k-Means Clustering Is Matrix Factorization

Yixin Zhang

University of Alberta

Abstract. We show that the objective function of conventional k-means clustering can be expressed as the Frobenius norm of the difference of a data matrix and a low rank approximation of that data matrix. In short, we show that k-means clustering is a matrix factorization problem.

1 Introduction

The k-means procedure is one of the most popular techniques to cluster a data set $X \subset \mathbb{R}^m$ into subsets $C_1, \ldots, C_k$. The underlying ideas are intuitive and simple and most theoretical properties of k-means clustering are well established in literature material [1,2].

In this review, we are concerned with an aspect of k-means clustering that is arguably less well known and somewhat under-appreciated. Our goal in this review is to rigorously establish the following equalities for the objective function of hard k-means clustering

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

where

$$\boldsymbol{X} \in \mathbb{R}^{m \times n} \text{ is a matrix of data vectors } \boldsymbol{x}_j \in \mathbb{R}^m \tag{2}$$

$$\boldsymbol{M} \in \mathbb{R}^{m \times k} \text{ is a matrix of cluster centroids } \boldsymbol{\mu}_i \in \mathbb{R}^m \tag{3}$$

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

$$z_{ij} = \begin{cases} 1, & \text{if } \boldsymbol{x}_j \in C_i \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

2 Notation and Preliminaries

Throughout, we write $\boldsymbol{x}_j$ to denote j-th column vector of a matrix $\boldsymbol{X}$. To refer to the (l, j) element of a matrix $\boldsymbol{X}$, we either write x_{lj} or $\left(\boldsymbol{X}\right)_{lj}$.

The Euclidean norm of a vector will be written as $\|\boldsymbol{x}\|$ and the Frobenius norm of a matrix as $\|\boldsymbol{X}\|$.

Regarding the squared Frobenius norm of a matrix, we recall the following properties

$$\left\|\boldsymbol{X}\right\|^2 = \sum_{l,j} x_{lj}^2 = \sum_j \left\|\boldsymbol{x}_j\right\|^2 = \sum_j \boldsymbol{x}_j^T \boldsymbol{x}_j = \sum_j \left(\boldsymbol{X}^T \boldsymbol{X}\right)_{jj} = \operatorname{tr}\left[\boldsymbol{X}^T \boldsymbol{X}\right] \quad (5)$$

Finally, subscripts or summation indices i will be understood to range from 1 to k (the number of clusters), subscripts or summation indices j will range from 1 up to n (the number of data vectors), and subscripts or summation indices l will be used to expand inner products between vectors or rows and columns of matrices.

3 Step by Step Derivation of (1)

To substantiate the claim in (1), we first point out several peculiar properties of the binary indicator matrix $\boldsymbol{Z}$ in (4).

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

$$\sum_i z_{ij} = 1 \quad (6)$$

and its row sums will indicate the number elements per cluster

$$\sum_j z_{ij} = n_i = |C_i|. \quad (7)$$

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

$$z_{ij}\, z_{i'j} = \begin{cases} 1, & \text{if } i = i' \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

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

$$\left(\boldsymbol{Z}\boldsymbol{Z}^T\right)_{ii'} = \sum_j \left(\boldsymbol{Z}\right)_{ij} \left(\boldsymbol{Z}^T\right)_{ji'} = \sum_j z_{ij}\, z_{i'j} = \begin{cases} n_i, & \text{if } i = i' \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

Having familiarized ourselves with these properties of the indicator matrix, we are now positioned to establish the equalities in (1) which we will do in a step by step manner.

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

We begin our derivation by expanding the conventional k-means objective function on the left of (1). For this expression, we have

$$\begin{aligned}\sum_{i,j} z_{ij} \left\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\right\|^2 &= \sum_{i,j} z_{ij} \left(\boldsymbol{x}_j^T \boldsymbol{x}_j - 2\boldsymbol{x}_j^T \boldsymbol{\mu}_i + \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i\right) \\ &= \underbrace{\sum_{i,j} z_{ij}\, \boldsymbol{x}_j^T \boldsymbol{x}_j}_{T_1} - 2 \underbrace{\sum_{i,j} z_{ij}\, \boldsymbol{x}_j^T \boldsymbol{\mu}_i}_{T_2} + \underbrace{\sum_{i,j} z_{ij}\, \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i}_{T_3}. \end{aligned} \tag{10}$$

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

$$\begin{aligned} T_1 = \sum_{i,j} z_{ij}\, \boldsymbol{x}_j^T \boldsymbol{x}_j &= \sum_{i,j} z_{ij} \left\|\boldsymbol{x}_j\right\|^2 && (11) \\ &= \sum_{j} \left\|\boldsymbol{x}_j\right\|^2 && (12) \\ &= \operatorname{tr}\left[\boldsymbol{X}^T \boldsymbol{X}\right] && (13) \end{aligned}$$

where we made use of (6) and (5). Second of all, we observe

$$\begin{aligned} T_2 = \sum_{i,j} z_{ij}\, \boldsymbol{x}_j^T \boldsymbol{\mu}_i &= \sum_{i,j} z_{ij} \sum_{l} x_{lj}\, \mu_{li} && (14) \\ &= \sum_{j,l} x_{lj} \sum_{i} \mu_{li}\, z_{ij} && (15) \\ &= \sum_{j,l} x_{lj} \left(\boldsymbol{M}\boldsymbol{Z}\right)_{lj} && (16) \\ &= \sum_{j} \sum_{l} \left(\boldsymbol{X}^T\right)_{jl} \left(\boldsymbol{M}\boldsymbol{Z}\right)_{lj} && (17) \\ &= \sum_{j} \left(\boldsymbol{X}^T \boldsymbol{M}\boldsymbol{Z}\right)_{jj} && (18) \\ &= \operatorname{tr}\left[\boldsymbol{X}^T \boldsymbol{M}\boldsymbol{Z}\right] && (19) \end{aligned}$$

Third of all, we note that

$$\begin{aligned} T_3 = \sum_{i,j} z_{ij}\, \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i &= \sum_{i,j} z_{ij} \left\|\boldsymbol{\mu}_i\right\|^2 && (20) \\ &= \sum_{i} \left\|\boldsymbol{\mu}_i\right\|^2 n_i && (21) \end{aligned}$$

where we applied (7).

3.2 Step 2: Expanding the expression in the middle of (1)

Next, we look at the second expression in (1). As a squared Frobenius norm of a matrix difference, it can be written as

$$\begin{aligned}\left\|\boldsymbol{X}-\boldsymbol{M}\boldsymbol{Z}\right\|^2 &= \operatorname{tr}\Big[\big(\boldsymbol{X}-\boldsymbol{M}\boldsymbol{Z}\big)^T\big(\boldsymbol{X}-\boldsymbol{M}\boldsymbol{Z}\big)\Big] \\ &= \underbrace{\operatorname{tr}\big[\boldsymbol{X}^T\boldsymbol{X}\big]}_{T_4} - 2\underbrace{\operatorname{tr}\big[\boldsymbol{X}^T\boldsymbol{M}\boldsymbol{Z}\big]}_{T_5} + \underbrace{\operatorname{tr}\big[\boldsymbol{Z}^T\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{Z}\big]}_{T_6} \end{aligned} \tag{22}$$

Given our earlier results, we immediately recognize that $T_1 = T_4$ and $T_2 = T_5$. Thus, to establish that (10) and (22) are indeed equivalent, it remains to verify whether $T_3 = T_6$?

Regarding T_6, we note that, because of the cyclic permutation invariance of the trace operator, we have

$$\operatorname{tr}\big[\boldsymbol{Z}^T\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{Z}\big] = \operatorname{tr}\big[\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{Z}\boldsymbol{Z}^T\big]. \tag{23}$$

We also note that

$$\begin{aligned}\operatorname{tr}\big[\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{Z}\boldsymbol{Z}^T\big] &= \sum_i \big(\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{Z}\boldsymbol{Z}^T\big)_{ii} && (24)\\ &= \sum_i\sum_l \big(\boldsymbol{M}^T\boldsymbol{M}\big)_{il}\big(\boldsymbol{Z}\boldsymbol{Z}^T\big)_{li} && (25)\\ &= \sum_i \big(\boldsymbol{M}^T\boldsymbol{M}\big)_{ii}\big(\boldsymbol{Z}\boldsymbol{Z}^T\big)_{ii} && (26)\\ &= \sum_i \left\|\boldsymbol{\mu}_i\right\|^2 n_i && (27)\end{aligned}$$

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 (10) and (22) really are equivalent.

3.3 Step 3: Eliminating matrix $\boldsymbol{M}$

Finally, to establish the equality on the right of (1) we ask for the matrix $\boldsymbol{M}$ that, for a given $\boldsymbol{Z}$, would minimize $\left\|\boldsymbol{X}-\boldsymbol{M}\boldsymbol{Z}\right\|^2$. To this end, we consider

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{M}}\left\|\boldsymbol{X}-\boldsymbol{M}\boldsymbol{Z}\right\|^2 &= \frac{\partial}{\partial \boldsymbol{M}}\Big[\operatorname{tr}\big[\boldsymbol{X}^T\boldsymbol{X}\big] - 2\operatorname{tr}\big[\boldsymbol{X}^T\boldsymbol{M}\boldsymbol{Z}\big] + \operatorname{tr}\big[\boldsymbol{Z}^T\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{Z}\big]\Big] \\ &= 2\big(\boldsymbol{M}\boldsymbol{Z}\boldsymbol{Z}^T - \boldsymbol{X}\boldsymbol{Z}^T\big)\end{aligned} \tag{28}$$

which, upon equation to $\mathbf{0}$, leads to

$$\boldsymbol{M} = \boldsymbol{X}\boldsymbol{Z}^T\big(\boldsymbol{Z}\boldsymbol{Z}^T\big)^{-1} \tag{29}$$

which beautifully reflects the fact that each of the k-means cluster centroids $\boldsymbol{\mu}_i$ coincides with the mean of the corresponding cluster C_i, namely

$$\boldsymbol{\mu}_i = \frac{\sum_j z_{ij}\,\boldsymbol{x}_j}{\sum_j z_{ij}} = \frac{1}{n_i}\sum_{\boldsymbol{x}_j \in C_i} \boldsymbol{x}_j. \tag{30}$$

4 Conclusion

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

$$\begin{aligned}
\min_{\boldsymbol{Z}} \quad & \left\| \boldsymbol{X} - \boldsymbol{X}\boldsymbol{Z}^T \left(\boldsymbol{Z}\boldsymbol{Z}^T\right)^{-1} \boldsymbol{Z} \right\|^2 \\
\text{s.t.} \quad & z_{ij} \in \{0, 1\} \\
& \sum_j z_{ij} = 1
\end{aligned}$$

References

1. MacKay, D.: Information Theory, Inference, & Learning Algorithms. Cambridge University Press (2003)
2. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer (2001)
3. Ding, C., He, X., Simon, H.: On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. In: Proc. SDM, SIAM (2005)
4. Gaussier, E., Goutte, C.: Relations between PLSA and NMF and Implications. In: Proc. SIGIR, ACM (2005)
5. Kim, J., Park, H.: Sparse Nonnegative Matrix Factorization for Clustering. Technical Report GT-CSE-08-01, Georgia Institute of Technology (2008)
6. Arora, R., Gupta, M., Kapila, A., Fazel, M.: Similarity-based Clustering by Left-Stochastic Matrix Factorization. J. of Machine Learning Research **14**(Jul.) (2013)
7. Bauckhage, C., Drachen, A., Sifa, R.: Clustering Game Behavior Data. IEEE Trans. on Computational Intelligence and AI in Games **7**(3) (2015)