Clustering

Clustering techniques apply when there is no class to be predicted but rather when the instances are to be divided into natural groups. These clusters presumably reflect some mechanism at work in the domain from which instances are drawn, a mechanism that causes some instances to bear a stronger resemblance to each other than they do to the remaining instances. Clustering naturally requires different techniques to the classification and association learning methods.

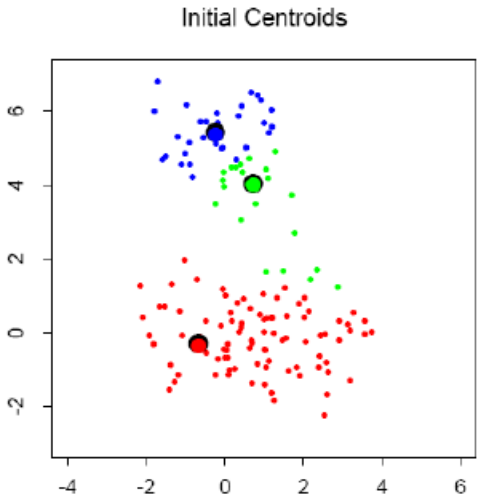
Iterative distance-based clustering (k-means)

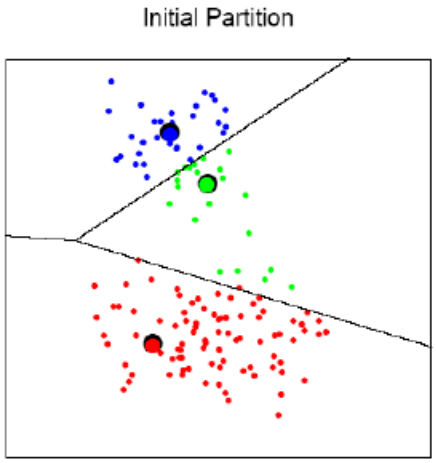
k is the specified in advance number of clusters are being sought. Then k points are chosen at random as cluster centers. All instances are assigned to their closest cluster center according to the ordinary Euclidean distance metric. Next the centroid, or mean, of the instances in each cluster is calculated—this is the “means” part. These centroids are taken to be new center values for their respective clusters. Finally, the whole process is repeated with the new cluster centers. Iteration continues until the same points are assigned to each cluster in consecutive rounds, at which stage the cluster centers have stabilized and will remain the same .

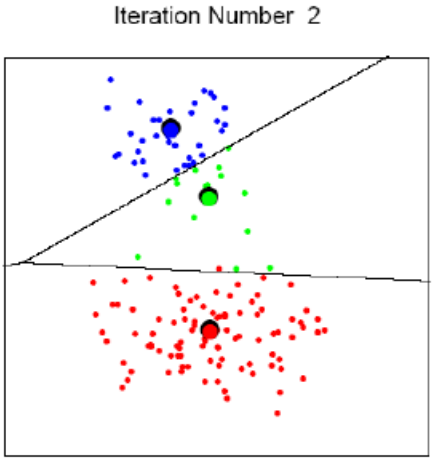
$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

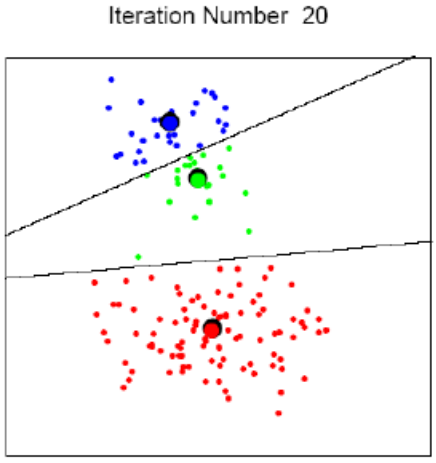
Euclidean distance

This clustering method easily proving that choosing the cluster center to be the centroid minimizes the total squared distance from each of the cluster’s points to its center. Once the iteration has stabilized, each point is assigned to its nearest cluster center, so the overall effect is to minimize the total squared distance from all points to their cluster centers. But the minimum is a local one; there is no guarantee that it is the global minimum.


Initial Centroids
6
4
2
0
-2
-4
-2
0
2
4
6


Initial Partition


Iteration Number 2


Iteration Number 20

The final clusters are quite sensitive to the initial cluster centers. Completely different arrangements can arise from small changes in the initial random choice. In fact, this is true of all practical clustering techniques: it is almost always infeasible to find globally optimal clusters.

But in order to increase the chance of finding a global minimum, we often run the algorithm ,several times, each with a random initial values (centroids). Every squared distance is taking into account, and the chosen initial centroids are the ones with minimum total squared distance.
The k-means clustering algorithm usually requires several iterations, each involving finding the distance of k cluster centers from every instance to determine its cluster which means that its speed it relatively slow.

Strength

- *Relatively efficient*: *O*(*tkn*), where *n* is # instances, and *t* is # iterations. Normally, *k*, *t* << *n*.
- Often terminates at a *local optimum*. The *global optimum* may be found using techniques such as: *simulated annealing* or *genetic algorithms*

Weakness

- Applicable only when *mean* is defined; what about categorical data?
- Need to specify c, the *number* of clusters, in advance
- Unable to handle noisy data and *outliers*
- Not suitable to discover clusters with *non-convex shapes*

In visual vocabulary construction , the vocabulary is a way of constructing a feature vector for classification that relates “new” descriptors in query images to descriptors previously seen in training. One extreme of this approach would be to compare each query descriptor to all training descriptors: this seems impractical given the huge number of training descriptors involved .
In practice it was found that the best trade-offs of accuracy and computational efficiency are obtained for intermediate sizes of clustering.
Most clustering or vector quantization algorithms* are based on iterative square-error

*__Vector quantization__ is a classical quantization technique from signal processing which allows the modeling of probability density functions by the distribution of prototype vectors.
A vector quantizer maps k-dimensional vectors in the vector space R^K into a finite set of vectors $Y = \{y_i: i = 1, 2, ..., N\}$. Each vector y_i is called a code vector or a codeword. and the set of all the codewords is called a codebook.

partitioning or on hierarchical techniques. Square-error partitioning algorithms -the one described earlier- attempt to obtain the partition which minimizes the within-cluster scatter or maximizes the between-cluster scatter.
There exist methods allowing to automatically estimating the number of clusters. For example, Pelleg et al [24] use cluster splitting to do it, where the splitting decision is done by computing the Bayesian Information Criterion.
However, in the present case we do not really know anything about the density or the compactness of our clusters. Moreover, we are not even interested in a “correct clustering” in the sense of feature distributions, but rather in accurate categorization. We therefore run k-means several times with different number of desired representative vectors (k) and different sets of initial cluster centers. We select the final clustering giving the lowest empirical risk in categorization [25].

References:

[24] D. Pelleg and A. Moore. X-Means: Extending K-means with Efficient Estimation of the Number of Clusters, International Conference on Machine Learning, 2000.
[25] V. Vapnik. Statistical Learning Theory. Wiley, 1998