Clueless k-means

François fleuret

January 23, 2015

1 Objective

Given a labeled training set

$$(x_n, y_n) \in \mathbb{R}^D \times \{1, \ldots, Q\}, \ n = 1, \ldots, N$$

we want to estimate centroids

$$c_k \in \mathbb{R}^D, \ k = 1, \ldots, K.$$

such that the clustering is "as poorly informative" about y as possible.

This means that while we still want the clusters to be "compact" geometrically, we want the values of ys to be equally represented in all clusters, so that knowing to which cluster a sample belongs does say a lot about its position x as usual, but does not say anything about its class y.

2 Standard k-Means

We reformulate the standard k-means, which is completely unsupervised, hence does not take the y_n into account, as follows:

Let

$$\gamma(n, k) \in \{0, 1\}, \ n = 1, \ldots, N, \ k = 1, \ldots, K$$

stands for the "memberships" of x_n to the cluster defined by c_k, with $\gamma(n, k) = 0$ standing for "x_n does not belong to cluster k", and $\gamma(n, k) = 1$ stands for "x_n belongs to cluster k".

The k-means procedure consists of alternating the two following steps:

1. Re-estimating the centroids

$$\forall k, c_k \leftarrow \frac{\sum_n \gamma(n,k) x_n}{\sum_n \gamma(n,k)}$$

2. Re-estimating the memberships of each sample

$$\forall n, \gamma(n,1), \ldots, \gamma(n,K) \leftarrow \underset{g_1,\ldots,g_K \in [0,1], \sum_k g_k = 1}{\operatorname{argmin}} \sum_k g_k \|x_n - c_k\|^2$$

   Under that form, and even if we have relaxed the constraints for the γ to be integers, that step always picks a value in $\{0,1\}^K$.

Interestingly the second step is a linear optimization under linear constraints.

3 Clueless k-Means

We want to enforce that the classes are equally represented in all clusters, that is

$$\forall k, q, \ \frac{\sum_n \gamma(n,k) \mathbf{1}_{\{y_n=q\}}}{\sum_n \gamma(n,k)} = \frac{\sum_n \mathbf{1}_{\{y_n=q\}}}{N},$$

which is equivalent to

$$\forall k, q, \ \sum_n \left(\mathbf{1}_{\{y_n=q\}} - \frac{1}{N} \sum_m \mathbf{1}_{\{y_m=q\}} \right) \gamma(n,k) = 0. \tag{1}$$

A related criterion that we call "absolute" forces the fraction of samples from each class to be the same in each cluster:

$$\forall k, q, \ \sum_n \gamma(n,k) \mathbf{1}_{\{y_n=q\}} = \sum_n \frac{1}{K} \mathbf{1}_{\{y_n=q\}}. \tag{2}$$

We can add (1) or (2) as constraints to the second step described in the previous section, which remains a linear optimization under linear constraints and can be solved with a standard LP optimizer.

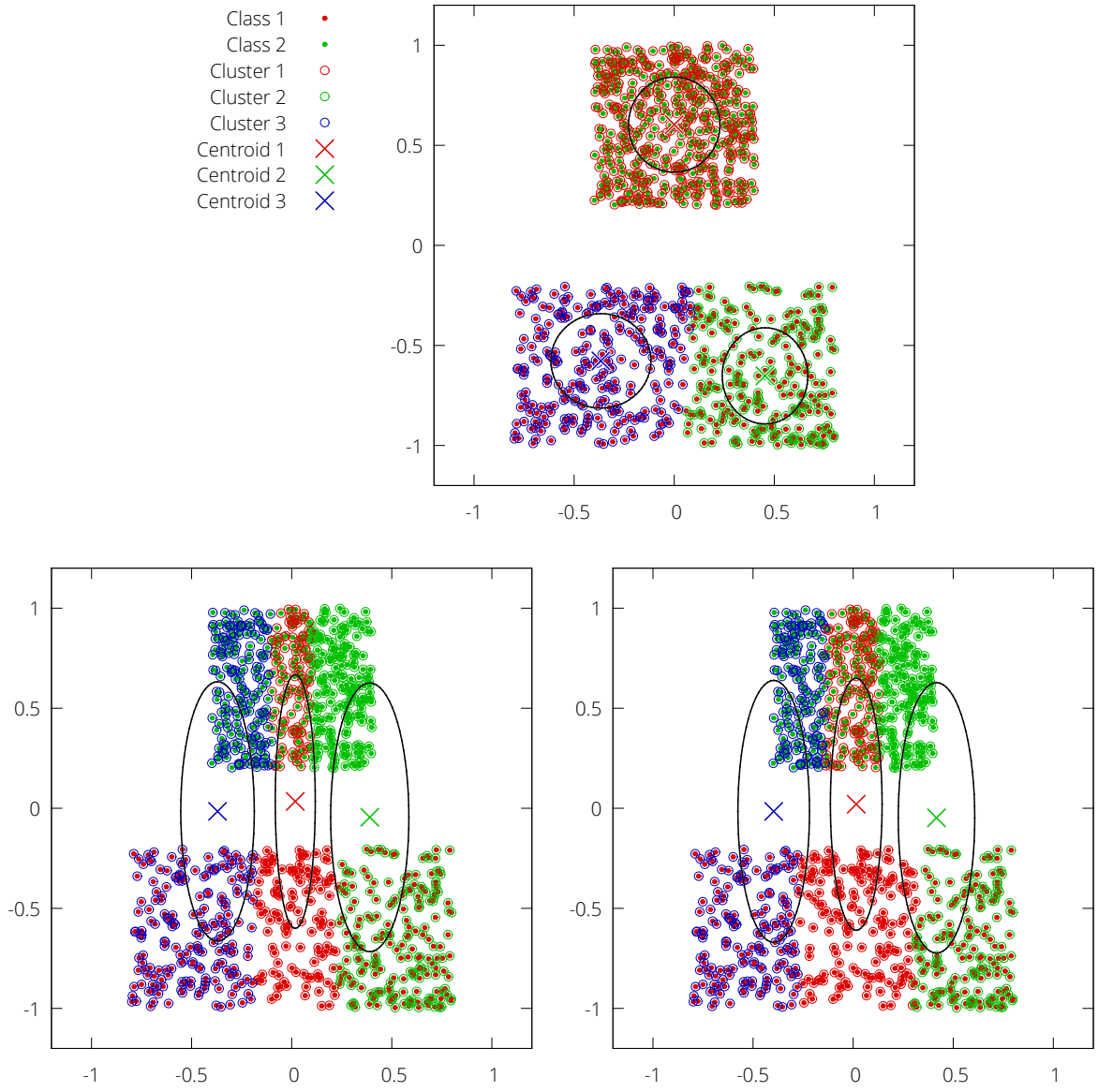

Figure 1: Results on a synthetic $2d$ problem, where the population is composed of two classes, one with samples uniformly distributed over a square at the top, and one with samples uniformly distributed in a rectangle at the bottom. We compare the standard k-means (top graph) with the two variants of the clueless k-means (bottom left graph for the standard one, bottom right for the "absolute" one), using three clusters. As expected in the bottom figure, the clusters are chosen such that they all capture samples from the two classes.