k-means clustering

Wikipedia

2019년 03월 06일

Table of Contents

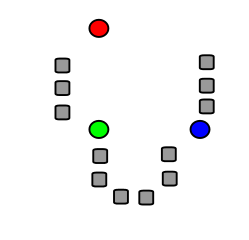
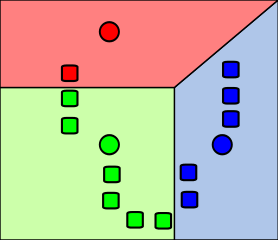
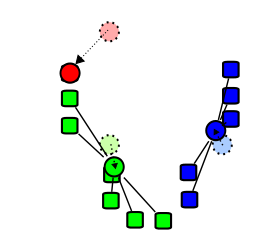
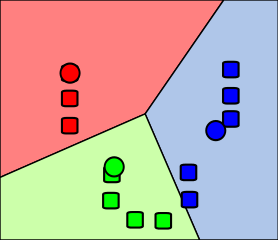
# Description

Given a set of observations , where each observation is a *d*-dimensional real vector, k-means clustering aims to partition the *n* observations into sets so as to minimize the within-cluster sum of squares (WCSS) (i.e. variance). Formally, the objective is to find:

# History

The term “k-means” was first used by James MacQueen in 1967(MacQueen 1967), though the idea goes back to Hugo Steinhaus in 1957(Steinhaus 1956). The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation, though it wasn’t published outside of Bell Labs until 1982(Lloyd 1982). In 1965, E. W. Forgy published essentially the same method, which is why it is sometimes referred to as Lloyd-Forgy(Forgy 1965).

# Algorithms

1. Forgy, E. 1965. “Cluster Analysis of Multivariate Data: Efficiency Versus Interpretability of Classification.” *Biometrics*, no. 3: 768–69.
2. Lloyd, S. 1982. “Least Squares Quantization in Pcm.” *IEEE Transactions on Information Theory*, no. 2. NJ, USA: IEEE Press Piscataway: 129–37.
3. MacQueen, J. 1967. “Some Methods for Classification and Analysis of Multivariate Observations.” In *Berkeley Symposium on Mathematical Statistics and Probability*, 281–97. University of California Press.
4. Steinhaus, H. 1956. “Sur La Division Des Corps Matériels En Parties.” *Bull. Acad. Polon. Sci.*, 801–4.