Lecture Notes on Data Science: k-Means Clustering Is Matrix Factorization

Christian Bauckhage
B-IT, University of Bonn

In this note, we show that k-means clustering can be understood as a constrained matrix factorization problem. This insight will later allow us to recognize that k-means clustering is but a specific latent factor model and closely related to techniques such as non-negative matrix factorization or archetypal analysis.

Introduction

Previously, we discussed[1] that hard k-means clustering of a data set $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^m$ into k clusters $C_1, \ldots, C_k$ boils down to the problem of finding appropriate cluster centroids $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ and that these will minimize the following objective function

$$E(k) = \sum_{i=1}^{k} \sum_{j=1}^{n} z_{ij} \left\| \boldsymbol{x}_j - \boldsymbol{\mu}_i \right\|^2 \tag{1}$$

where

$$z_{ij} = \begin{cases} 1, & \text{if } \boldsymbol{x}_j \in C_i \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

[1] C. Bauckhage. Lecture Notes on Data Science: k-Means Clustering, 2015b. DOI: 10.13140/RG.2.1.2829.4886; and C. Bauckhage. Lecture Notes on Data Science: k-Means Clustering Is Gaussian Mixture Modeling, 2015a. DOI: 10.13140/RG.2.1.3033.2646

Our purpose in this note is to show that there is yet another way of how to formalize the k-means objective in (1).

To this end, we note that we may understand the binary indicator variables z_{ij} in (2) as the elements of an ***indicator matrix*** $\boldsymbol{Z} \in \mathbb{R}^{k \times n}$.

We also observe that we may think of the given data points $\boldsymbol{x}_j$ as the columns of a ***data matrix***

data matrix

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_n \end{bmatrix} \in \mathbb{R}^{m \times n} \tag{3}$$

and that we may furthermore introduce a ***centroid matrix***

centroid matrix

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{\mu}_1 & \boldsymbol{\mu}_2 & \ldots & \boldsymbol{\mu}_k \end{bmatrix} \in \mathbb{R}^{m \times k} \tag{4}$$

whose columns correspond to the cluster centroids that are to be determined.

Given the matrices defined in (2), (3), and (4), we will show that the k-means objective function in (1) can indeed be written as

alternative form of the k-means objective function

$$\sum_{i=1}^{k} \sum_{j=1}^{n} z_{ij} \left\| \boldsymbol{x}_j - \boldsymbol{\mu}_i \right\|^2 = \left\| \boldsymbol{X} - \boldsymbol{M}\boldsymbol{Z} \right\|_F^2 \tag{5}$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm.

In other words, we will show that k-means clustering is a matrix factorization problem! If there were two appropriate matrices $\boldsymbol{M}$ and $\boldsymbol{Z}$ that would minimize the right hand side of (5), the data matrix $\boldsymbol{X}$ could be approximated as $\boldsymbol{X} \approx \boldsymbol{M}\boldsymbol{Z}$.

Exercise: convince yourself that $\boldsymbol{M}\boldsymbol{Z}$ is a $m \times n$ matrix.

Proving Equation (5)

In this section, we will prove that our claim in (5) does indeed hold. The basic idea is to expand both sides of the equation into several, more elementary terms and to show that the expressions we obtain for the left- and right hand side are indeed equivalent.

Yet, before we set out to do so, we will remind ourselves of general properties of the Frobenius norm and point out some of the peculiar features of the binary indicator matrix $\boldsymbol{Z}$.

General Properties of the Squared Frobenius Norm of a Matrix

Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ be any real valued matrix of m rows and n columns. To denote individual elements of such a matrix, we either write a_{ij} or $(\boldsymbol{A})_{ij}$ and to refer to the j-th column vector of $\boldsymbol{A}$, we write $\boldsymbol{a}_j$.

The ***squared Frobenius norm*** of $\boldsymbol{A}$ is defined as

$$\|\boldsymbol{A}\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \tag{6}$$

and we recall the following properties

$$\|\boldsymbol{A}\|_F^2 = \sum_{j=1}^{n} \|\boldsymbol{a}_j\|^2 = \sum_{j=1}^{n} \boldsymbol{a}_j^T \boldsymbol{a}_j = \sum_{j=1}^{n} (\boldsymbol{A}^T \boldsymbol{A})_{jj} = \operatorname{tr}[\boldsymbol{A}^T \boldsymbol{A}]. \tag{7}$$

Since our derivation below will frequently allude to the identities in (7), readers are encouraged to verify (7) for themselves.

Exercise: convince yourself that all the equalities in (7) do hold.

Peculiar Properties of the Indicator Matrix $\boldsymbol{Z}$

If the clusters $C_1, \dots C_k$ have distinct cluster centroids $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k$, each of the n columns of $\boldsymbol{Z}$ will contain a single element that is 1 and $k-1$ elements that are 0. Accordingly, each column j of $\boldsymbol{Z}$ will sum to one

$$\sum_{i=1}^{k} z_{ij} = 1 \tag{8}$$

and the k different row sums will indicate the number of elements per cluster, that is, for each row i of $\boldsymbol{Z}$, we have

$$\sum_{j=1}^{n} z_{ij} = |C_i| = n_i. \tag{9}$$

Moreover, since $z_{ij} \in \{0, 1\}$ and each column of $\boldsymbol{Z}$ only contains a single 1, the rows of $\boldsymbol{Z}$ are pairwise perpendicular because

$$z_{ij}\, z_{i'j} = \begin{cases} 1, & \text{if } i = i' \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

which is then to say that the matrix $\boldsymbol{Z}\boldsymbol{Z}^T$ is a diagonal matrix where

$$(\boldsymbol{Z}\boldsymbol{Z}^T)_{ii'} = \sum_{j} (\boldsymbol{Z})_{ij} (\boldsymbol{Z}^T)_{ji'} = \sum_{j} z_{ij}\, z_{i'j} = \begin{cases} n_i, & \text{if } i = i' \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

HAVING FAMILIARIZED OURSELVES with these properties of the indicator matrix, we are now positioned to establish the equalities in (5) which we will do in a step by step manner.

Step 1: Expanding the expression on the left of (5)

We begin by expanding the traditional k-means objective on the left of (5). For this expression, we have

$$\begin{aligned}\sum_{i,j} z_{ij} \left\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\right\|^2 &= \sum_{i,j} z_{ij} \left(\boldsymbol{x}_j^T \boldsymbol{x}_j - 2\boldsymbol{x}_j^T \boldsymbol{\mu}_i + \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i\right) \\ &= \underbrace{\sum_{i,j} z_{ij}\, \boldsymbol{x}_j^T \boldsymbol{x}_j}_{T_1} \underbrace{-2 \sum_{i,j} z_{ij}\, \boldsymbol{x}_j^T \boldsymbol{\mu}_i}_{T_2} + \underbrace{\sum_{i,j} z_{ij}\, \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i}_{T_3}. \end{aligned} \tag{12}$$

This expansion leads to further insights, if we examine the three terms T_1, T_2, and T_3 one by one.

FIRST OF ALL, we find

$$\begin{aligned} T_1 = \sum_{i,j} z_{ij}\, \boldsymbol{x}_j^T \boldsymbol{x}_j &= \sum_{i,j} z_{ij} \left\|\boldsymbol{x}_j\right\|^2 && (13)\\ &= \sum_{j} \left\|\boldsymbol{x}_j\right\|^2 && (14)\\ &= \operatorname{tr}\left[\boldsymbol{X}^T \boldsymbol{X}\right] && (15)\end{aligned}$$

where we made use of (8) and (7).

SECOND OF ALL, we observe

$$\begin{aligned} T_2 = \sum_{i,j} z_{ij}\, \boldsymbol{x}_j^T \boldsymbol{\mu}_i &= \sum_{i,j} z_{ij} \sum_{l} x_{lj}\, \mu_{li} && (16)\\ &= \sum_{j,l} x_{lj} \sum_{i} \mu_{li}\, z_{ij} && (17)\\ &= \sum_{j,l} x_{lj} \left(\boldsymbol{M}\boldsymbol{Z}\right)_{lj} && (18)\\ &= \sum_{j} \sum_{l} \left(\boldsymbol{X}^T\right)_{jl} \left(\boldsymbol{M}\boldsymbol{Z}\right)_{lj} && (19)\\ &= \sum_{j} \left(\boldsymbol{X}^T \boldsymbol{M}\boldsymbol{Z}\right)_{jj} && (20)\\ &= \operatorname{tr}\left[\boldsymbol{X}^T \boldsymbol{M}\boldsymbol{Z}\right]. && (21)\end{aligned}$$

THIRD OF ALL, we note that

$$\begin{aligned} T_3 = \sum_{i,j} z_{ij}\, \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i &= \sum_{i,j} z_{ij} \left\|\boldsymbol{\mu}_i\right\|^2 && (22)\\ &= \sum_{i} \left\|\boldsymbol{\mu}_i\right\|^2 n_i && (23)\end{aligned}$$

where we applied (9).

Step 2: Expanding the expression on the right of (5)

Next, we look at the expression on the right hand side of (5). As a squared Frobenius norm of a matrix difference, it can be written as

$$\begin{aligned}\left\|\boldsymbol{X}-\boldsymbol{M}\boldsymbol{Z}\right\|^2 &= \operatorname{tr}\Big[\big(\boldsymbol{X}-\boldsymbol{M}\boldsymbol{Z}\big)^T\big(\boldsymbol{X}-\boldsymbol{M}\boldsymbol{Z}\big)\Big] \\ &= \underbrace{\operatorname{tr}\big[\boldsymbol{X}^T\boldsymbol{X}\big]}_{T_4} - 2\underbrace{\operatorname{tr}\big[\boldsymbol{X}^T\boldsymbol{M}\boldsymbol{Z}\big]}_{T_5} + \underbrace{\operatorname{tr}\big[\boldsymbol{Z}^T\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{Z}\big]}_{T_6} \end{aligned} \tag{24}$$

Given our results in (15) and (21), we immediately recognize that $T_1 = T_4$ and $T_2 = T_5$. Thus, to establish that (12) and (24) are indeed equivalent, it remains to verify whether $T_3 = T_6$?

Regarding term T_6, we note that, due to the cyclic permutation invariance of the trace operator, we have

$$\operatorname{tr}\big[\boldsymbol{Z}^T\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{Z}\big] = \operatorname{tr}\big[\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{Z}\boldsymbol{Z}^T\big]. \tag{25}$$

We also note that

$$\begin{aligned}\operatorname{tr}\big[\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{Z}\boldsymbol{Z}^T\big] &= \sum_i \big(\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{Z}\boldsymbol{Z}^T\big)_{ii} && (26)\\ &= \sum_i\sum_l \big(\boldsymbol{M}^T\boldsymbol{M}\big)_{il}\big(\boldsymbol{Z}\boldsymbol{Z}^T\big)_{li} && (27)\\ &= \sum_i \big(\boldsymbol{M}^T\boldsymbol{M}\big)_{ii}\big(\boldsymbol{Z}\boldsymbol{Z}^T\big)_{ii} && (28)\\ &= \sum_i \left\|\boldsymbol{\mu}_i\right\|^2 n_i && (29)\end{aligned}$$

where we used the fact that $\boldsymbol{Z}\boldsymbol{Z}^T$ is diagonal. This result, however, shows that $T_3 = T_6$ and, consequently, that (12) and (24) really are equivalent.

Summary and Outlook

Using rather tedious yet straightforward algebra, we have shown that the problem of hard k-means clustering can be understood as the following constrained matrix factorization problem

$$\begin{aligned}\underset{\boldsymbol{M},\boldsymbol{Z}}{\operatorname{argmin}} \quad & \left\|\boldsymbol{X}-\boldsymbol{M}\boldsymbol{Z}\right\|^2 \\ \text{s.t.} \quad & z_{ij}\in\{0,1\} \\ & \sum_i z_{ij} = 1\end{aligned} \tag{30}$$

where

$$\boldsymbol{X}\in\mathbb{R}^{m\times n} \text{ is a matrix of data vectors } \boldsymbol{x}_j\in\mathbb{R}^m \tag{31}$$

$$\boldsymbol{M}\in\mathbb{R}^{m\times k} \text{ is a matrix of cluster centroids } \boldsymbol{\mu}_i\in\mathbb{R}^m \tag{32}$$

$\boldsymbol{Z}\in\mathbb{R}^{k\times n}$ is a matrix of binary indicator variables such that

$$z_{ij} = \begin{cases} 1, \text{ if } \boldsymbol{x}_j\in C_i \\ 0, \text{ otherwise.}\end{cases} \tag{33}$$

AT THIS POINT, readers who are not accustomed with the idea of matrix factorization for data analysis might be wondering what we could possibly gain from this insight.

Admittedly, the formulation of the k-means clustering problem in (30) appears to be more complicated and less intuitive than those fond in the textbooks. However, in later notes, we will see that the expression in (30) allows for seamless insights into several important properties of the k-means clustering problem that are otherwise more difficult to uncover [2].

[2] C. Bauckhage. k-Means Clustering via the Frank-Wolfe Algorithm. In *Proc. KDML-LWDA*, 2016

References

C. Bauckhage. Lecture Notes on Data Science: k-Means Clustering Is Gaussian Mixture Modeling, 2015a. DOI: 10.13140/RG.2.1.3033.2646.

C. Bauckhage. Lecture Notes on Data Science: k-Means Clustering, 2015b. DOI: 10.13140/RG.2.1.2829.4886.

C. Bauckhage. k-Means Clustering via the Frank-Wolfe Algorithm. In *Proc. KDML-LWDA*, 2016.