}(Z^\top DD^\top Z) = \sum_{c=1}^{k} \frac{Y_{\cdot c}^\top DD^\top Y_{\cdot c}}{|Y_{\cdot c}|} = \sum_{c=1}^{k} \frac{1}{|\mathcal{C}_c|} \sum_{j,l \in \mathcal{C}_c} D_{j\cdot}D_{l\cdot}^\top$$

DRAWBACK: INNER PRODUCT SIMILARITIES RESULT IN CONVEX CLUSTERS



Nodes are positioned at their coordinates, the strength of lines indicates the similarity $sim(x, y) = \langle x, y \rangle$

Optimization

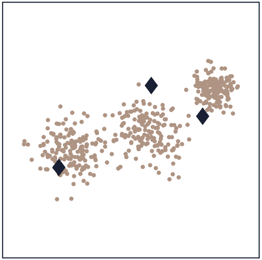
The k-**MEANS** objective introduces **BINARY CONSTRAINTS** to matrix factorization.

Binary constraints make every **FEASIBLE** binary matrix into a **LOCAL MINIMUM**.

Hess et al. 2021

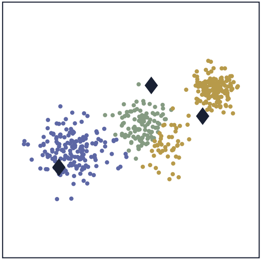
The well known k-means algorithm offers an elegant solution to the optimization problem: **ALTERNATING MINIMIZATION**.

K-MEANS: INITIALIZATION



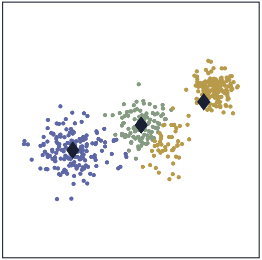
K-MEANS: CLUSTER ASSIGNMENT

$$\min_{Y} \|D - YX^\top\|^2 \text{ s.t. } Y \in \mathbb{1}^{m \times k}$$



K-MEANS: CENTROIDS UPDATE

$$\min_{X} \; \|D - YX^\top\|^2 \text{ s.t. } X \in \mathbb{R}^{n \times k}$$



Alternating Minimization for k-Means

$$X_{t+1} = \arg\min_X \|D - Y_t X^\top\|^2 \qquad \text{s.t. } X \in \mathbb{R}^{n\times k}$$

$$Y_{t+1} = \arg\min_Y \|D - Y X_{t+1}^\top\|^2 \qquad \text{s.t. } Y \in \mathbb{1}^{m\times k}$$

The **exclusivity assumption** makes the analytical computation of optimal X_{t+1} and Y_{t+1} possible.

Pros:

- fast convergence
- no hyperparameters except for rank

Cons:

- sensitive to initialization
- only applicable to partitioning clusters

Lloyd 1982

Conclusions

DISCUSSION

- k-means is like a prototype of a data mining method
- The introduction of binary constraints make the MF result interpretable as clustering (in contrast to the fuzzy coefficients of NMF)
- k-means is connected to DNN classification **Hess et al. 2020**
- k-means has an interpretation as a special case of a Gaussian mixture model

Some references part IIIa

Bauckhage 2015	Proof that k-means is matrix factorization
Pompili et al. 2014	Comparison of orthogonal NMF to k-means
Telgarsky & Vattani 2010	Discussion of Hartigans coordinate descent for k-means