Published on *STAT 897D* (https://onlinecourses.science.psu.edu/stat857)

Home > 12.1.4 - Classification by K-means

12.1.4 - Classification by K-means

The primary application of k-means is clustering or unsupervised classification. K-means alone is not designed for classification, but we can adapt it for the purpose of supervised classification.

If we use k-means to classify data, there are two schemes.

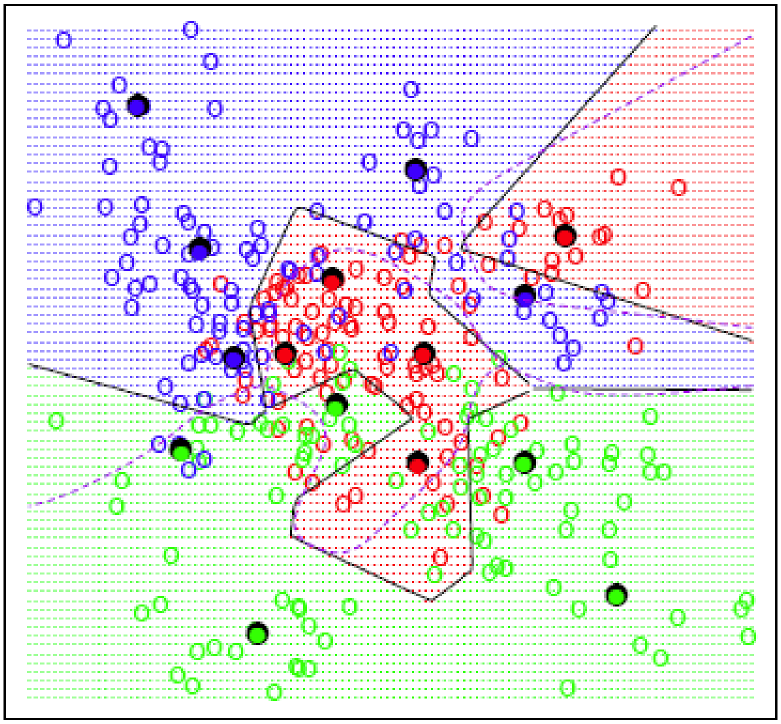
One method used is to separate the data according to class labels and apply k-means to every class separately. If we have two classes, we would perform k-means twice, once for each group of data. At the end, we acquire a set of prototypes for each class. When we have a new data point, we put all of the prototypes together and find which one is closest to the new data point. This prototype is associated with a class because the prototypes are created by clustering each class of data individually. The class of this prototype is taken as the class of the new data point.

Here is a review of the basic steps involved:

- Apply k-means clustering to the training data in each class separately, using R prototypes per class.
- Assign a class label to each of the $K \times R$ prototypes.
- Classify a new feature vector x to the class of the closest prototype.

This above approach to using k-means for classification is referred to as Scheme 1.

Below is a result from the textbook using this scheme. There are three classes green, red, and blue. The authors applied k-means using 5 prototypes for each class. We can see below that for each class, the 5 prototypes chosen are shown by filled circles.



Inspect! [1]

According to the classification scheme, for any new point, among these 15 prototypes, we would find the one closest to this new point. Then, depending on the color code of that prototype, the corresponding class will be assigned to the new point.

The black lines are classification boundaries generated by the k-means algorithm. Specifically, these are the classification boundaries induced by the set of prototypes based on the nearest neighbor. The decision boundary between any two prototypes based on the nearest neighbor rule is linear. Every prototype occupies some region in the space. The region around each prototype is sometimes called the voronoi region and is bounded by hyperplanes. Because we have more than one prototype for each class, the classification boundary between the two classes is connected segments of the straight lines, which gives a zigzag look.

Another Approach for using K-Means - Scheme 2

The second scheme of classification by k-means is to put all the data points together and perform k-means once. There's no guarantee that points in the same group are of the same class because we conducted k-means on the class blended data. To associate a prototype with a class, we count the number of data points in each class that are assigned to this prototype. The dominant class with the most data points is associated with the prototype. During the classification of a new data point, the procedure then goes in the same way as Scheme 1.

We new summarize the steps of Scheme 2:

- Apply k-means clustering to the entire training data, using *M* prototypes.
- For each prototype, count the number of samples from each class that are assigned to

this prototype. Associate the prototype with the class that has the highest count.
- Classify a new feature *x* to the class of the closest prototype.

This alternative approach is referred to as Scheme 2.

Source URL: https://onlinecourses.science.psu.edu/stat857/node/113

Links:
[1] https://onlinecourses.science.psu.edu/stat857/javascript:popup_window('/stat857/sites/onlinecourses.science.psu.edu.stat857/files/lesson06/Kmeans_viewlet_swf.html', 'kmeans', 704, 652);