K-means Erratum for Elements of Statistical Learning

Tim van Erven

November 23, 2015

The book claims that K-means minimizes the within-point scatter, which in (14.31) is expressed as

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

where $\bar{x}_k$ is the mean of the k-th cluster and N_k is the number of observations assigned to the k-th cluster.

This is close, but not quite what the definition should be. In fact, K-means minimizes the sum of squared distances to the cluster means:

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

Note the absence of the factor N_k.