K-Means Clustering Notation Guide

Todd W. Neller
Gettysburg College
tneller@gettysburg.edu

Introduction

It is rare in any computational field or topic for a standard notation to be established and followed uniformly. This is indeed the case for teaching materials concerning the k-means clustering algorithm. In this guide, we provide notation translations for and commentary on a select set of resources instructor and students may find helpful. It is our hope that this will enable the instructor/student to more easily gain insight from these excellent supplementary materials.

Our Slides

number of data points n

data point dimensions d

data point $\mathbf{x}_i, i \in \{1, \ldots, n\}$

number of clusters k

data point cluster number $C(\mathbf{x}_i) \in \{1, \ldots, k\}$

centroid $\mu_j, j \in \{1, \ldots, k\}$

optimization function The WCSS (Within-Cluster Sum-of-Squares) measure is $\sum_{i=1}^{n} \left\|\mathbf{x}_i - \mu_{C(\mathbf{x}_i)}\right\|^2$.

Textbooks

The Elements of Statistical Learning

In section 14.3.6 of "The Elements of Statistical Learning" (Hastie, Tibshirani, and Friedman 2001)[1], we see the following notational differences:

number of data points N

data point dimensions p

data point $x_i, i \in \{1, \ldots, N\}$

number of clusters K

data point cluster number $C(i) \in \{1, \ldots, K\}, i \in \{1, \ldots, N\}$

centroid $\bar{x}_k$ and $m_k, k \in \{1, \ldots, K\}$

optimization function Within-point scatter

$$W(C) = \sum_{k=1}^{K} N_k \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2,$$

where N_k is the number of points assigned to cluster k.

An Introduction to Statistical Learning with Applications in R

In section 10.3.1 of "An Introduction to Statistical Learning with Applications in R" (James et al. 2014)[2], we see the following notational differences:

number of data points n

data point dimensions p

data point $x_i, i \in \{1, \ldots, n\}$

number of clusters K

data point cluster number

$$i \in C_k, k \in \{1, \ldots, K\},$$

if and only if point x_i is assigned to cluster number k

centroid $\bar{x}_k, k \in \{1, \ldots, K\}$

optimization function Total within-cluster variation

$$\sum_{k=1}^{K} W(C_k),$$

where within-cluster variation

$$W(C) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2,$$

where $|C_k|$ is the size of the kth cluster set.

Introduction to Machine Learning

In section 7.3 of "Introduction to Machine Learning" (Alpaydin 2010), we see the following notational differences:

number of data points N

data point dimensions d

[1] http://statweb.stanford.edu/~tibs/ElemStatLearn/

[2] http://www-bcf.usc.edu/~gareth/ISL/

data point $\mathbf{x}^t, t \in \{1, \ldots, N\}$

number of clusters k

data point cluster number i such that $b_i^t = 1$, where

$$b_i^t = \begin{cases} 1 & \text{if } \|\mathbf{x}^t - \mathbf{m}_i\| = \min_j \|\mathbf{x}^t - \mathbf{m}_j\| \\ 0 & \text{otherwise} \end{cases}$$

centroid $\mathbf{m}_i, i \in \{1, \ldots, K\}$

optimization function (not specified)

Internet Resources

Wikipedia

In Wikipedia's "k-means clustering" article (Wikipedia 2015), we see the following notational differences:

number of data points n

data point dimensions d

data point $\mathbf{x}_i$ or $x_i, i \in \{1, \ldots, n\}$

number of clusters k

data point cluster number i for $x_p \in S_i^{(t)}$ in iteration t where cluster

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \forall j, 1 \leq j \leq k \right\}$$

centroid μ_i or $m_i(t)$ in iteration t, $i \in \{1, \ldots, k\}$

optimization function $\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mu_i\|^2$

OnMyPhd

In OnMyPhD.com's "K-Means Clustering" article (Conceição 2015), we see the following notational differences:

number of data points n

data point dimensions d

data point $\mathbf{x}_i, i \in \{1, \ldots, n\}$

number of clusters k

data point cluster number $\mathbf{x}_j$ is in cluster i if and only if $j \in \mathbf{c}_i$ where

$$\mathbf{c}_i = \{j : d(\mathbf{x}_j, \mu_i) \leq d(\mathbf{x}_j, \mu_l), l \neq i, j = 1, \ldots, n\}$$

where $d(\mathbf{x}_j, \mu_i)$ is the Euclidean distance between $\mathbf{x}_j$ and μ_i.

centroid $\mu_i, i \in \{1, \ldots, k\}$

optimization function $\arg\min_{\mathbf{c}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathbf{c_i}} \|\mathbf{x} - \mu_i\|^2$

Commentary

number of data points All sources use n or N.

data point dimensions Statistical Learning sources use p but all others use d.

data point All sources use vector x (bold or not) with a subscript indicating the index in the training set, except for Alpaydin who uses a superscript.

number of clusters All sources use k or K.

data point cluster number Significant variation in notation exists for relating data points to cluster numbers. These include a mapping of point to cluster number (Hastie, Tibshirani, and Friedman 2001, our slides), clusters defined as a set of points (Wikipedia 2015), clusters defined as a set of point indices (James et al. 2014; Conceição 2015), and clusters defined by minimum distance to centroids (Alpaydin 2010). While these look different on the surface, these are equivalent interpretations.

It is also worth noting that all sources do not well define how one uniquely assigns a point to a cluster when the minimum distance centroid is not unique. In practice, breaking ties by minimum cluster index is not problematic, but mathematically, this appears to be an overlooked issue.

centroid All sources use $\bar{x}$ (mean of x's), or relate to the mean through the use of m or μ (mu).

optimization function Here we see the greatest variation of all. The process (beyond initialization) is uniformly described, but different authors vary significantly in the expression of *what* is minimized[3]. We find our expression to be simplest, and we encourage instructors to reinforce the simple idea of k-means clustering:

Alternate the following until no change occurs:

- Holding cluster assignments fixed, minimize WCSS by changing centroids to cluster means.
- Holding centroids fixed, minimize WCSS by changing cluster assignments according to closest centroids.

References

[Alpaydin 2010] Alpaydin, E. 2010. *Introduction to Machine Learning*. The MIT Press, 2nd edition.

[Conceição 2015] Conceição, H. 2015. OnMyPhd.com: K-means clustering. `http://www.onmyphd.com/?p=k-means.clustering`. Accessed: 2015-11-16.

[Hastie, Tibshirani, and Friedman 2001] Hastie, T.; Tibshirani, R.; and Friedman, J. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

[James et al. 2014] James, G.; Witten, D.; Hastie, T.; and Tibshirani, R. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

[Wikipedia 2015] Wikipedia. 2015. Wikipedia: k-means clustering. `https://en.wikipedia.org/wiki/K-means_clustering`. Accessed: 2015-11-16.

[3] It is not even expressed in the case of (Alpaydin 2010)