Proofs, Exercises and Literature - k-means

1 Proofs

1.1 Objectives of k-means

The objective of k-means is to minimize the *within cluster scatter*. Given a data matrix $D \in \mathbb{R}^{n\times d}$ and the number of clusters r, the task is to find clusters $\{\mathcal{C}_1,\ldots,\mathcal{C}_r\} \in \mathcal{P}_n$ which create a partition of $\{1,\ldots,n\}$, minimizing the distance between points within clusters:

$$\min_{\{\mathcal{C}_1,\ldots,\mathcal{C}_r\}\in\mathcal{P}_n} Dist(\mathcal{C}_1,\ldots,\mathcal{C}_r) = \sum_{s=1}^{r} \frac{1}{|\mathcal{C}_s|} \sum_{j,i\in\mathcal{C}_s} \|D_{j\cdot} - D_{i\cdot}\|^2 \tag{1}$$

Theorem 1 (k-means centroid objective)**.** *The k-means objective in Eq.* (1) *is equivalent to the following objective:*

$$\min \sum_{s=1}^{r} \sum_{i\in\mathcal{C}_s} \|D_{i\cdot} - X_{\cdot s}^\top\|^2 \qquad s.t.\ \ X_{\cdot s} = \frac{1}{|\mathcal{C}_s|}\sum_{i\in\mathcal{C}_s} D_{i\cdot}^\top, \{\mathcal{C}_1,\ldots,\mathcal{C}_r\} \in \mathcal{P}_n \tag{2}$$

Proof. The objective function in Eq. (1) returning the average distance of points within one cluster can be transformed as follows:

$$\begin{aligned}
Dist(\mathcal{C}_1,\ldots,\mathcal{C}_r) &= \sum_{s=1}^{r} \frac{1}{|\mathcal{C}_s|} \sum_{j\in\mathcal{C}_s} \sum_{i\in\mathcal{C}_s} \|D_{i\cdot} - D_{j\cdot}\|^2 \\
&= \sum_{s=1}^{r} \frac{1}{|\mathcal{C}_s|} \sum_{j\in\mathcal{C}_s} \sum_{i\in\mathcal{C}_s} \left(\|D_{i\cdot}\|^2 - 2D_{i\cdot}D_{j\cdot}^\top + \|D_{j\cdot}\|^2\right) \quad \text{(binomial formula)} \\
&= \sum_{s=1}^{r} \left(\frac{1}{|\mathcal{C}_s|} \sum_{j\in\mathcal{C}_s} \sum_{i\in\mathcal{C}_s} \|D_{i\cdot}\|^2 - \frac{1}{|\mathcal{C}_s|} \sum_{j\in\mathcal{C}_s} \sum_{i\in\mathcal{C}_s} 2D_{i\cdot}D_{j\cdot}^\top + \frac{1}{|\mathcal{C}_s|} \sum_{j\in\mathcal{C}_s} \sum_{i\in\mathcal{C}_s} \|D_{j\cdot}\|^2 \right) \\
&= \sum_{s=1}^{r} \left(\sum_{i\in\mathcal{C}_s} \|D_{i\cdot}\|^2 - 2\sum_{i\in\mathcal{C}_s} D_{i\cdot} \frac{1}{|\mathcal{C}_s|} \sum_{j\in\mathcal{C}_s} D_{j\cdot}^\top + \sum_{j\in\mathcal{C}_s} \|D_{j\cdot}\|^2 \right) \\
&= \sum_{s=1}^{r} \left(2\sum_{i\in\mathcal{C}_s} \|D_{i\cdot}\|^2 - 2\sum_{i\in\mathcal{C}_s} D_{i\cdot} \frac{1}{|\mathcal{C}_s|} \sum_{j\in\mathcal{C}_s} D_{j\cdot}^\top \right)
\end{aligned}$$

This transformation introduces the centroid to the objective, it is given by the term on the right:

$$Dist(\mathcal{C}_1,\ldots,\mathcal{C}_r) = 2\sum_{s=1}^{r}\left(\sum_{i\in\mathcal{C}_s}\|D_{i\cdot}\|^2 - \sum_{i\in\mathcal{C}_s} D_{i\cdot}\underbrace{\frac{1}{|\mathcal{C}_s|}\sum_{j\in\mathcal{C}_s} D_{j\cdot}^\top}_{X_{\cdot s}}\right)$$

$X_{\cdot s}$ is the centroid (the arithmetic mean position) of all points assigned to cluster $\mathcal{C}_s$. We rearrange the terms now, such that we can again apply the binomial formula for norms, where the norm is used to measure the distance of a point in a cluster to the corresponding centroid:

$$\begin{aligned}
Dist(\mathcal{C}_1,\ldots,\mathcal{C}_r) &= 2\sum_{s=1}^{r}\left(\sum_{i\in\mathcal{C}_s}\|D_{i\cdot}\|^2 - \underbrace{\underbrace{\sum_{i\in\mathcal{C}_s} D_{i\cdot}}_{|\mathcal{C}_s|X_{\cdot s}^\top}\underbrace{\frac{1}{|\mathcal{C}_s|}\sum_{j\in\mathcal{C}_s} D_{j\cdot}^\top}_{X_{\cdot s}}}_{\sum_{i\in\mathcal{C}_s}\|X_{\cdot s}\|^2}\right)\\
&= 2\sum_{s=1}^{r}\sum_{i\in\mathcal{C}_s}\left(\|D_{i\cdot}\|^2 - 2D_{i\cdot}X_{\cdot s} + \|X_{\cdot s}\|^2\right)\\
&= 2\sum_{s=1}^{r}\sum_{i\in\mathcal{C}_s}\|D_{i\cdot} - X_{\cdot s}^\top\|^2 && \text{(binomial formula)}
\end{aligned}$$

The step from the first to the second equation follows by adding and subtracting the term $\sum_{i\in\mathcal{C}_s}\|X_{\cdot s}\|^2 = \sum_{i\in\mathcal{C}_s} D_{i\cdot}X_{\cdot s}$. □

Theorem 2 (k-means MF objective)**.** *The k- means objective in Eq.* (1) *is equivalent to*

$$\min_{Y} RSS(X,Y) = \|D - YX^\top\|^2 \qquad s.t.\ Y \in \mathbb{1}^{n\times r}, X = D^\top Y(Y^\top Y)^{-1}$$

Proof. The matrix Y is a cluster-indicator matrix, indicating a partition of the n data points into r sets. For every data point with index $i \in \{1,\ldots,n\}$, there exists one cluster index s_i, such that $Y_{is_i} = 1$ and $Y_{is} = 0$ for $s \neq s_i$ (point i is assigned to cluster s_i). Using this notation, the objective function in Eq. (2), returning the distance of every point to its cluster centroid, is equal to

$$\begin{aligned}
\sum_{s=1}^{r}\sum_{i\in\mathcal{C}_s}\|D_{i\cdot} - X_{\cdot s}^\top\|^2 &= \sum_{i=1}^{n}\sum_{s=1}^{r} Y_{is}\|D_{i\cdot} - X_{\cdot s}^\top\|^2 && \text{(use indicator matrix)}\\
&= \sum_{i=1}^{n}\|D_{i\cdot} - X_{\cdot s_i}^\top\|^2 && \text{(only } Y_{is_i} = 1)\\
&= \sum_{i=1}^{n}\left\|D_{i\cdot} - \sum_{s=1}^{r} Y_{is}X_{\cdot s}^\top\right\|^2 && \text{(only } Y_{is_i} = 1)\\
&= \left\|D - \sum_{s=1}^{r} Y_{\cdot s}X_{\cdot s}^\top\right\|^2 && (3)\\
&= \|D - YX^\top\|^2 && \text{(outer product def. matrix product).}
\end{aligned}$$

Eq. (3) uses the composition of the squared Frobenius norm as a sum of the squared Euclidean vector norm over all rows. We have shown this connection in the math PDF for the regression lecture. □

2 Exercises

1. Compute a 2-means clustering on the movie rating matrix from the lecture step by step:

$$D = \begin{pmatrix} 5 & 3 & 1 & 1 \\ 3 & 1 & 5 & 3 \\ 2 & 1 & 5 & 3 \\ 4 & 3 & 4 & 2 \\ 5 & 5 & 3 & 1 \\ 3 & 1 & 5 & 3 \end{pmatrix}.$$

   You can use as initialization the centroids

$$X_0^\top = \begin{pmatrix} 5 & 3 & 1 & 1 \\ 3 & 1 & 5 & 3 \end{pmatrix}.$$

2. Show that for $Y \in \mathbb{1}^{n\times r}$ we have

$$Y^\top Y = \begin{pmatrix} |Y_{\cdot 1}| & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & |Y_{\cdot r}| \end{pmatrix}.$$

3 Recommended Literature

As always, the best exercise is to go through the lecture and see if you can follow the steps with pen and paper. If you feel like reading though, you can have a look at the following material:

Friedman, Hastie, and Tibshirani. The elements of statistical learning. 2001.

Here, the procedure of k-means is introduced for general (not necessarily Euclidean) distances. Note, that the objective of k-means is here a bit different, the distances of points within one cluster are not averaged!

14.3 Cluster Analysis

- **14.3.1** Proximity Matrices (This sets the stage for next lecture already)
- **14.3.2** Dissimilarities Based on Attributes
- **14.3.3** Object Dissimilarity
- **14.3.5** Combinatorial Algorithms
- **14.3.6** k-means

Linear Algebra and Optimization for Machine Learning by Charu C. Aggarwal

4.10.2 Block Coordinate Descent

4.10.3 K-Means as Block Coordinate Descent