# Imbalanced K Means Algorithm

Zou Xueyan 13253956

June 6, 2016

**Abstract**

This report is going to illustrate the relationship between the performance of Kmeans algorithm and the property of clusters (E.g. standard deviation, density ratio, distance) . . .

# 1 Introduction

In this section, parameters used in the report would be demonstrated, then comes the brief introduction of the experiment procedure.

## 1.1 Parameter Demonstration

In Figure 1, $(x_1, y_1) = (3, 3)$, which is set manually. Then base on $(x_1, y_1) = (3, 3)$, $(x_2, y_2) = (x_1 + d, y_1)$, where $d$ is the distance between two cluster centers.

$n_1$, $n_2$ is the number of data in $cluster_1$ and $cluster_2$ respectively. For example, point $(x_i, y_i)$ is in $cluster_1$, then $(x_i, y_i) = (x_1 + x_t, y_1 + y_t)$, where $t \in [0, 50)$, and $x_t$ is generated randomly with mean=0, standard deviation = $\sigma_1$($\sigma_1$ and $\sigma_2$ are the standard deviation of $cluster_1$ and $cluster_2$) in normal distribution.

Then we define $d = 4\sigma_1 + 4\sigma_2$, which could offer more than 99.9% confidence that two clusters will not overlap.

$(x_j, y_j)$ is the initial seed point of Kmeans algorithm. In order to exclude the influence of the initial position of seed points, $(x_j, y_j) = (x_1, y_1 + 2/3 *$

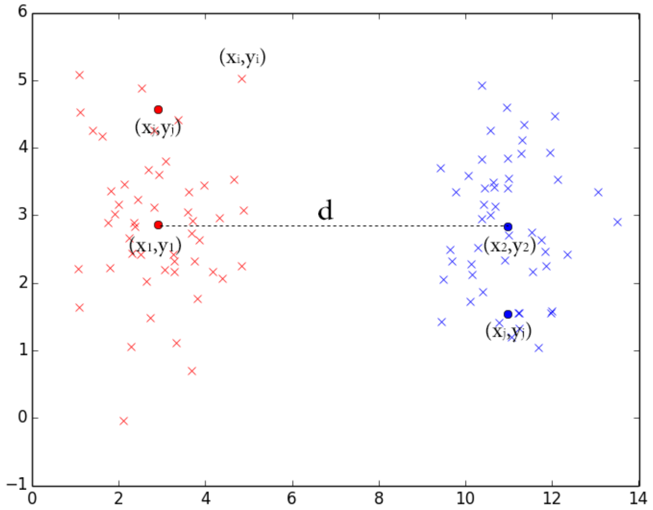

Figure 1: Initial cluster

$max(y_t)$) so that the seed points are ensured to appear in the clusters.

The cluster density $\rho$ is defined as the number of points inside unit area.

$$\rho_1 = \frac{n_1}{\pi\sigma_1^2} \tag{1}$$

For the measurement of error $\varepsilon$, we define it to be the distance between the real center of two clusters and the final position of the seed points. The cluster density $\rho$ is defined as the number of points inside unit area.

$$\sqrt{(x_1 - x_j)^2 + (y_1 - y_j)^2} \tag{2}$$

## 1.2 Experiment Introduction

The following experiment is to explore the performance of Kmeans Algorithm in relation with standard deviation, density ratio and distance.

The first experiment is going to estimate the relationship with error and standard deviation, which gives out a linear result say $y = kx + b$. Then the second experiment is to explore what factors influence $k$ and $b$.

## 2 Experiment: Error and standard deviation

In this section, standard deviation is increased in a learning rate for each iteration(Iteration means a whole procedure of Kmeans Algorithm). The pseudo code of the procedure is listed below. After executing the program, we plot the content of output in figure 2.

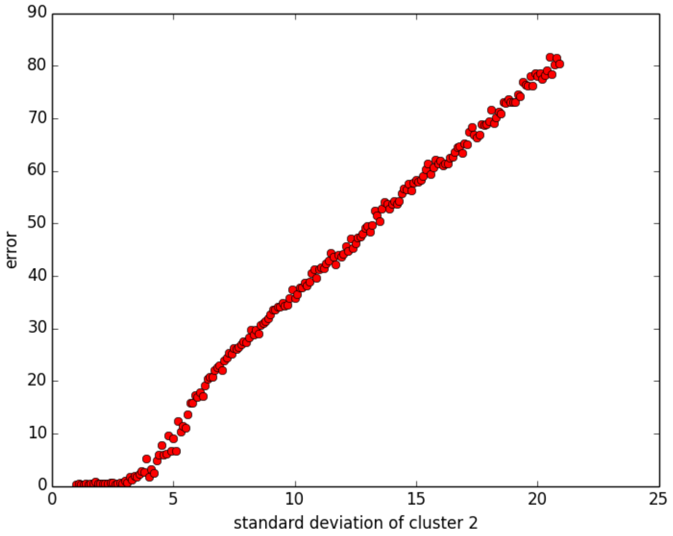

Figure 2: standard deviation and error

```
$n_1 = 50$, $n_2 = 50$
$\sigma_1 = \sigma_2 = 1$
learnRate=0.1
$\varepsilon = 0$
T=1
output=[]

while T¡201:
    $d = 4\sigma_1 + 4\sigma_2$
    $n_2 = n_1/\sigma_1^2 * \sigma_2^2$
     KmeansAlgorithm()
    $\varepsilon$=calculateError()
     output.append([$\sigma_2$,$\varepsilon$])
    $\sigma_2 = \sigma_2 + learnRate$
    T=T+1
```

Figure 2 shows a perfect linear relationship between $\varepsilon$ and $\sigma_2$. After doing the linear regression, which shows the following result:

$$y = 4.43\sigma_2 - 9.446 \qquad (r = 0.996) \tag{3}$$

This equation reveals that there is a positive linear relationship between error and standard deviation. After getting this result, the next step is to figure out what could influence the slope and the intersection point with y axis.

## 3 Experiment: k,b and desity ratio

According to the previous section, we define the density ratio $dr$ as following:

$$\rho_1 = \frac{n_1}{\pi\sigma_1^2} \qquad \rho_2 = \frac{n_2}{\pi\sigma_2^2} \tag{4}$$

$$dr = \frac{\rho_2}{\rho_1} \tag{5}$$

After each iteration of section 2, we increase $dr$ by 1. Then we could generate a set of data of $dr$, $k$, $b$ and $r$ (correlation coefficient). We plot $lg(dr)$, $\varepsilon$, and $1/lg(dr+1)$, $\varepsilon$ in figure 3 and figure 4

So that we could conclude $1/lg(dr+1)$, $lg(dr)$ have positive linear relationships with k, b respectively.

The relationship between error and distance is a little bit complicated, which is still study in progress.

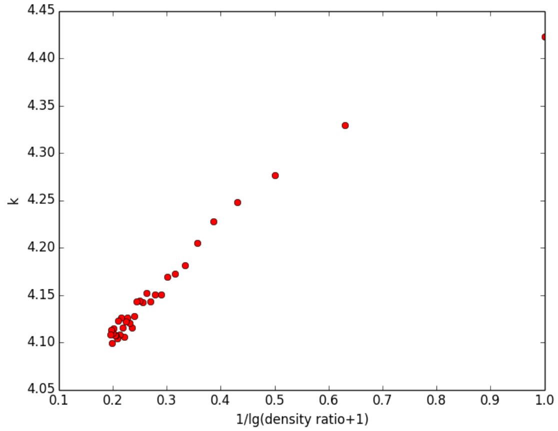

Figure 3: 1/lg(density ratio+1) and k, r=0.976

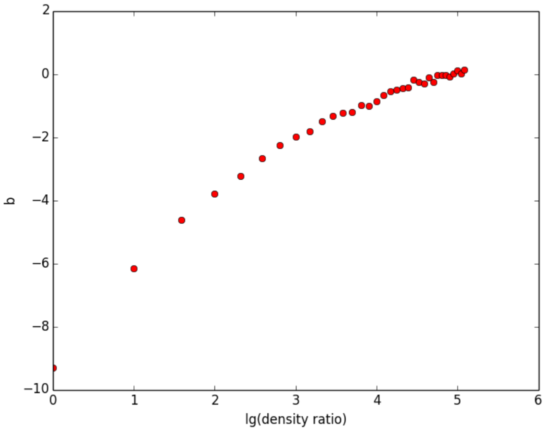

Figure 4: lg(density ratio) and b, r=0.966