MACHINE LEARNING

Kernel for Clustering

kernel K-Means

Interactive lecture and exercises

Kernel K-means: Algorithm

Kernel K-means algorithm is also an iterative procedure:

1. **Initialization:** pick K clusters (random assignment of points to a cluster, or use K-means at initialization)

2. **Assignment Step:** Assign each data point to its "closest" centroid (E-step).

$$\arg\min_k d\left(x, C^k\right) = \min_k \left(k\left(x,x\right) - \frac{2\sum_{x^j \in C^k} k\left(x, x^j\right)}{m_k} + \frac{\sum_{x^j, x^l \in C^k} k\left(x^j, x^l\right)}{\left(m_k\right)^2} \right)$$

3. **Update Step:** Update the list of points belonging to each centroid (M-step)

4. Go back to step 2 and repeat the process until the clusters are stable.

Interpreting the objective function

What is the influence of this term on the clustering (when using the RBF kernel)?

$$\arg\min_k d\left(x, C^k\right) = \min_k \left(\boxed{k\left(x,x\right)} - \frac{2\sum_{x^j \in C^k} k\left(x, x^j\right)}{m_k} + \frac{\sum_{x^j, x^l \in C^k} k\left(x^j, x^l\right)}{\left(m_k\right)^2} \right)$$

A. It gives more weight to points close to the cluster.

B. It gives less weight to points close to the cluster.

C. It has no influence. ✓

D. I do not know.

$k\left(x,x\right)$ depends only on the query datapoint x

It is the same for all clusters and hence has no influence on cluster allocation.

Interpreting the objective function

What is the influence of this term on the clustering (when using the RBF kernel)?

$$\arg\min_k d\left(x, C^k\right) = \min_k \left(\mathrm{k}\left(x, x\right) - \frac{2 \sum_{x^j \in C^k} \mathrm{k}\left(x, x^j\right)}{m_k} + \frac{\sum_{x^j, x^l \in C^k} \mathrm{k}\left(x^j, x^l\right)}{\left(m_k\right)^2} \right)$$

A. The closer the point x is to the cluster, the larger is this term. ✓

B. The denser the cluster, the larger. ✓

C. The more spread out the cluster is, the larger.

D. I do not know.

$$0 < k\left(x, x^i\right) = e^{-\frac{\left\|x - x^i\right\|^2}{\sigma^2}} \le 1$$

The closer the point to the center of the cluster, the larger the kernel.

Interpreting the objective function

What is the influence of this term on the clustering (when using the RBF kernel)?

$$\arg\min_k d\left(x, C^k\right) = \min_k \left(k\left(x, x\right) - \frac{2 \sum_{x^j \in C^k} k\left(x, x^j\right)}{m_k} + \frac{\sum_{x^j, x^l \in C^k} k\left(x^j, x^l\right)}{\left(m_k\right)^2} \right)$$

A. The closer the point x is to the cluster, the larger is this term.

B. The denser the cluster, the larger. ✓

C. The more spread out the cluster is, the larger.

D. I do not know.

$$0 < \frac{\sum_{x^j, x^l \in C^k} k\left(x^j, x^l\right)}{\left(m_k\right)^2} \leq 1$$

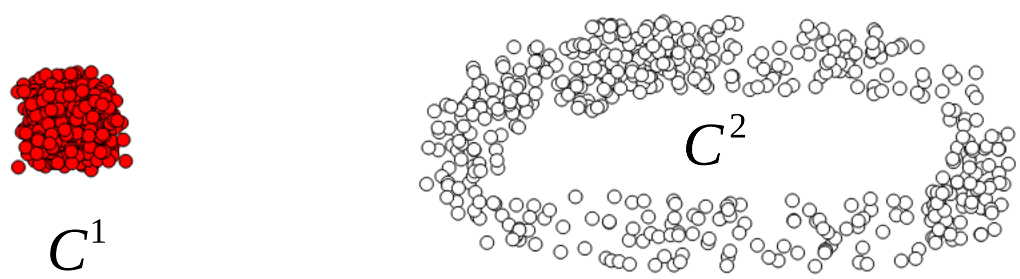
The closer the points are to one another in the same cluster, the larger the sum.

Kernel K-means: interpreting the solution

Density versus number of points

$$\arg\min_k d\left(x, C^k\right) = \min_k \left(k\left(x, x\right) - \frac{2\sum_{x^j \in C^k} k\left(x, x^j\right)}{m_k} + \frac{\sum_{x^j, x^l \in C^k} k\left(x^j, x^l\right)}{\left(m_k\right)^2} \right)$$

C^1 has same number of points than C^2, but is denser.

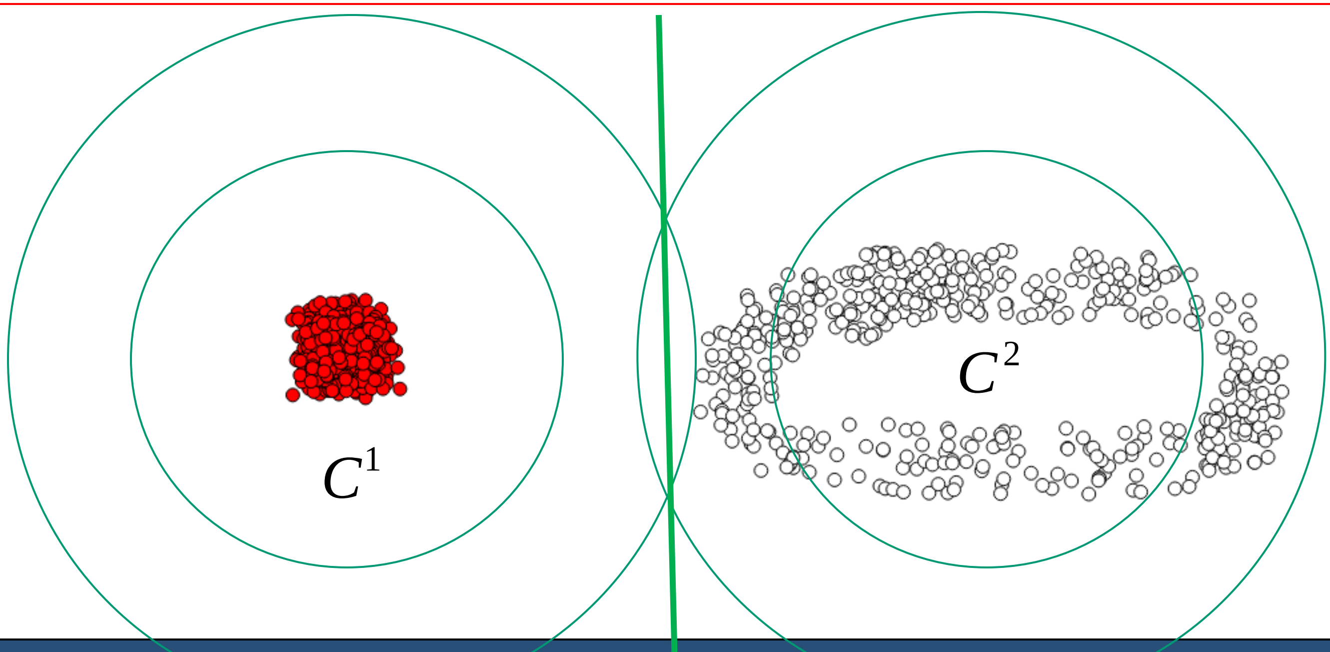


Kernel K-means: interpreting the solution

Density versus number of points

$$\arg\min_k d\left(x, C^k\right) = \min_k \left(\mathrm{k}\left(x, x\right) - \frac{2\sum_{x^j \in C^k} \mathrm{k}\left(x, x^j\right)}{m_k} + \frac{\sum_{x^j, x^l \in C^k} \mathrm{k}\left(x^j, x^l\right)}{\left(m_k\right)^2} \right)$$

Cutoff like classical K-means.

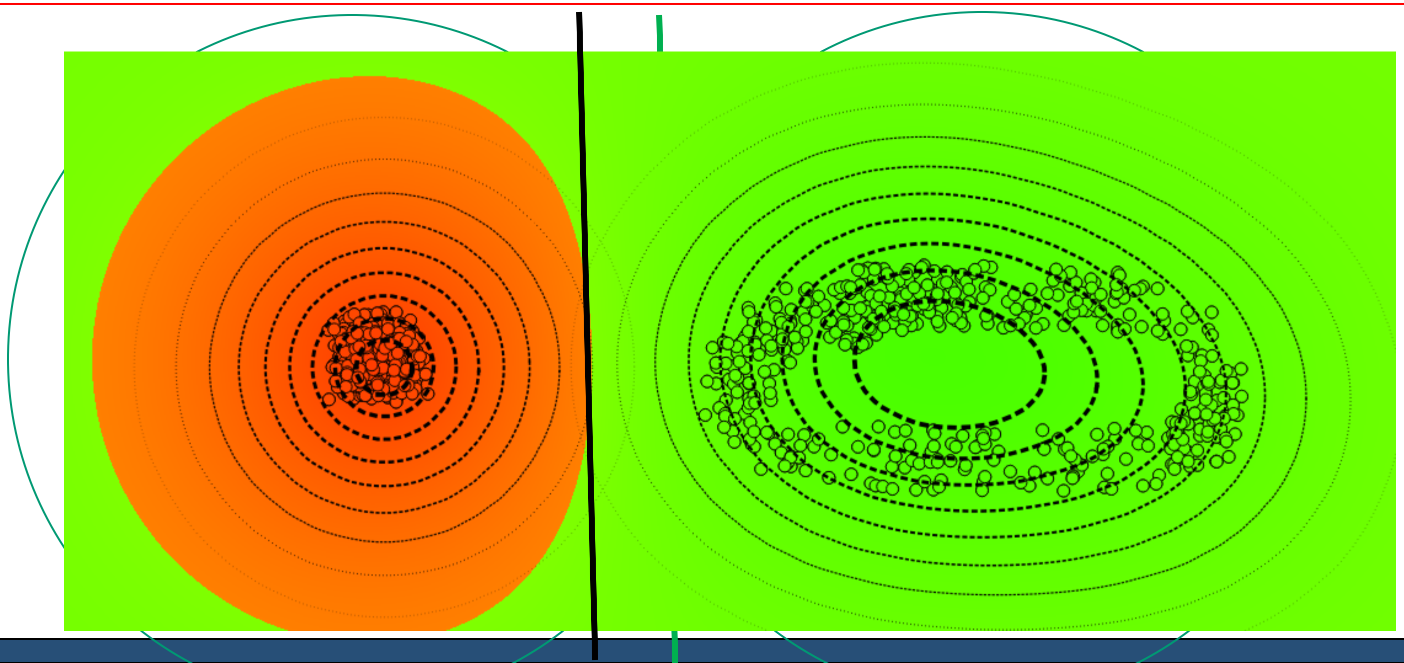


Kernel K-means: interpreting the solution

Density versus number of points

$$\arg\min_k d\left(x, C^k\right) = \min_k \left(k\left(x, x\right) - \frac{2 \sum_{x^j \in C^k} k\left(x, x^j\right)}{m_k} + \frac{\sum_{x^j, x^l \in C^k} k\left(x^j, x^l\right)}{\left(m_k\right)^2} \right)$$

Increases effect of spread out clusters



Kernel K-means: interpreting the solution

Assume that C2 has now twice more points than C1, does this affect the result?

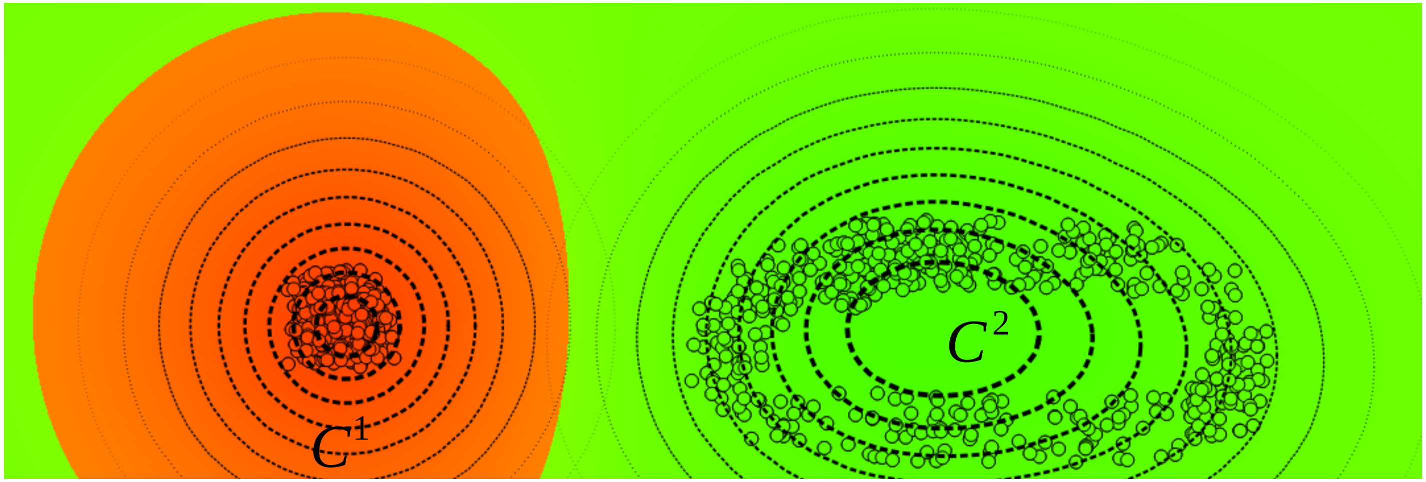
A. Yes

B. No

C. I do not know

Terms are unchanged

$$\arg\min_k d\left(x, C^k\right) = \min_k \left(k(x,x) - \frac{2\sum_{x^j \in C^k} k\left(x, x^j\right)}{m_k} + \frac{\sum_{x^j, x^l \in C^k} k\left(x^j, x^l\right)}{\left(m_k\right)^2} \right)$$



There is no difference if the additional datapoints are superimposed to the previous group.

The additional number is taken into account in the normalization.

Kernel K-means: interpreting the solution

Assume that C2 has now twice more points than C1, does this affect the result?

A. Yes

B. No

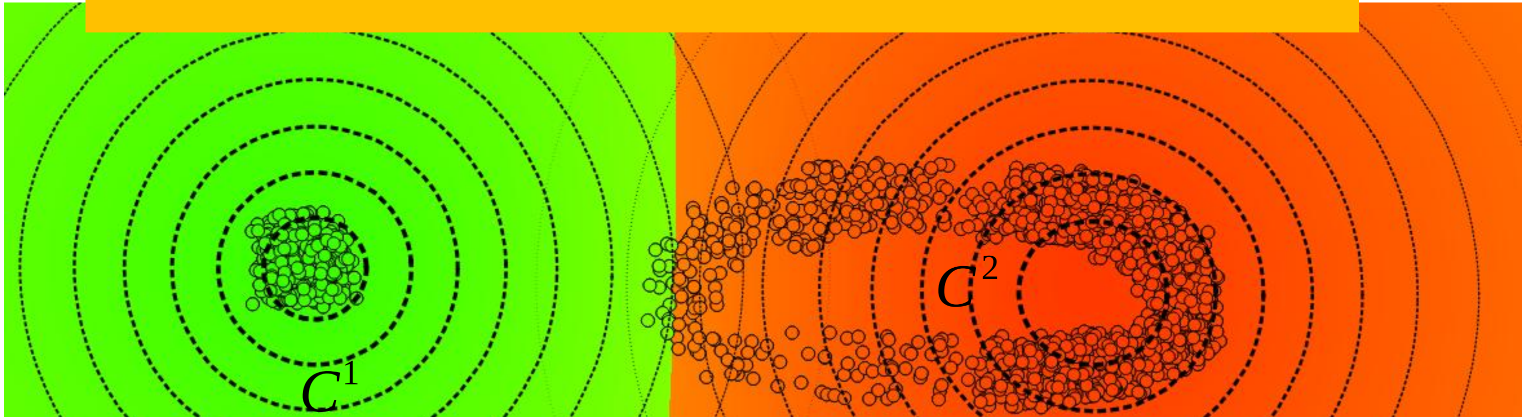
C. I do not know

Term decreases

Term increases

$$\arg\min_k d\left(x, C^k\right) = \min_k \left(k\left(x,x\right) - \frac{2\sum_{x^j \in C^k} k\left(x,x^j\right)}{m_k} + \frac{\sum_{x^j,x^l \in C^k} k\left(x^j,x^l\right)}{\left(m_k\right)^2} \right)$$

What if C^2 has twice more points than C^1, in the outer part.



In this case, this affects the result as it shifts the boundary, but this is not the result of having more points but of the centroid of C2 to shift to the right.

Kernel K-means: interpreting the solution

Density versus number of points

$$\arg\min_k d\left(x, C^k\right) = \min_k \left(\mathrm{k}(x,x) - \frac{2\sum_{x^j \in C^k} \mathrm{k}\left(x, x^j\right)}{m_k} + \frac{\sum_{x^j, x^l \in C^k} \mathrm{k}\left(x^j, x^l\right)}{\left(m_k\right)^2} \right)$$

Normalization factors cancel effect of # points and give a measure of average density / distance

Kernel K-means: interpreting the solution

With a polynomial kernel

$k\left(x^i, x^j\right) = \left(\left(x^i\right)^T x^j + c\right)^p \quad , c \in \mathbb{R}, \; p \in \mathbb{N}_+$

A: Affected by the position of the points from the origin (norm).

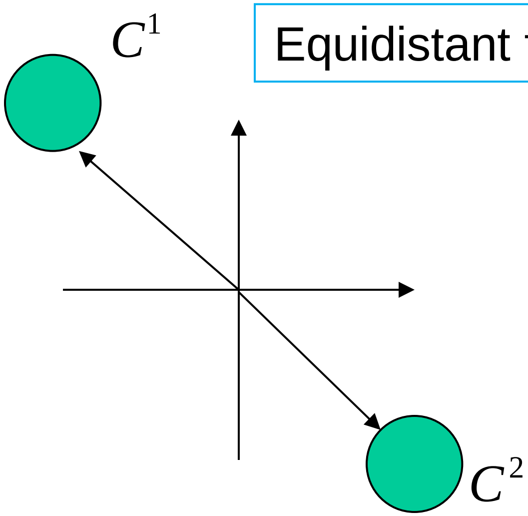
B: Affected by the relative angle across the points.

Norm - Positive value

$$\arg\min_k d\left(x, C^k\right) = \min_k \left(k(x,x) - \frac{2 \sum_{x^j \in C^k} k\left(x, x^j\right)}{m_k} + \frac{\sum_{x^j, x^l \in C^k} k\left(x^j, x^l\right)}{\left(m_k\right)^2} \right)$$

A datapoint will be assigned to the closest cluster in the closest partition.

Kernel K-means: interpreting the solution



Equidistant for *p=2, 4*, etc

Homogeneous Polynomial

$$k\left(x^i, x^j\right) = \left(\left(x^i\right)^T x^j\right)^p \ , \ p \in \mathbb{N}_+$$

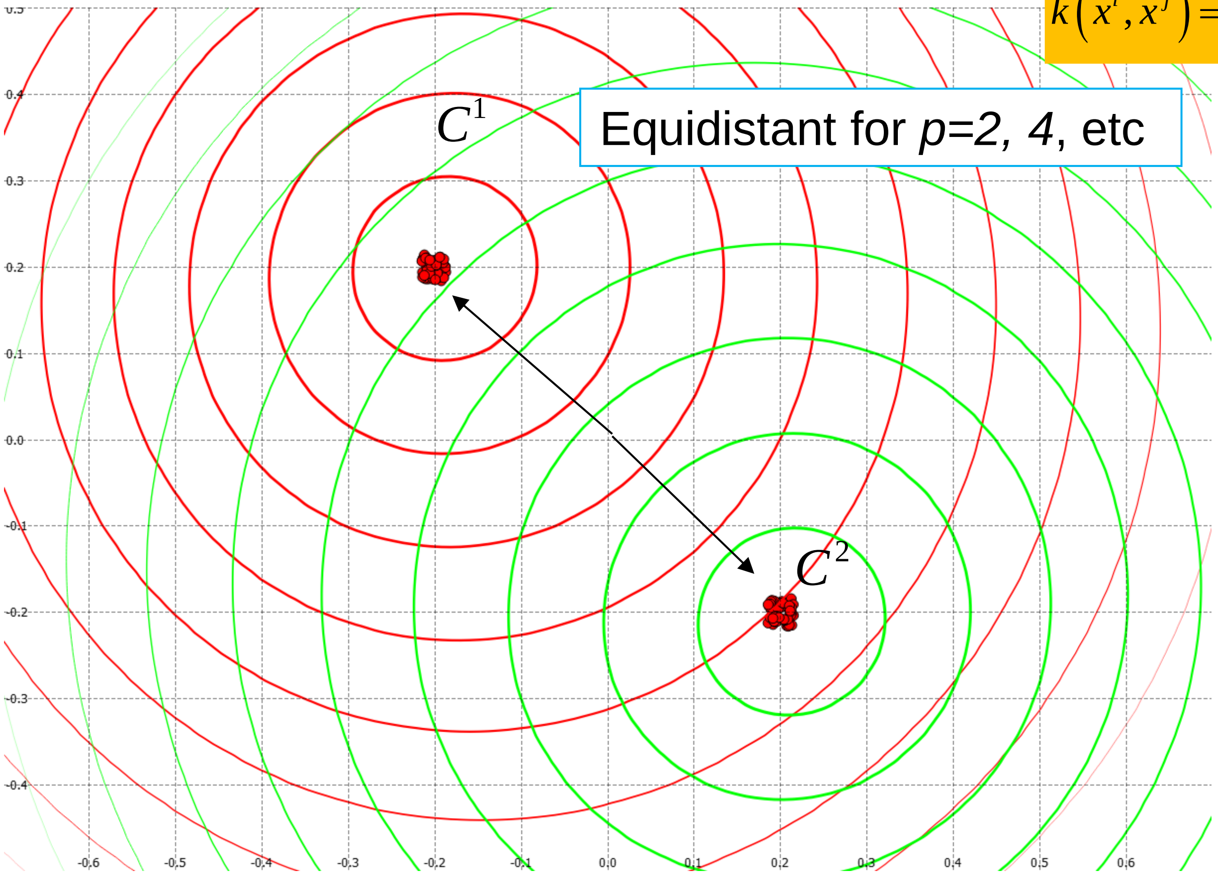
A: Affected by the position of the points from the origin (norm).

$$\arg\min_k d\left(x, C^k\right) = \min_k \left(\mathrm{k}\left(x, x\right) - \frac{2 \sum_{x^j \in C^k} \mathrm{k}\left(x, x^j\right)}{m_k} + \frac{\sum_{x^j, x^l \in C^k} \mathrm{k}\left(x^j, x^l\right)}{\left(m_k\right)^2} \right)$$

Kernel K-means: interpreting the solution

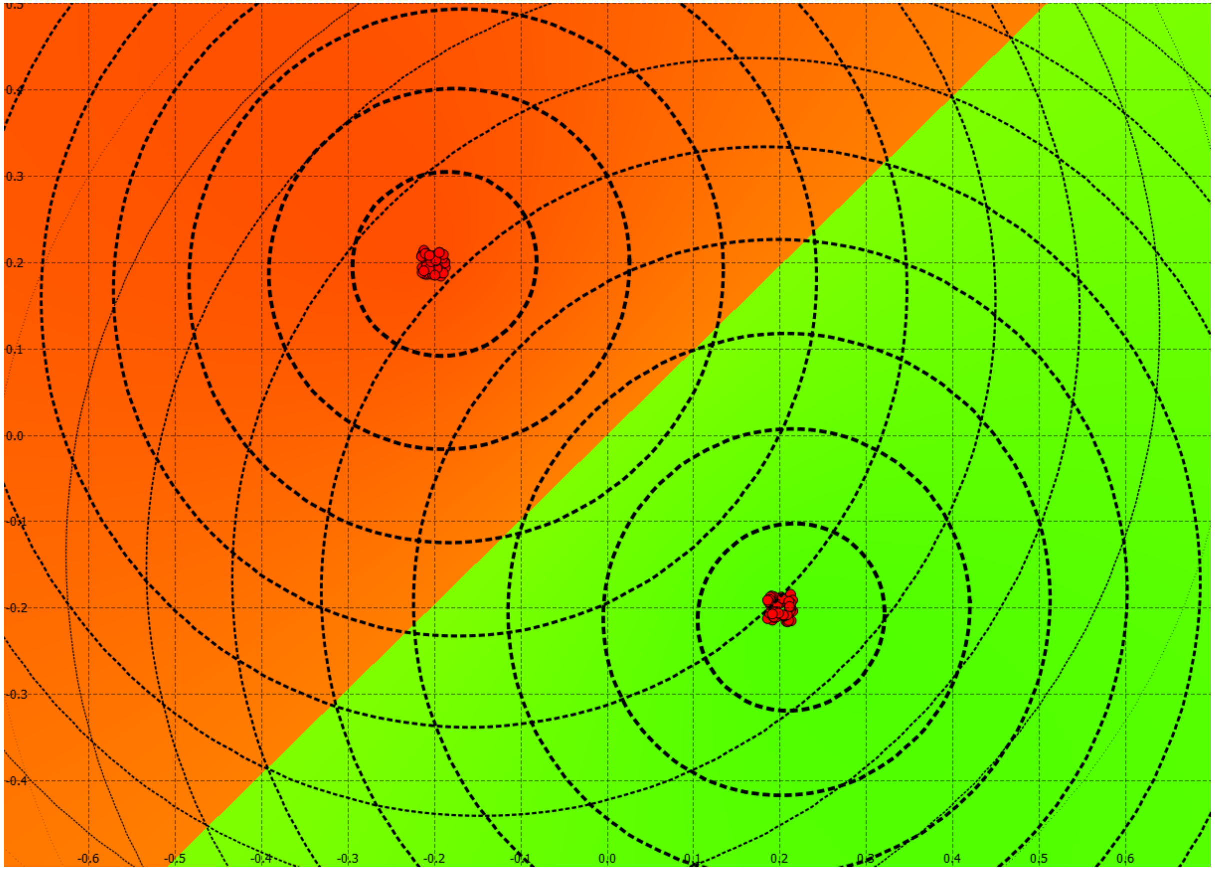
Homogeneous Polynomial

$$k\left(x^i, x^j\right) = \left(\left(x^i\right)^T x^j\right)^p \quad , \; p \in \mathbb{N}_+$$



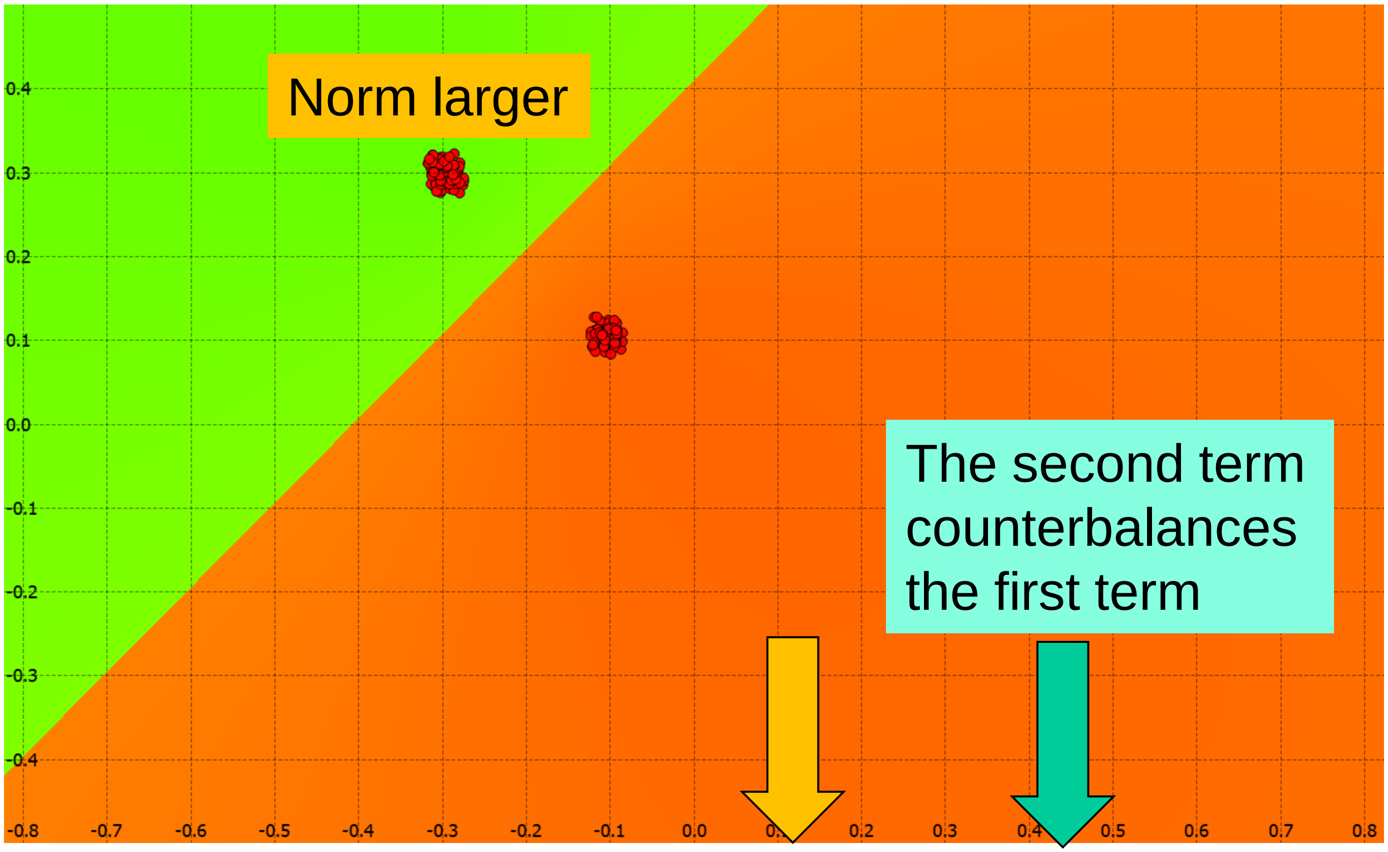
K=2, p=2

Kernel K-means: interpreting the solution



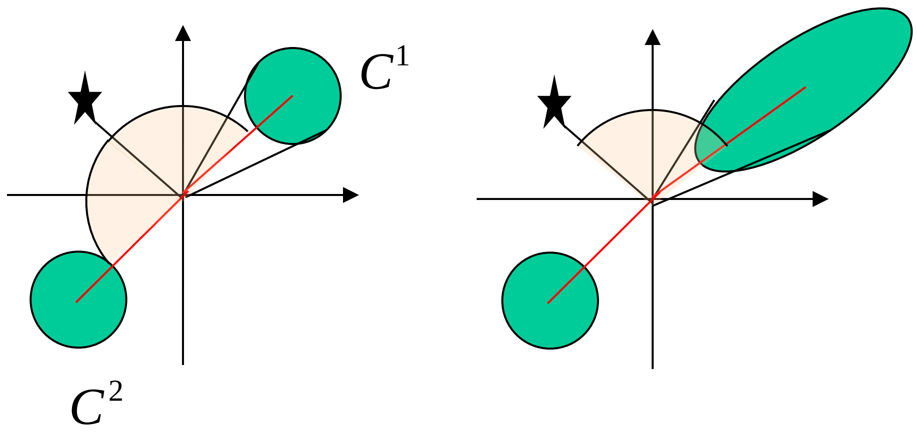
K=2, p=2

Kernel K-means: interpreting the solution



$$\arg\min_k d\left(x, C^k\right) = \min_k \left(k\left(x,x\right) - \frac{2 \sum_{x^j \in C^k} k\left(x, x^j\right)}{m_k} + \frac{\sum_{x^j, x^l \in C^k} k\left(x^j, x^l\right)}{\left(m_k\right)^2} \right)$$

Kernel K-means: interpreting the solution

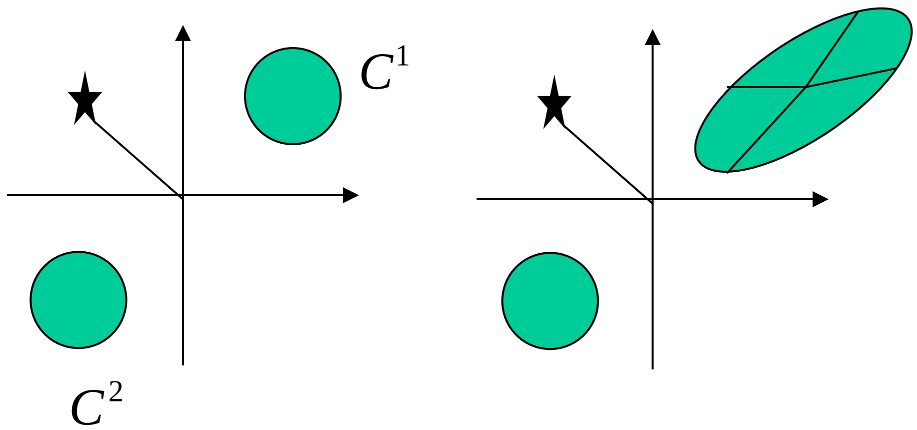


B: Affected by the relative angle across the points.

$$\arg\min_k d\left(x, C^k\right) = \min_k \left(k(x,x) - \frac{2\sum_{x^j \in C^k} k\left(x, x^j\right)}{m_k} + \frac{\sum_{x^j, x^l \in C^k} k\left(x^j, x^l\right)}{\left(m_k\right)^2} \right)$$

If the centroid of the cluster does not change, the term remains comparable.

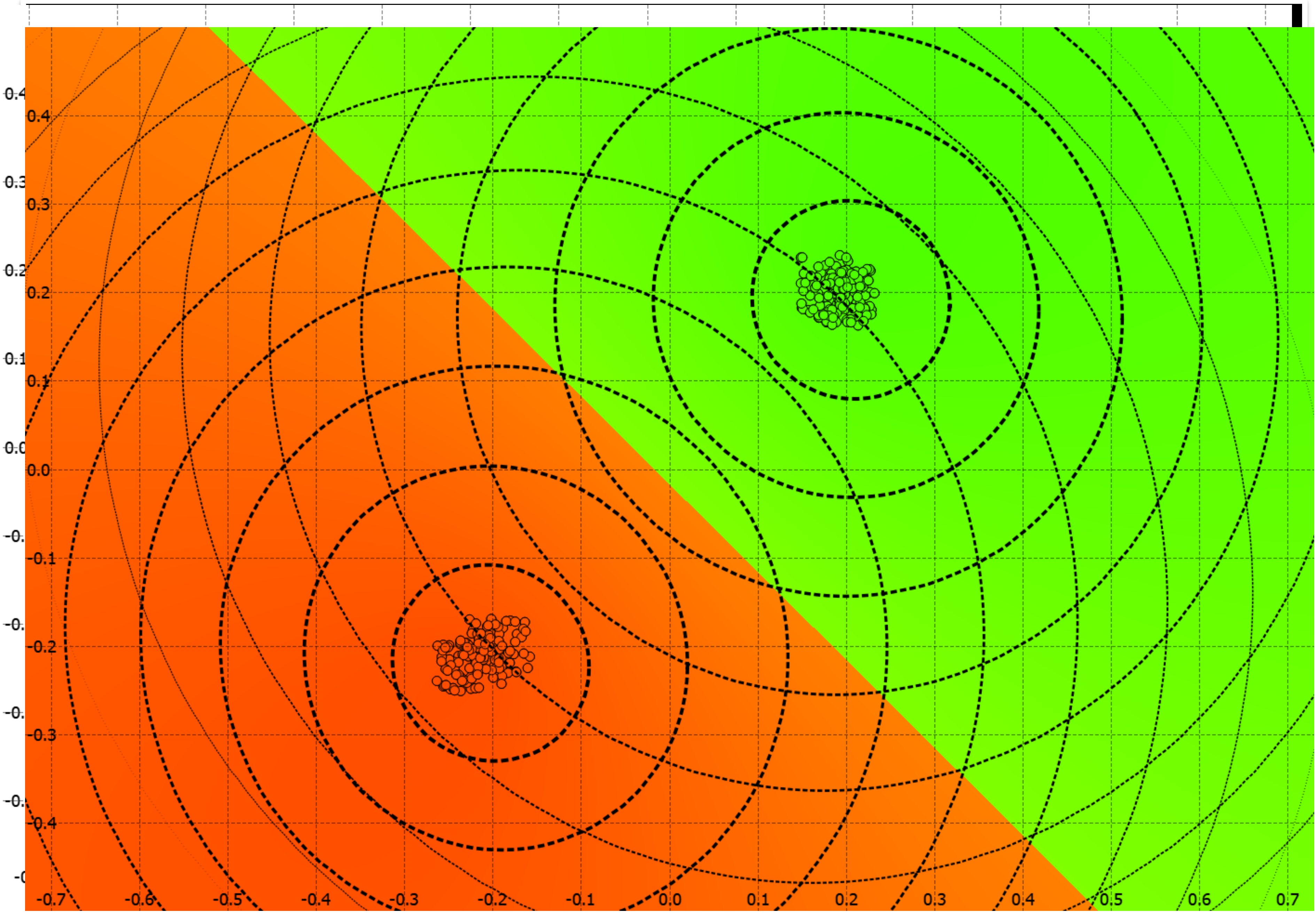
Kernel K-means: interpreting the solution

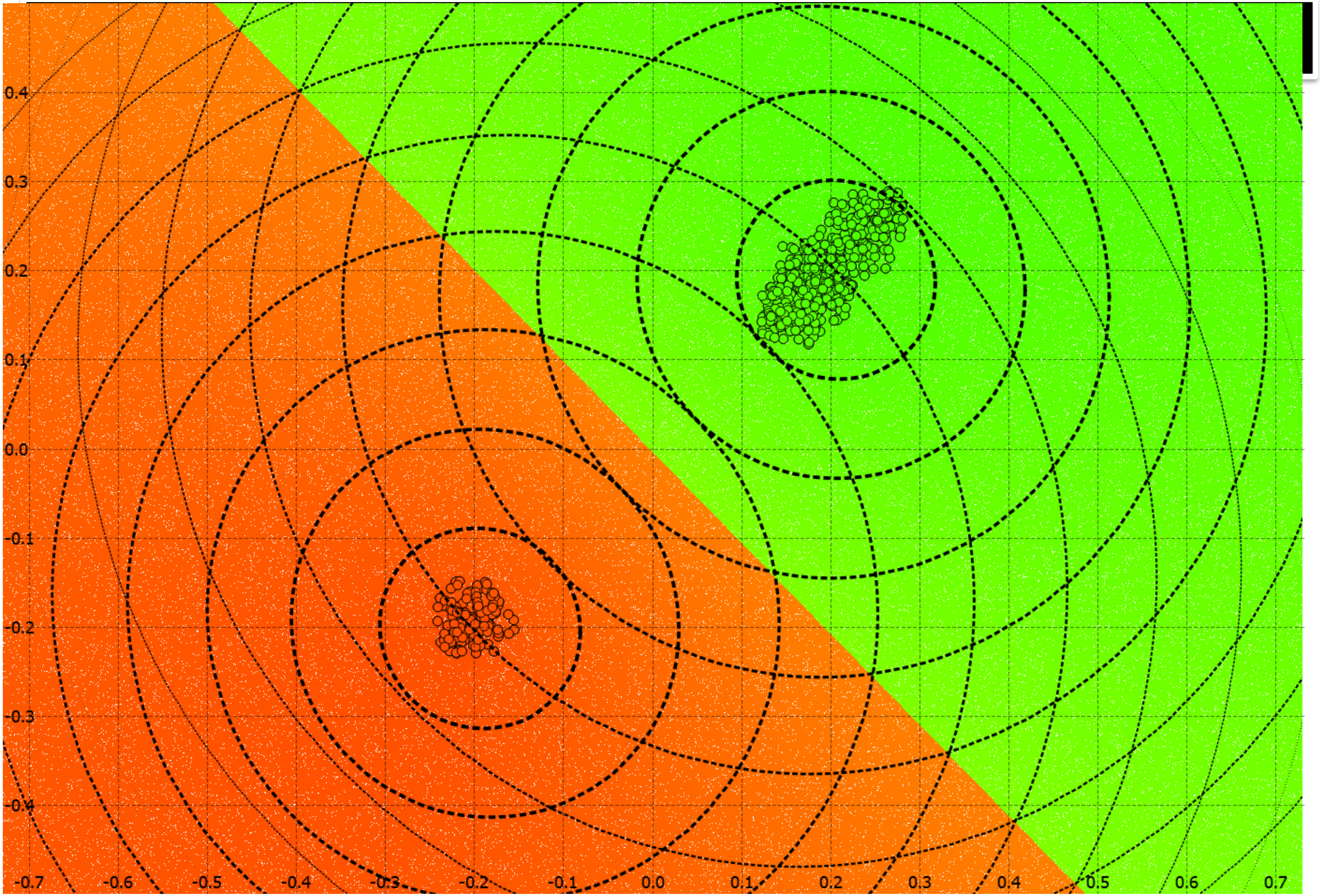


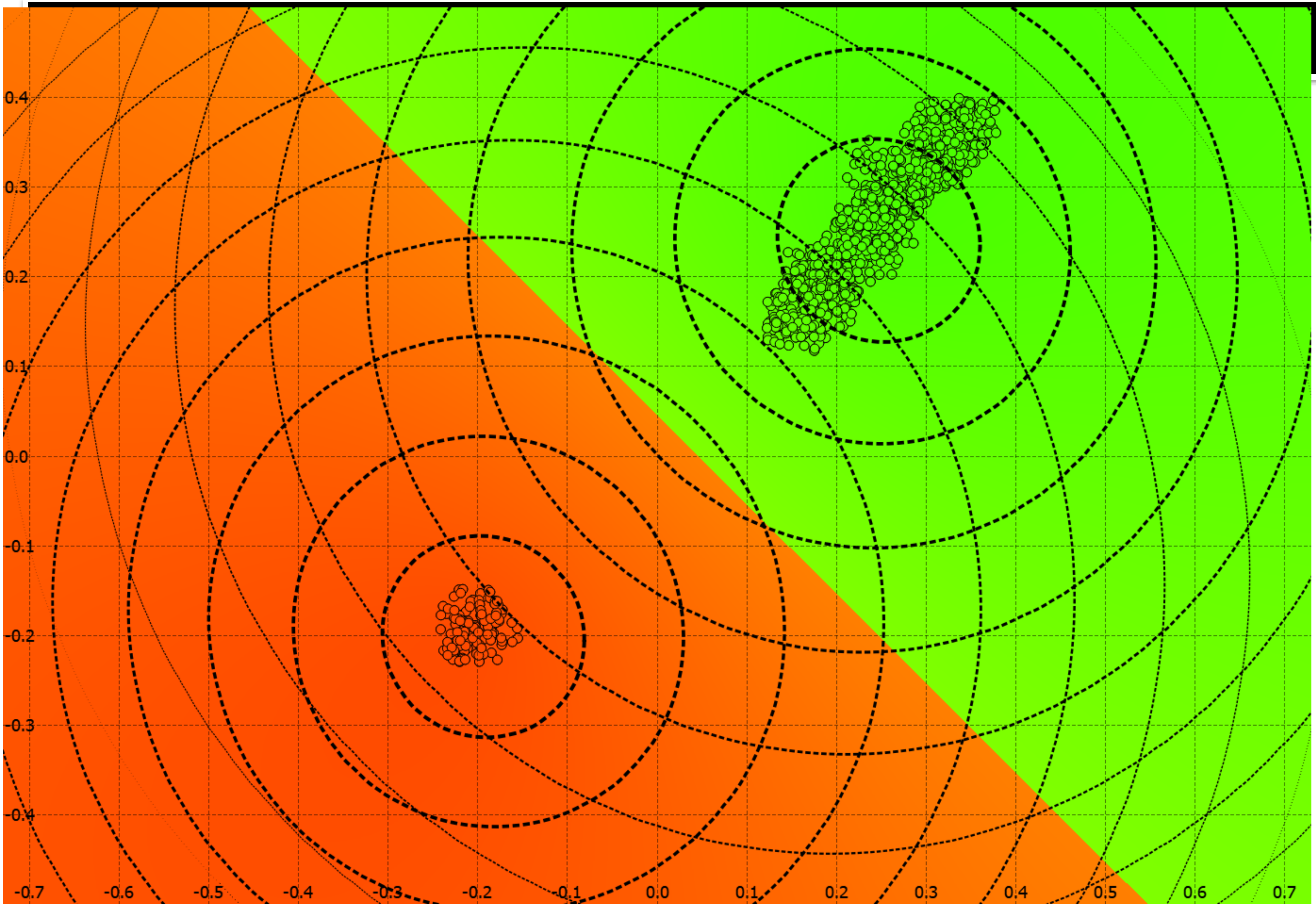
B: Affected by the relative angle across the points.

$$\arg\min_k d\left(x, C^k\right) = \min_k \left(k\left(x, x\right) - \frac{2 \sum_{x^j \in C^k} k\left(x, x^j\right)}{m_k} + \frac{\sum_{x^j, x^l \in C^k} k\left(x^j, x^l\right)}{\left(m_k\right)^2} \right)$$

The spread of the cluster affects both angle and distance between points within same cluster. Imbalance in point distribution may be exacerbated.

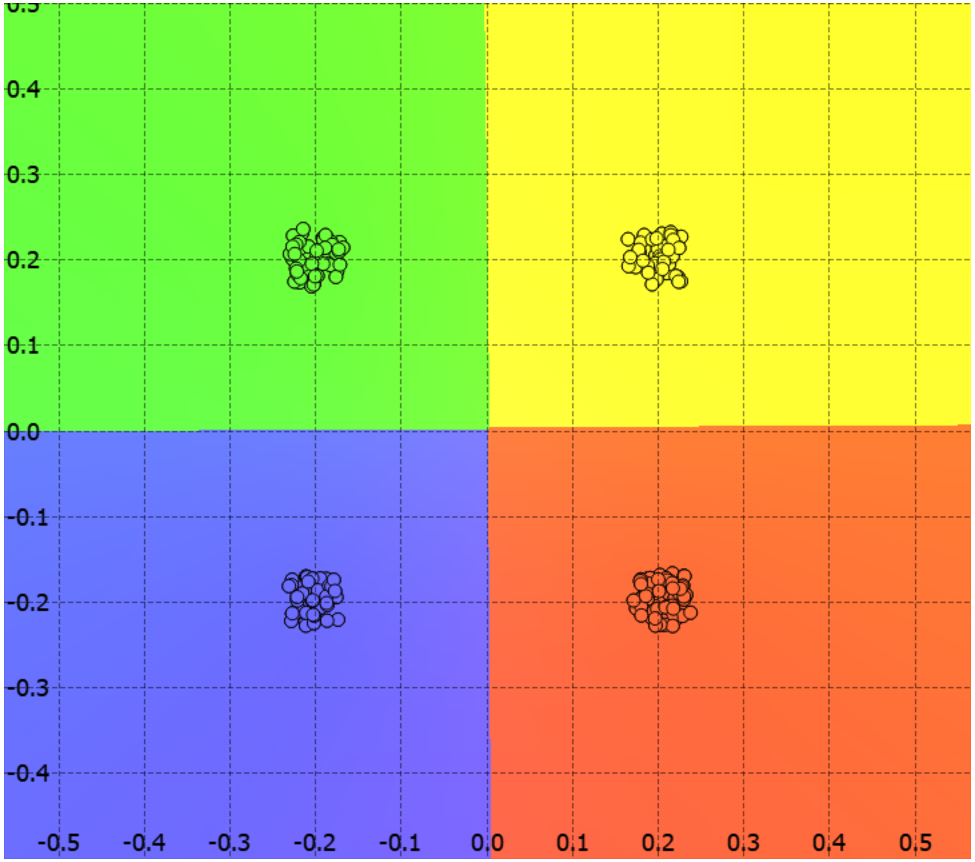

0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7


0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7


0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

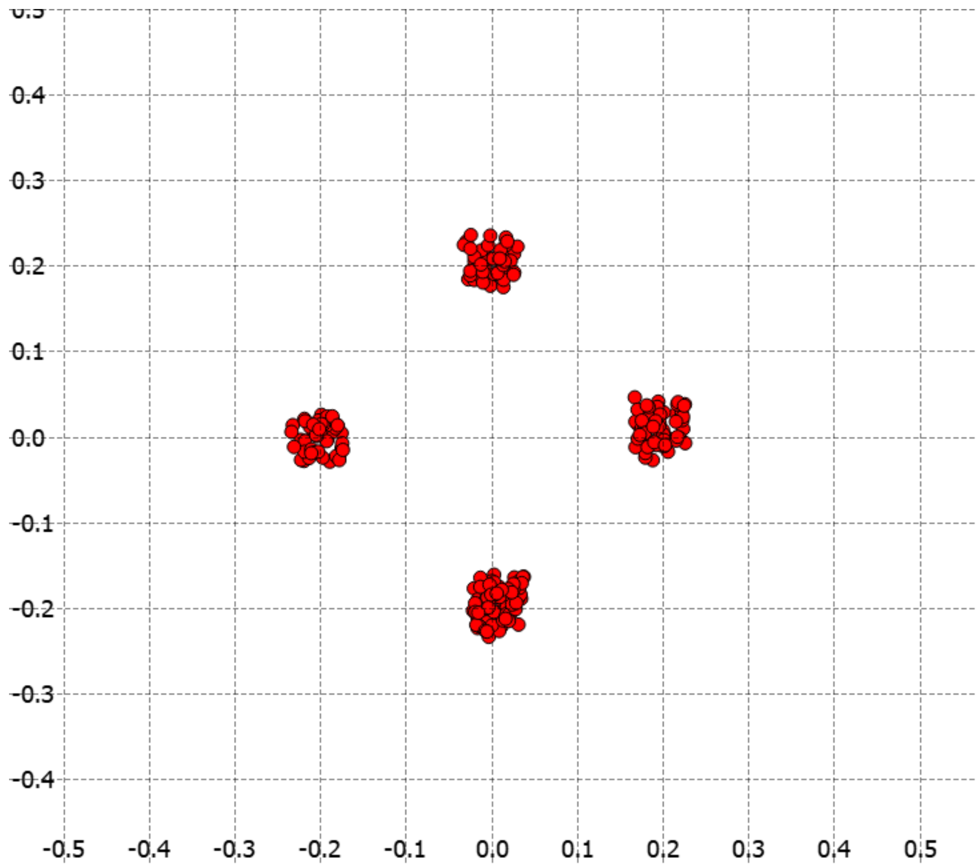
The spread of the cluster affects both angle and distance between points within same cluster. Imbalance in point distribution may be exacerbated.

Quadran partitioning



Partitioning with K=4 and homogeneous polynomial with p=1.

Will the partitioning be correct in this case too?



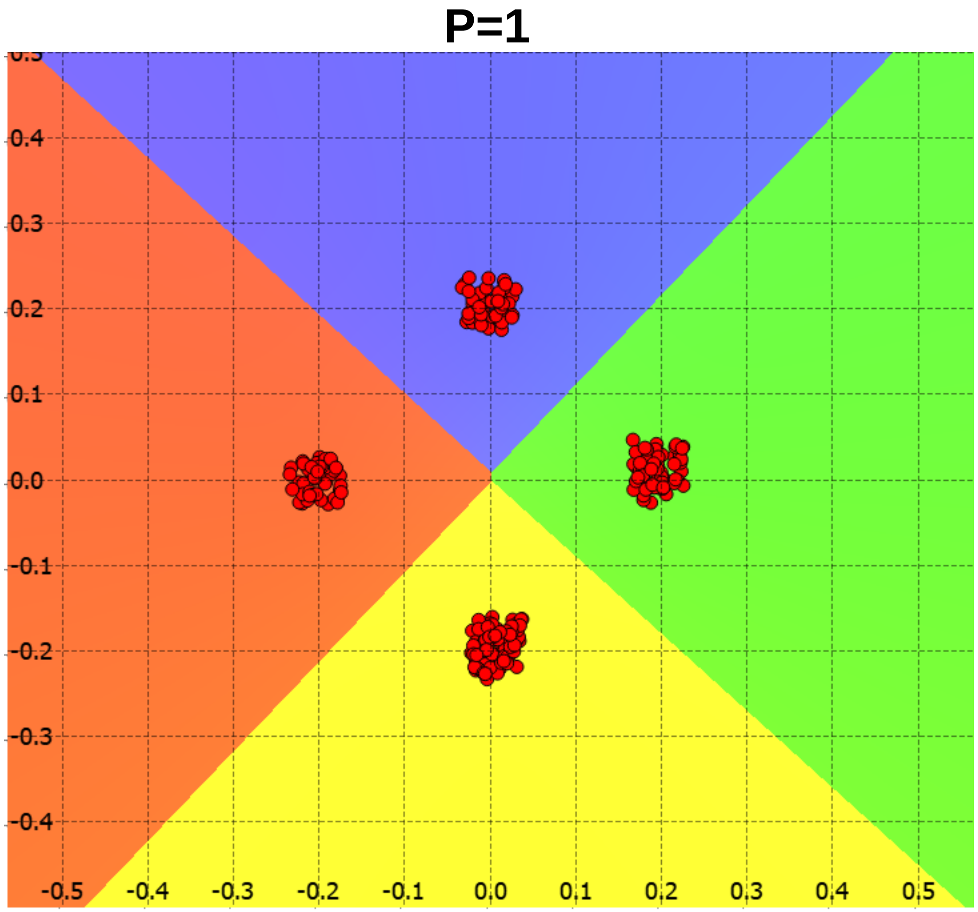
A. Yes

B. No

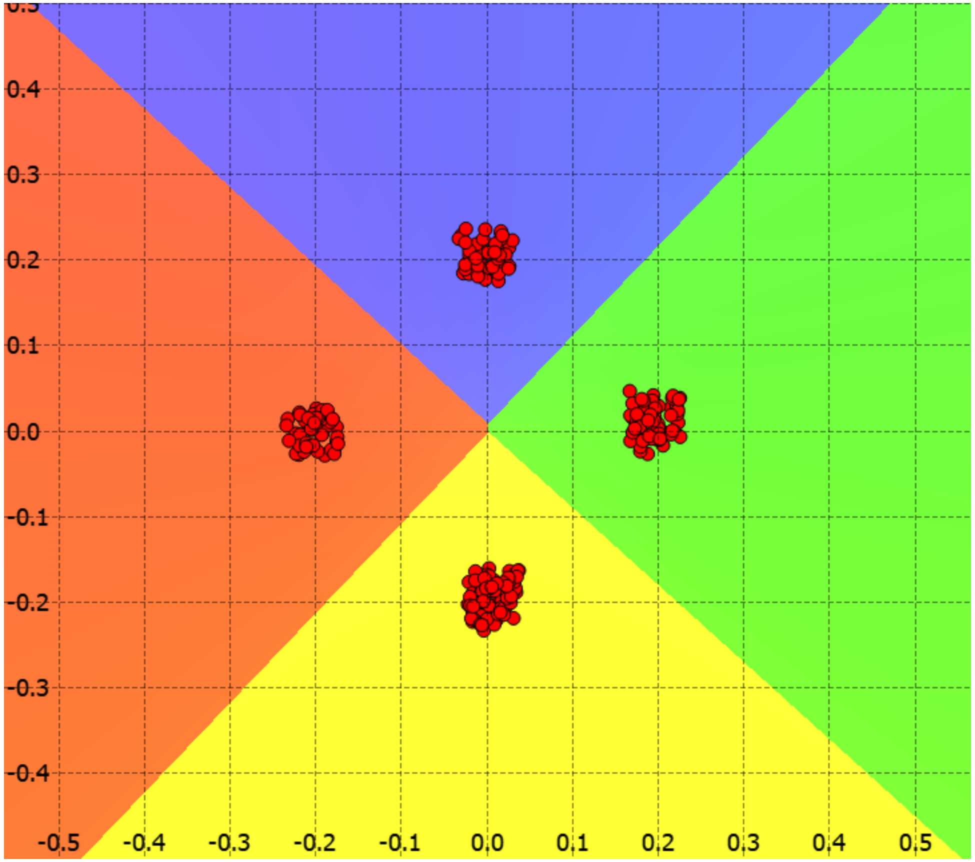
C. I do not know

Partitioning with K=4 and homogeneous polynomial with p=1.

Quadran partitioning



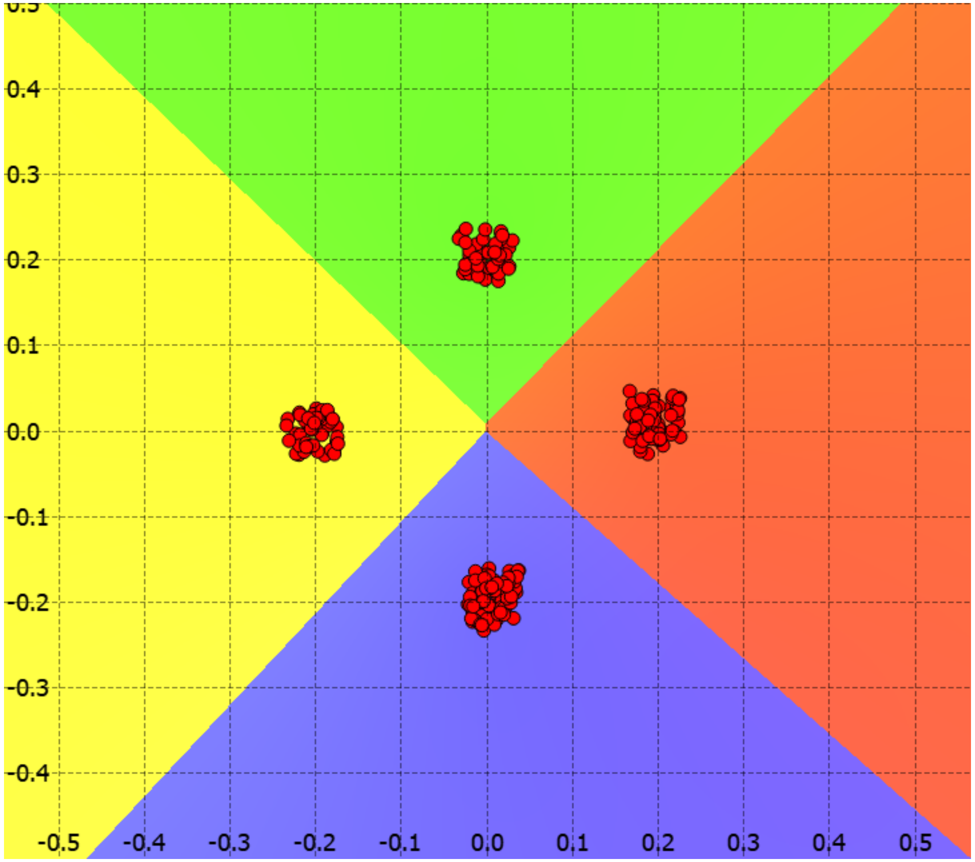
Does the boundary depend on the power of the polynomial p?



A. Yes
B. No
C. I do not know

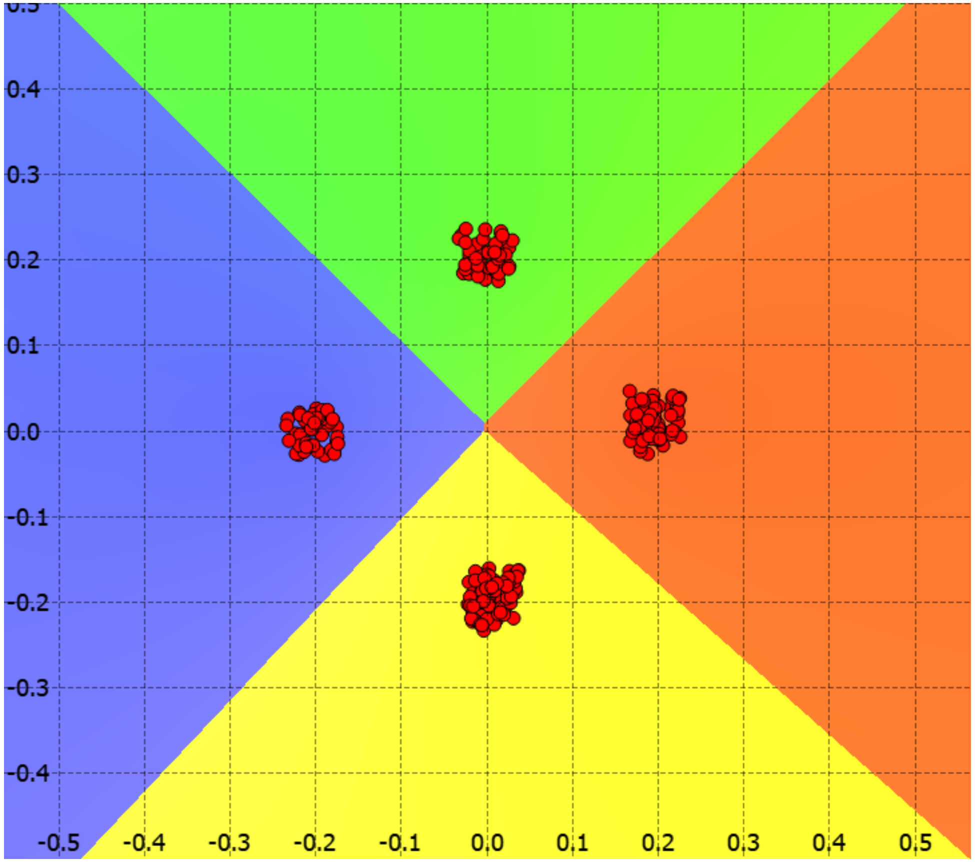
Quadran partitioning

P=2

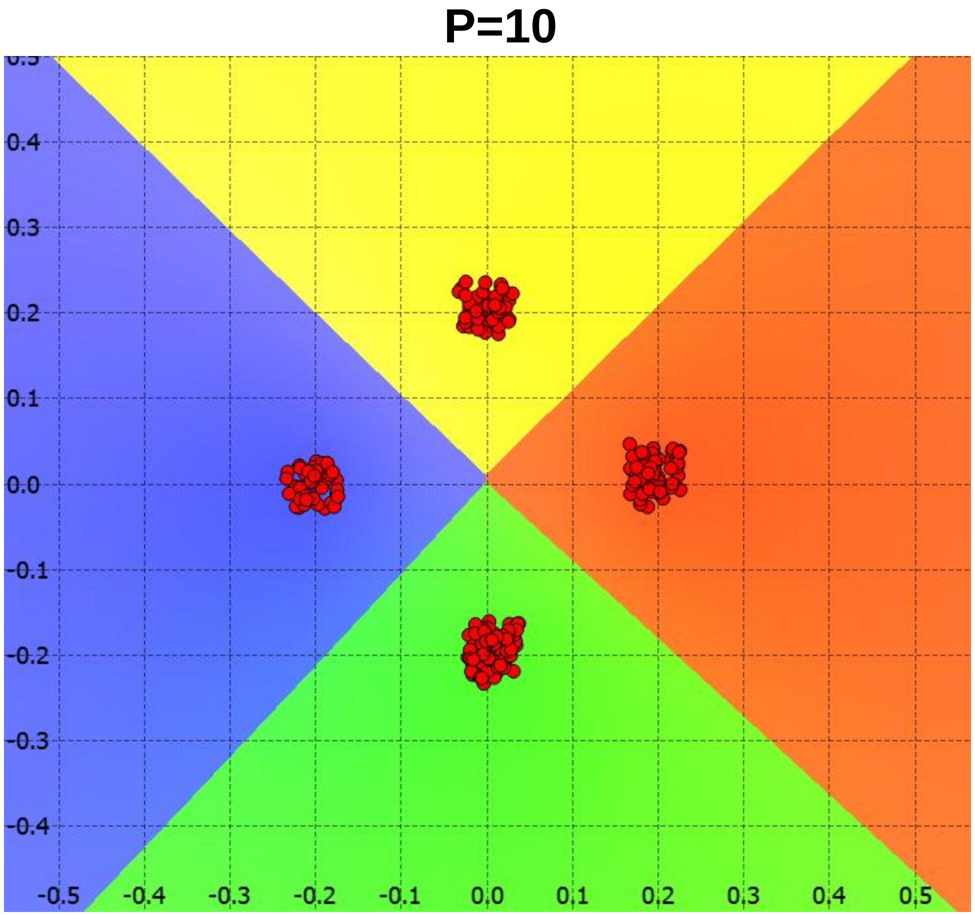


Quadran partitioning

P=3

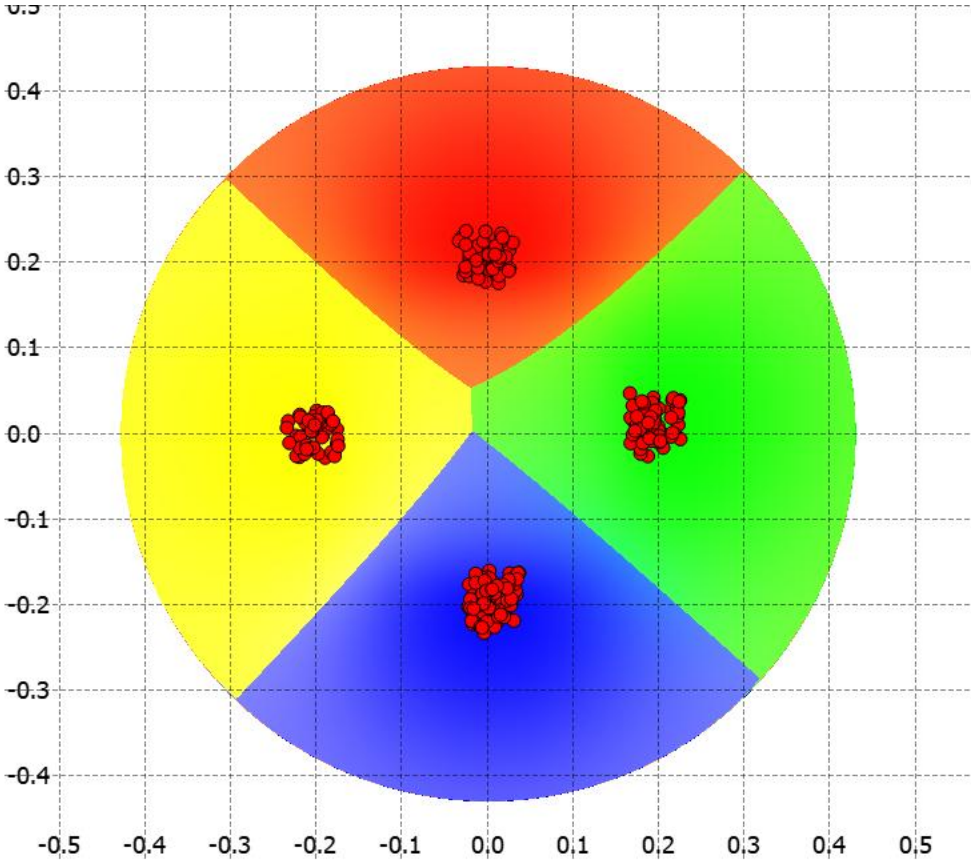


Quadran partitioning

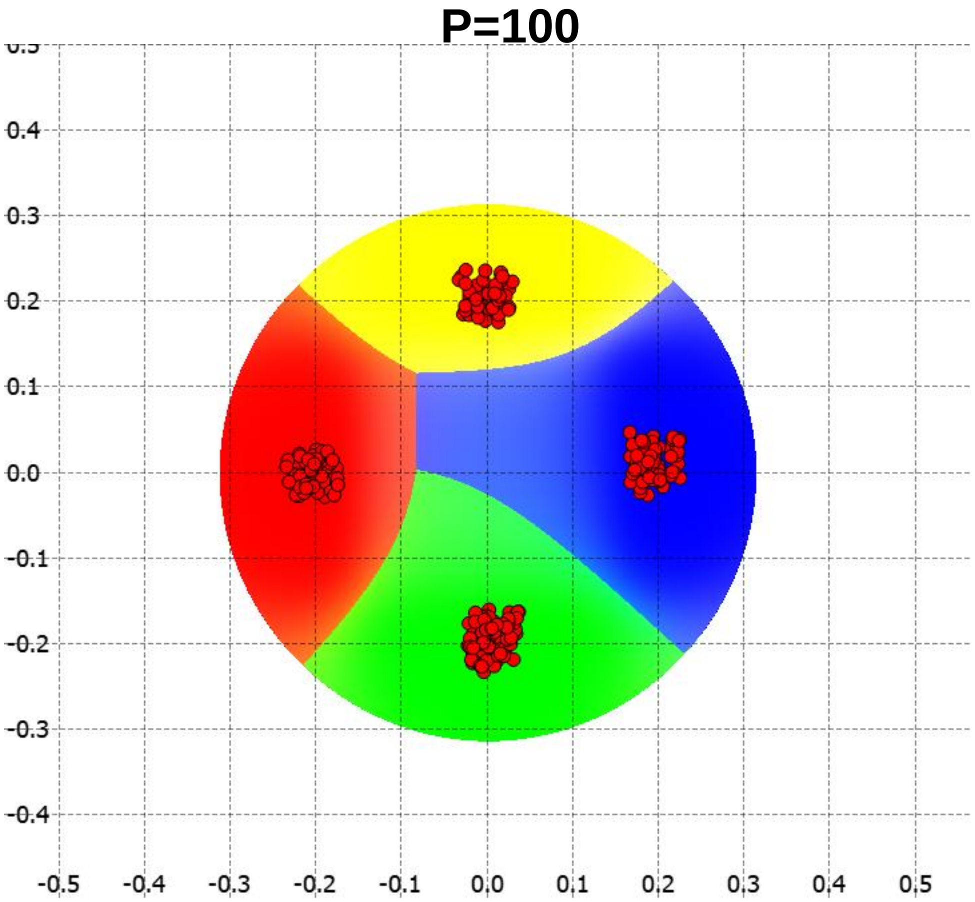


Quadran partitioning

P=55

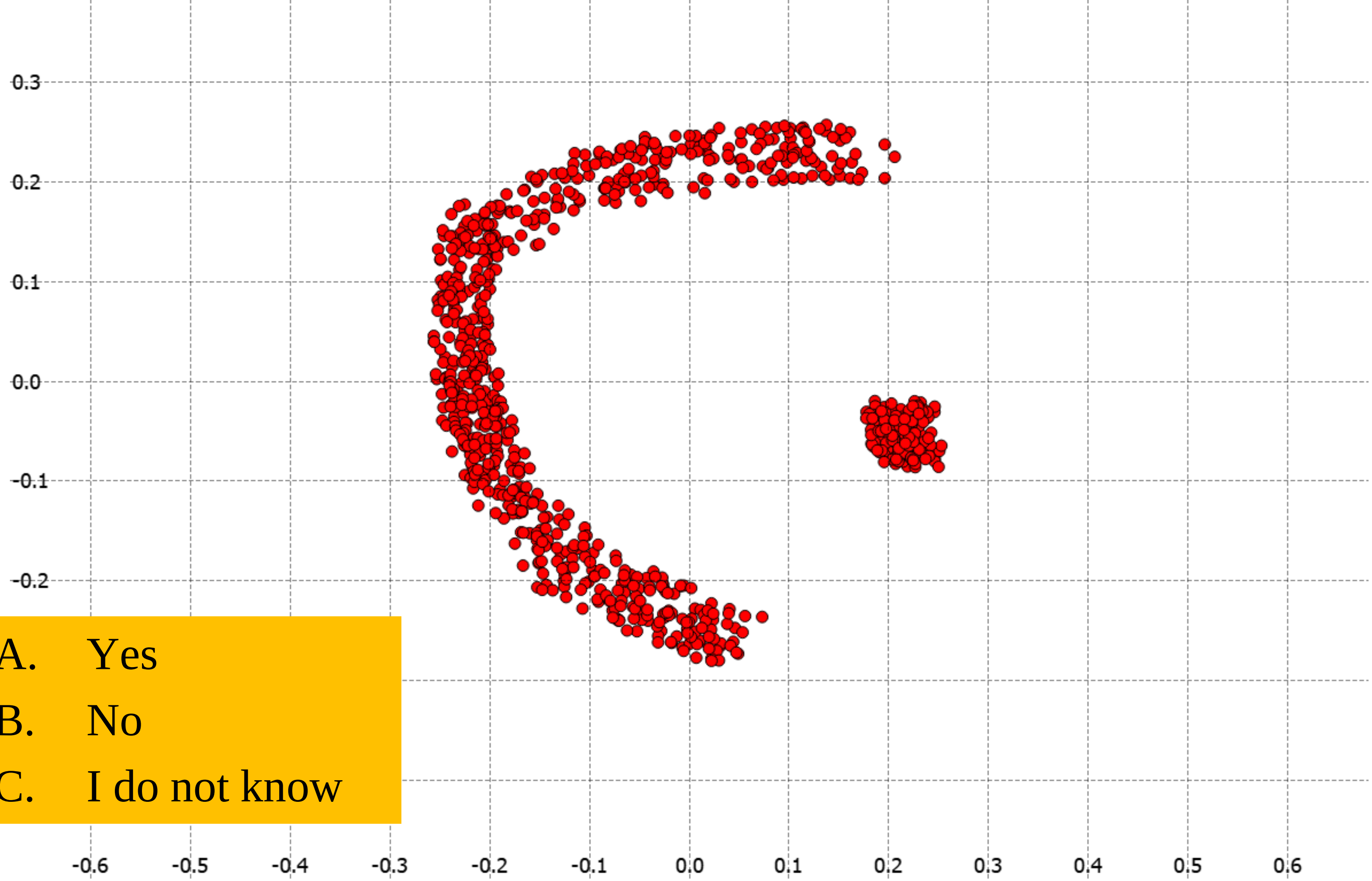


Quadran partitioning



Type of partitioning

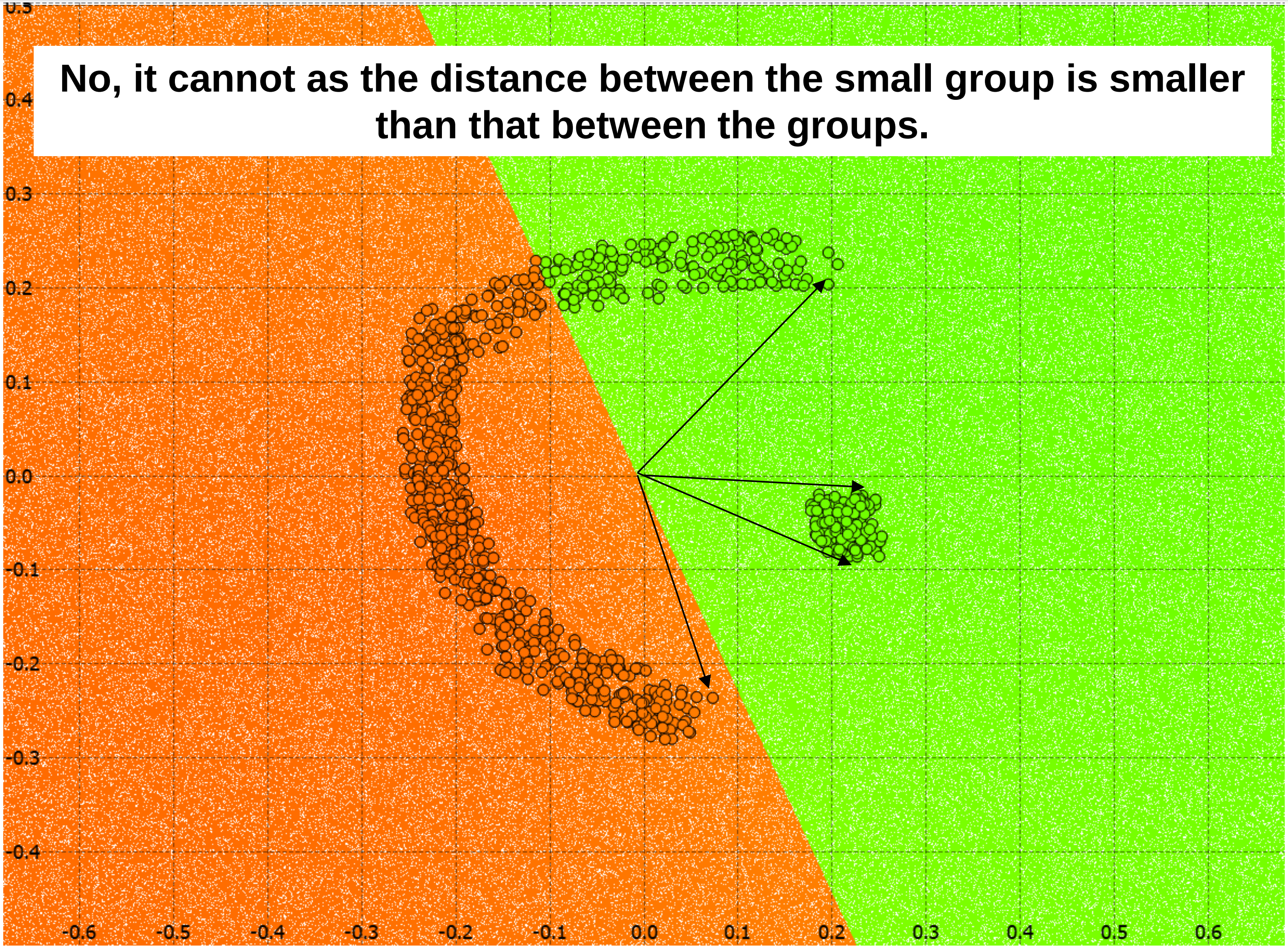
Can homogeneous polynomial kernel with p=1 separate the 2 groups?



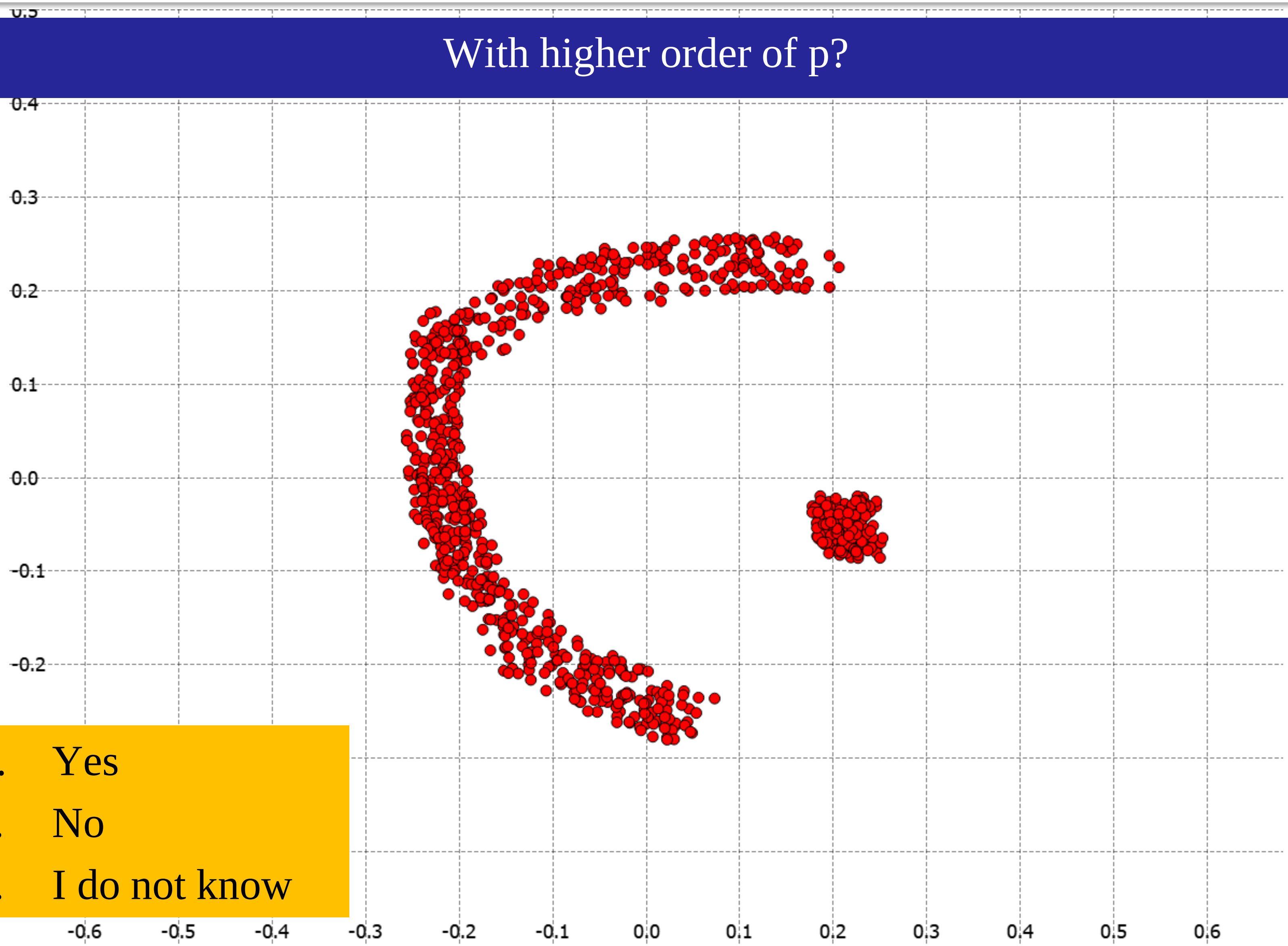
A. Yes
B. No
C. I do not know

Type of partitioning

No, it cannot as the distance between the small group is smaller than that between the groups.



Type of partitioning

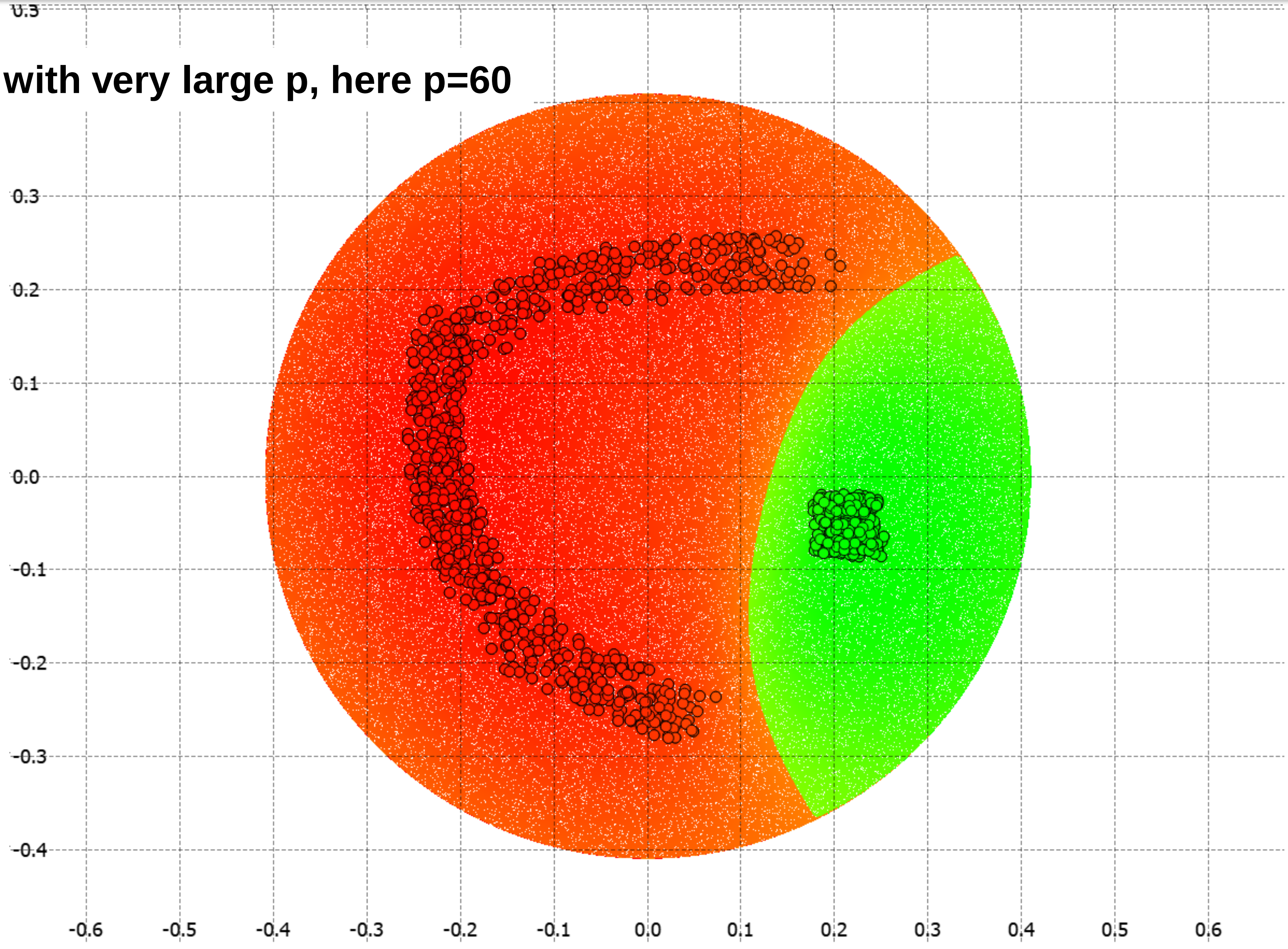


With higher order of p?

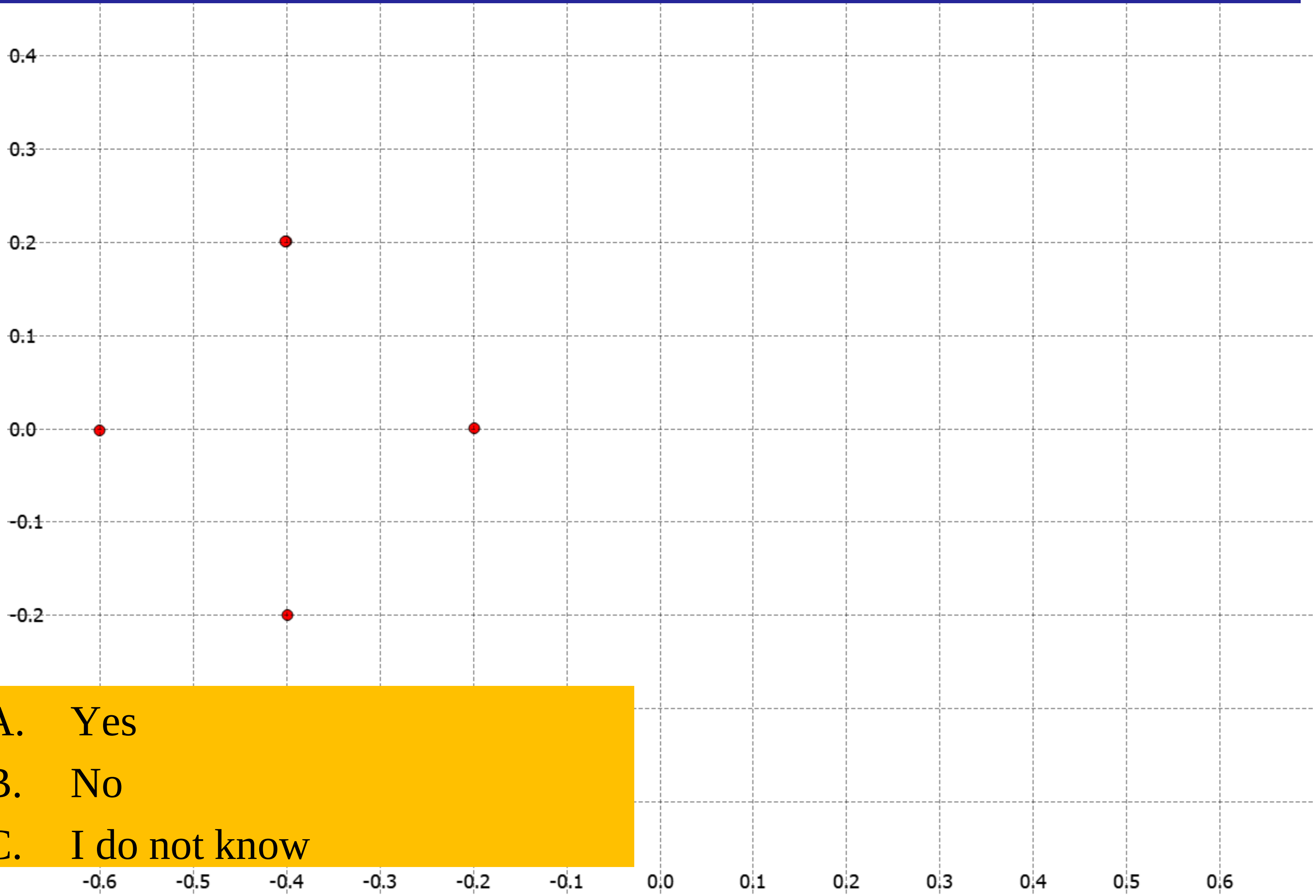
A. Yes
B. No
C. I do not know

Type of partitioning

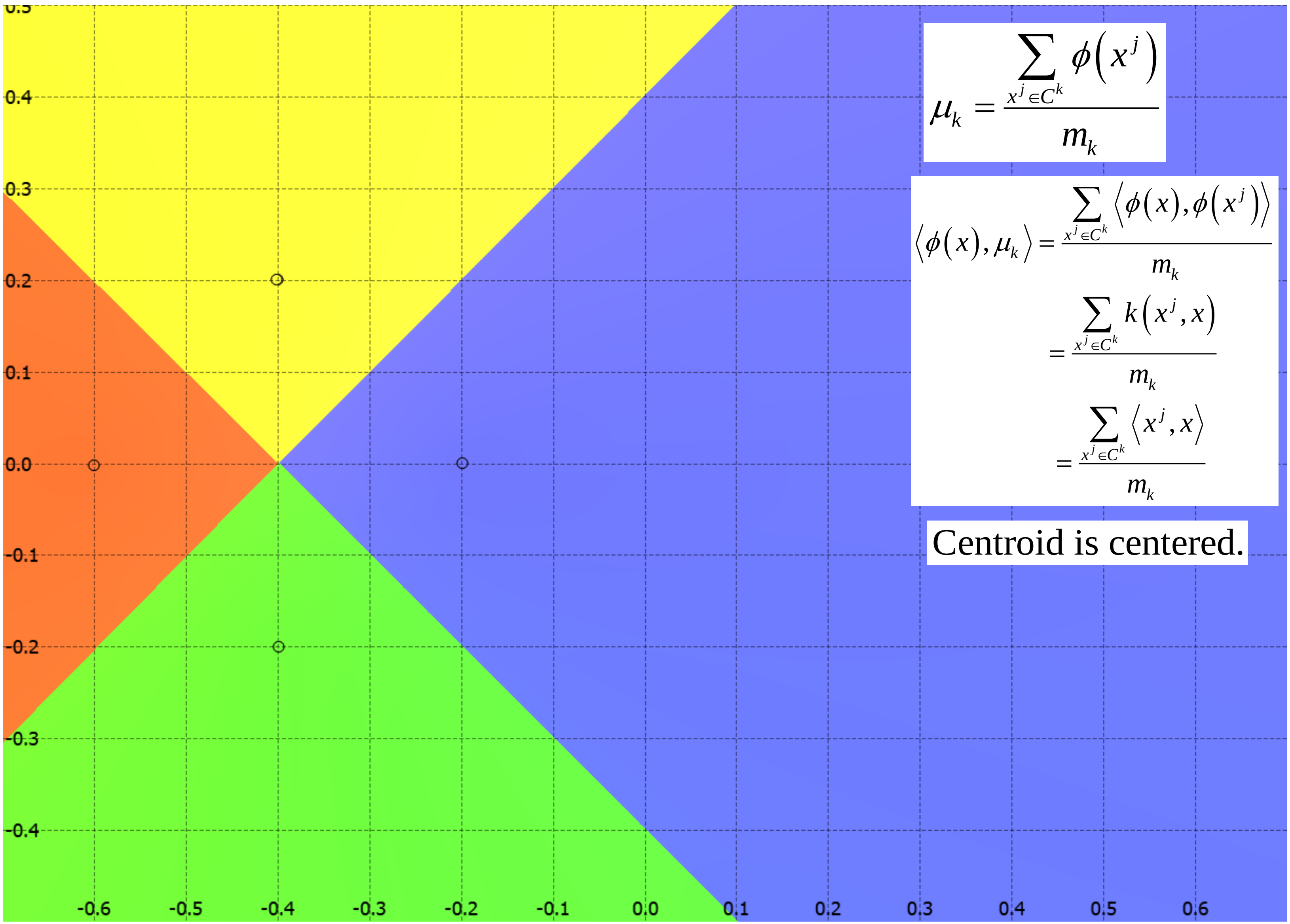
Yes with very large p, here p=60



Consider this group of points not centered, if you use K=4, homogeneous polynomial kernel p=1, will you get correct partitioning?



A. Yes
B. No
C. I do not know

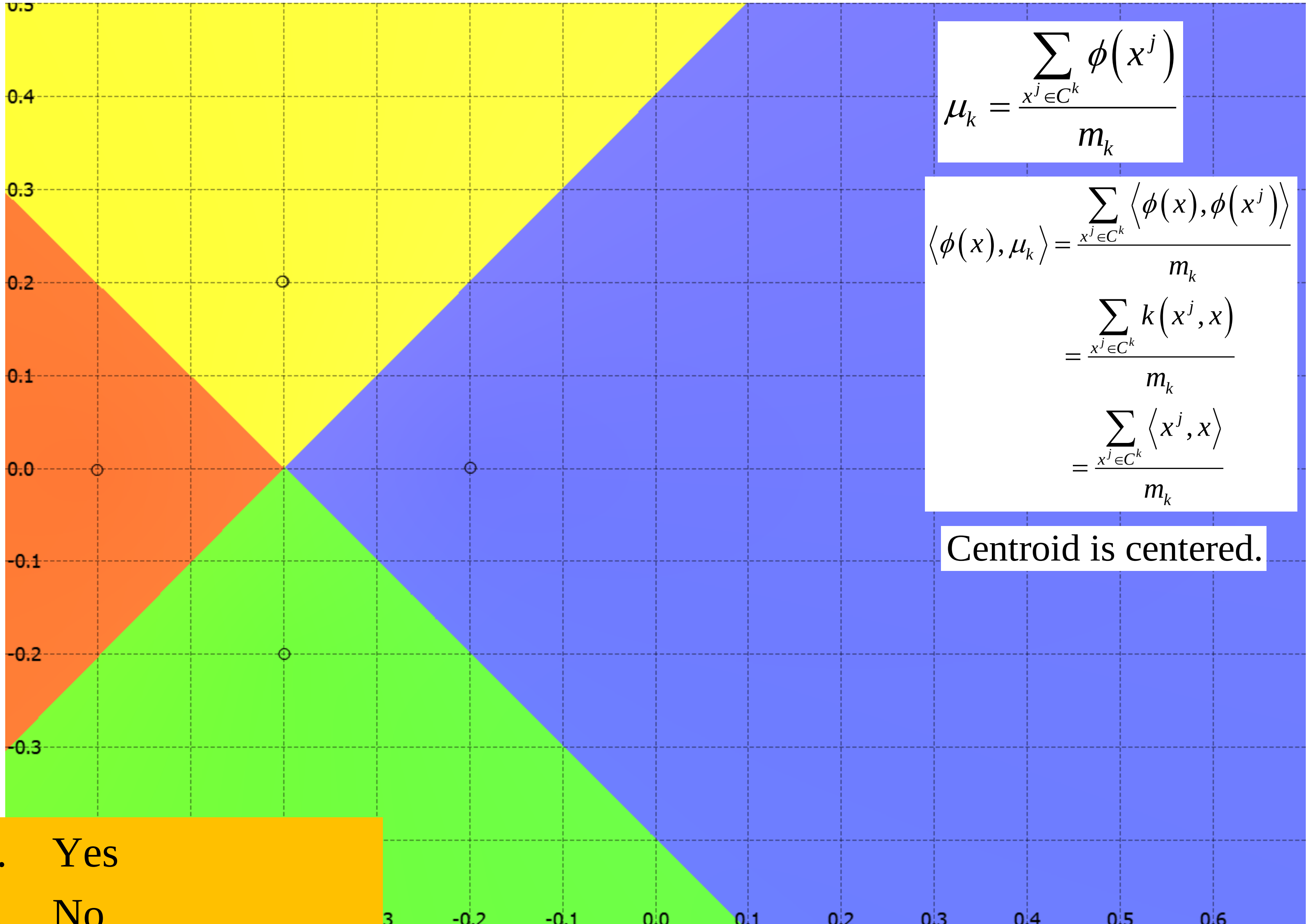


$$\mu_k = \frac{\sum_{x^j \in C^k} \phi\left(x^j\right)}{m_k}$$

$$\left\langle \phi(x), \mu_k \right\rangle = \frac{\sum_{x^j \in C^k} \left\langle \phi(x), \phi\left(x^j\right) \right\rangle}{m_k} = \frac{\sum_{x^j \in C^k} k\left(x^j, x\right)}{m_k} = \frac{\sum_{x^j \in C^k} \left\langle x^j, x \right\rangle}{m_k}$$

Centroid is centered.

Would the result change with p>1?

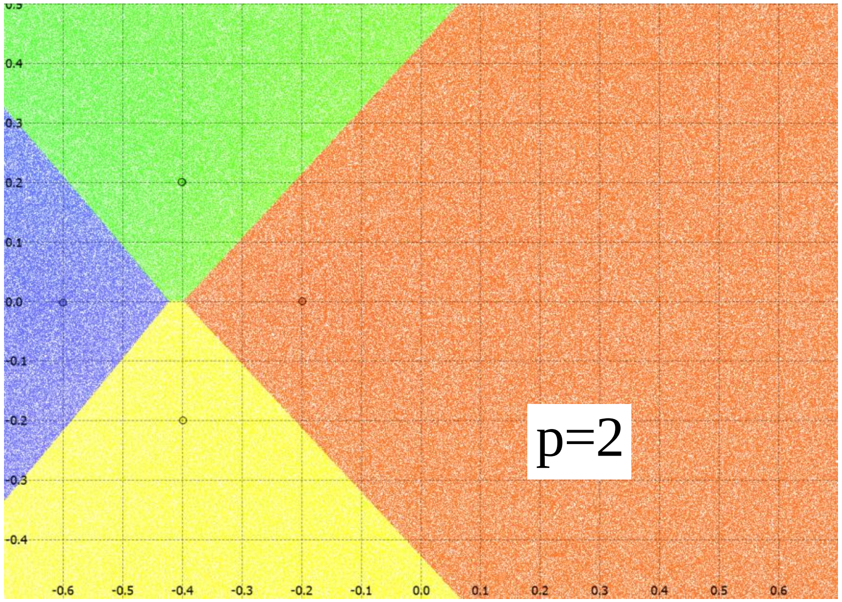


$$\mu_k = \frac{\sum_{x^j \in C^k} \phi\left(x^j\right)}{m_k}$$

$$\left\langle \phi(x), \mu_k \right\rangle = \frac{\sum_{x^j \in C^k} \left\langle \phi(x), \phi\left(x^j\right)\right\rangle}{m_k} = \frac{\sum_{x^j \in C^k} k\left(x^j, x\right)}{m_k} = \frac{\sum_{x^j \in C^k} \left\langle x^j, x\right\rangle}{m_k}$$

Centroid is centered.

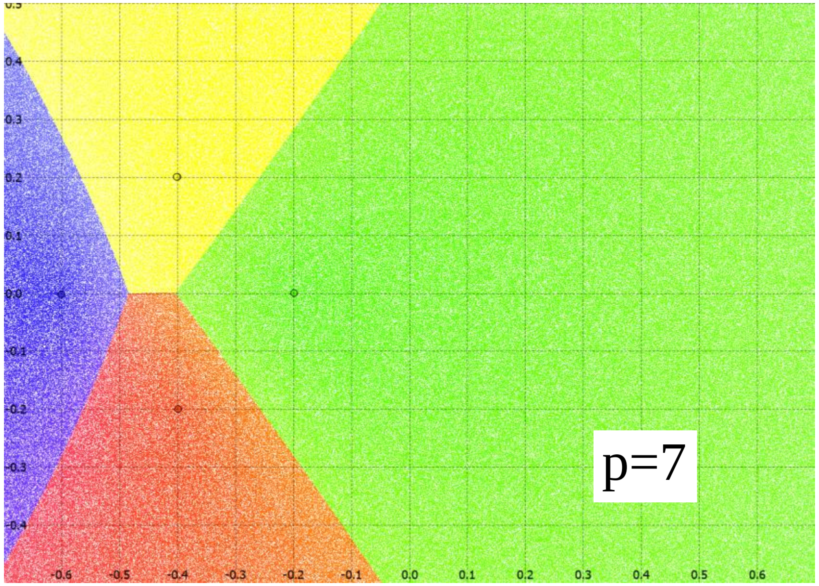
A. Yes
B. No
C. I do not know

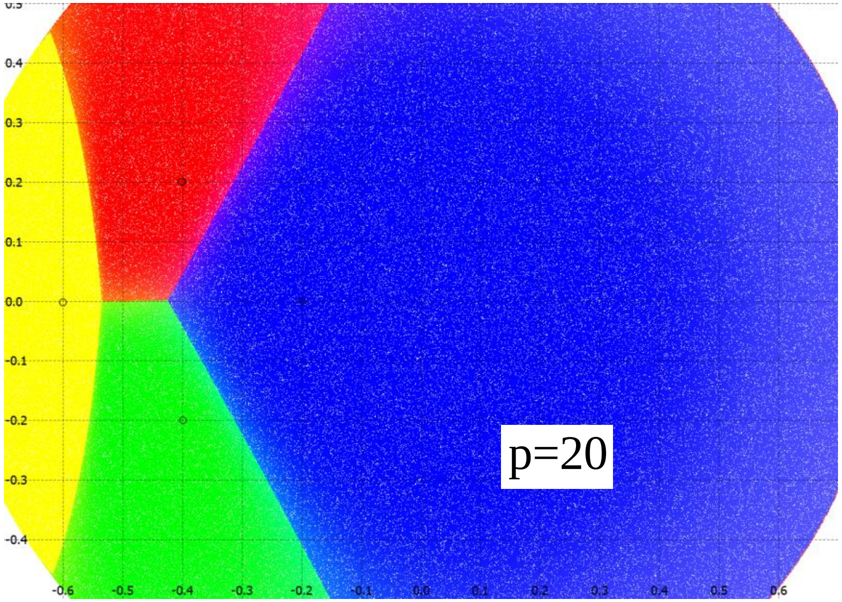


$$\langle \phi(x), \mu_k \rangle = \frac{\sum_{x^j \in C^k} \langle x^j, x \rangle^P}{m_k}$$

$$= \frac{\sum_{x^j \in C^k} \left\| x^j \right\|^P \left\| x \right\|^P \cos(\theta)^P}{m_k}$$

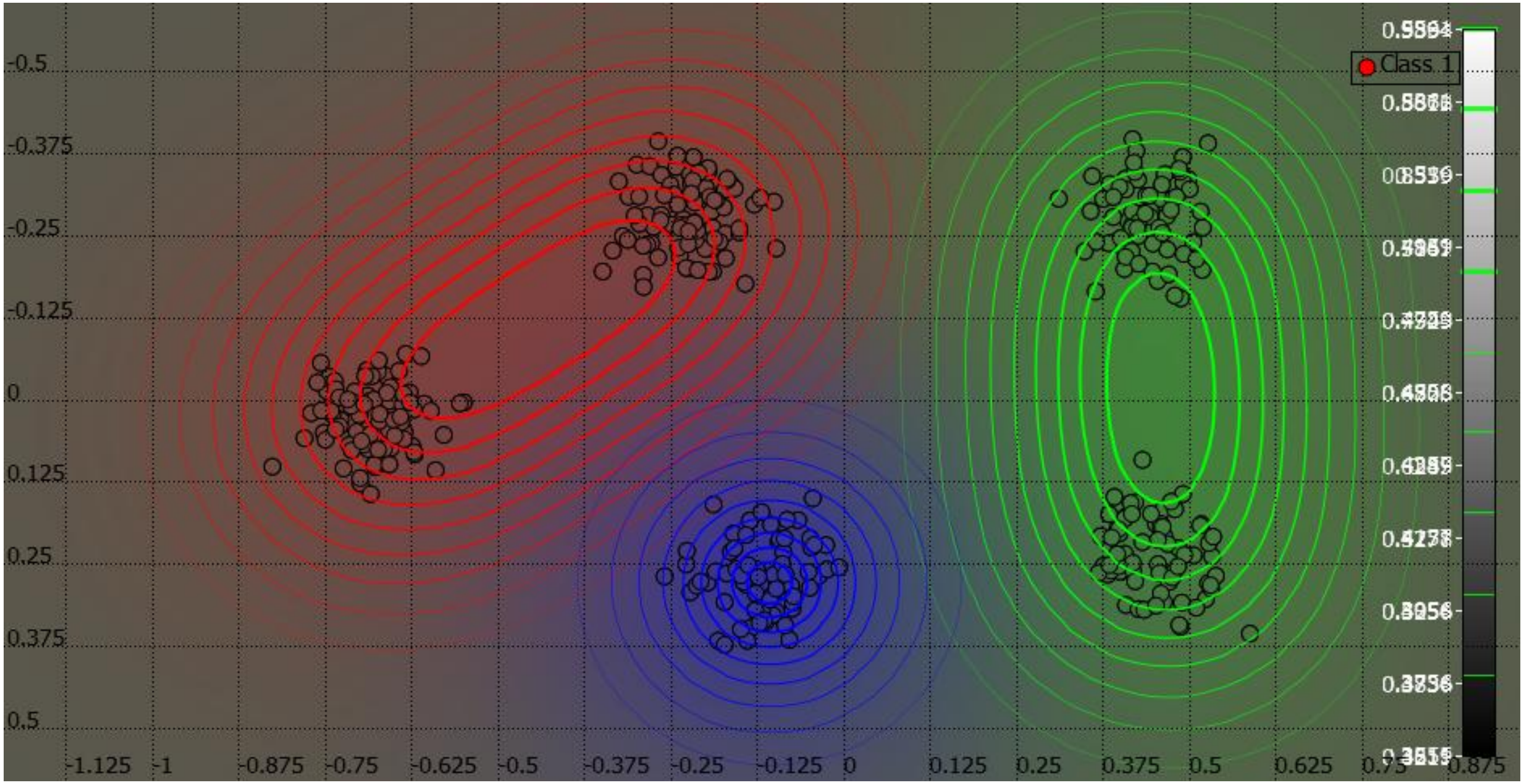
The higher *p*, the more curvy the boundaries.





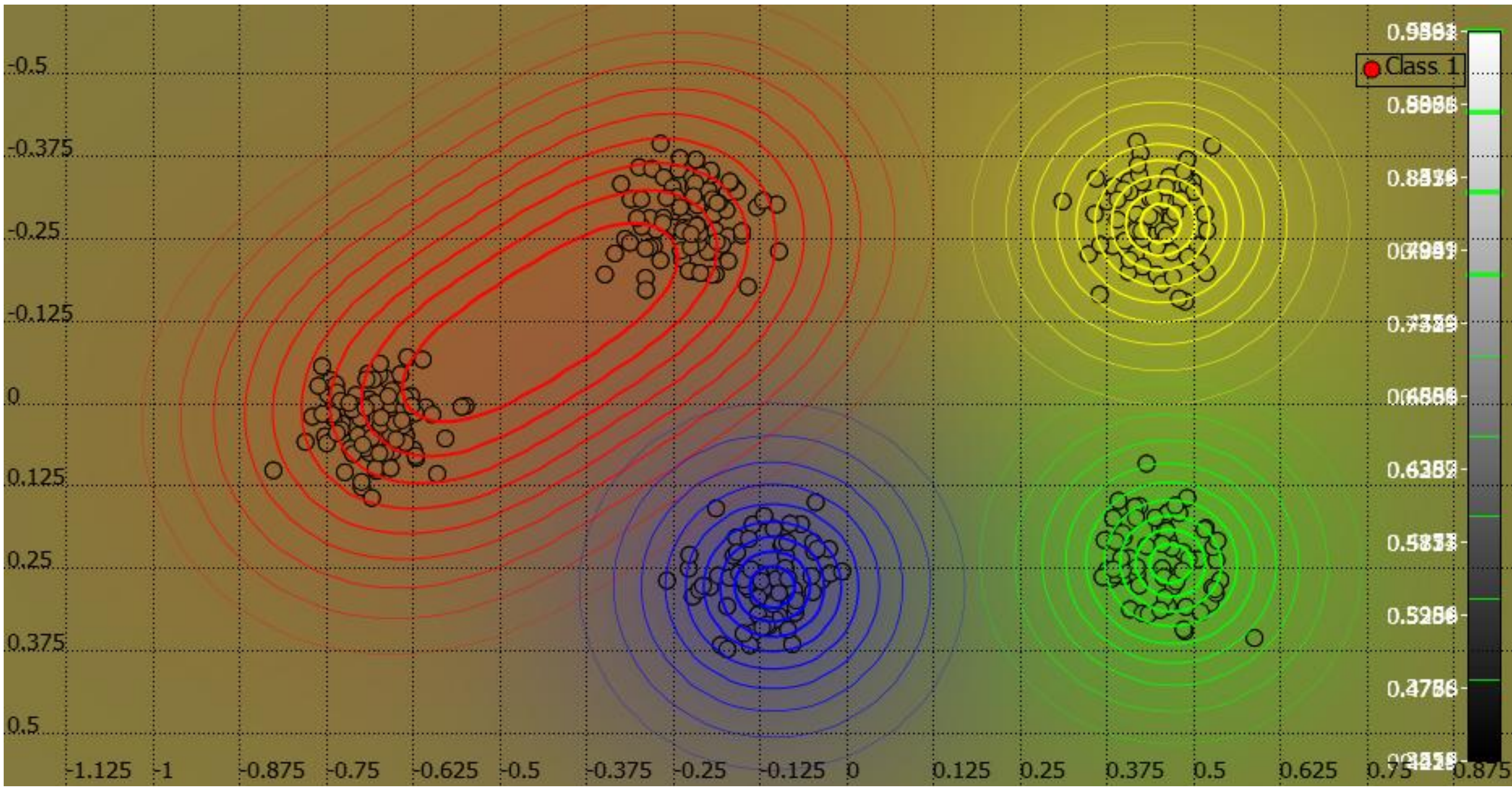
Kernel K-means: Limitations

Choice of number of Clusters in Kernel K-means is important



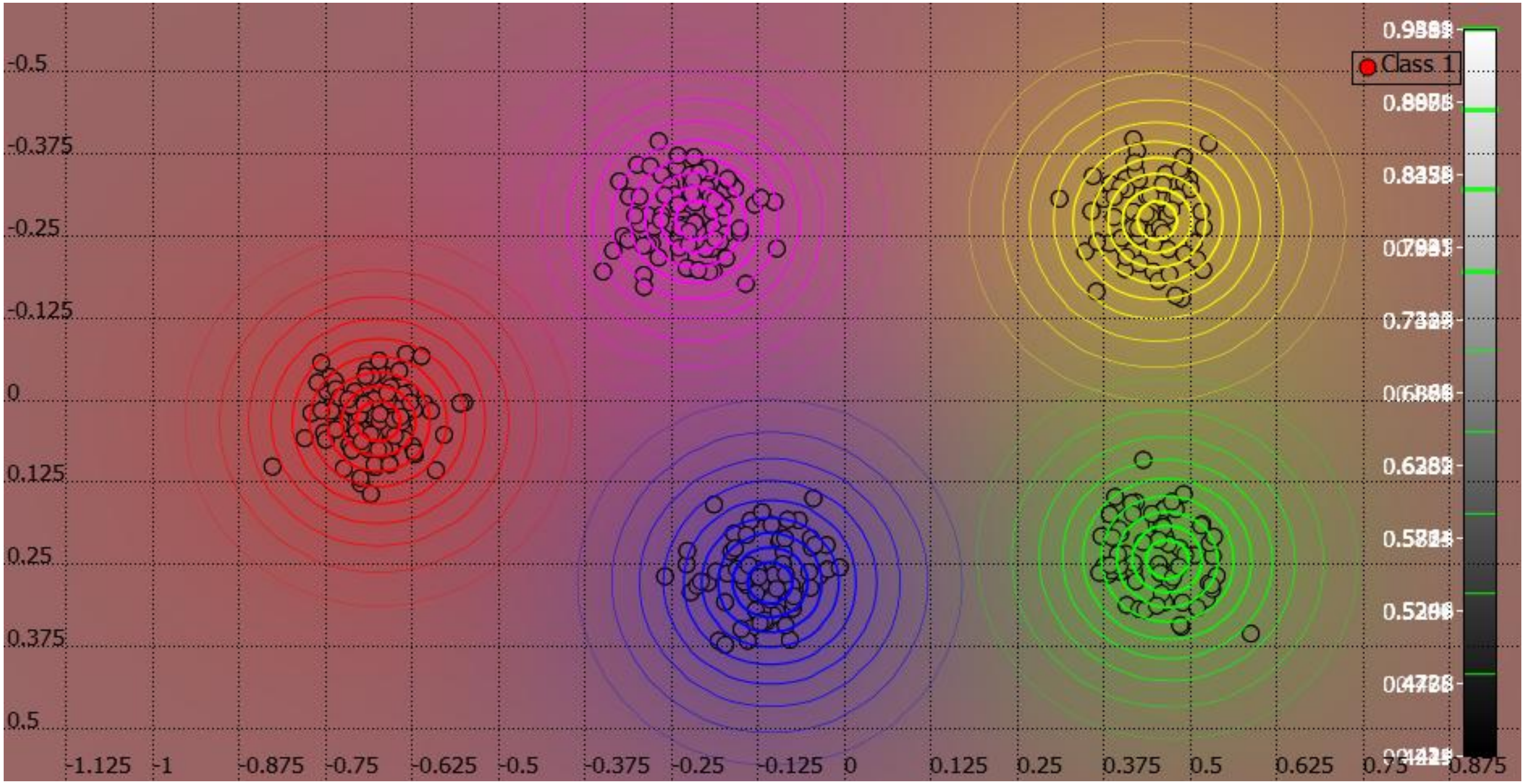
Kernel K-means: Limitations

Choice of number of Clusters in Kernel K-means is important

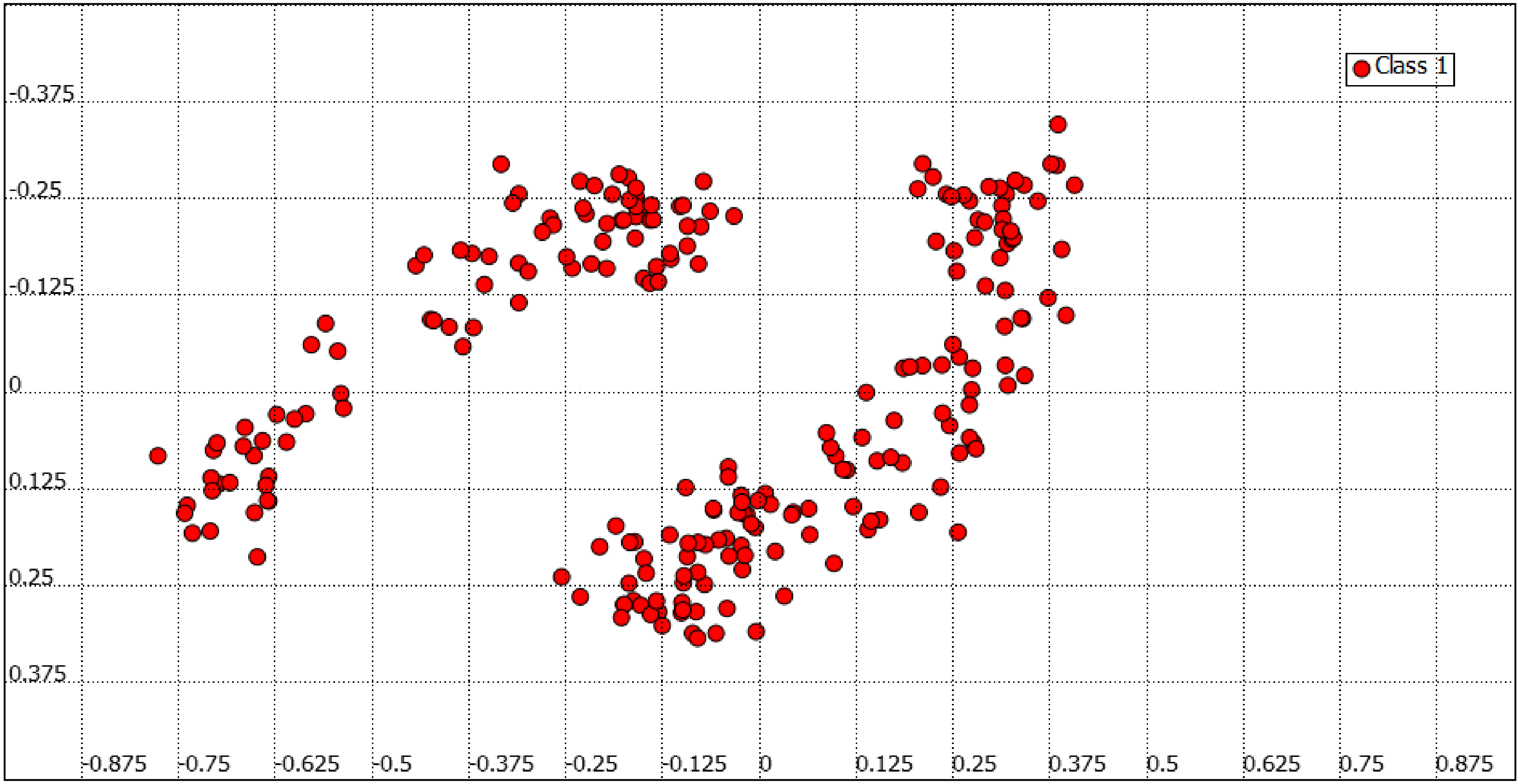


Kernel K-means: Limitations

Choice of number of Clusters in Kernel K-means is important

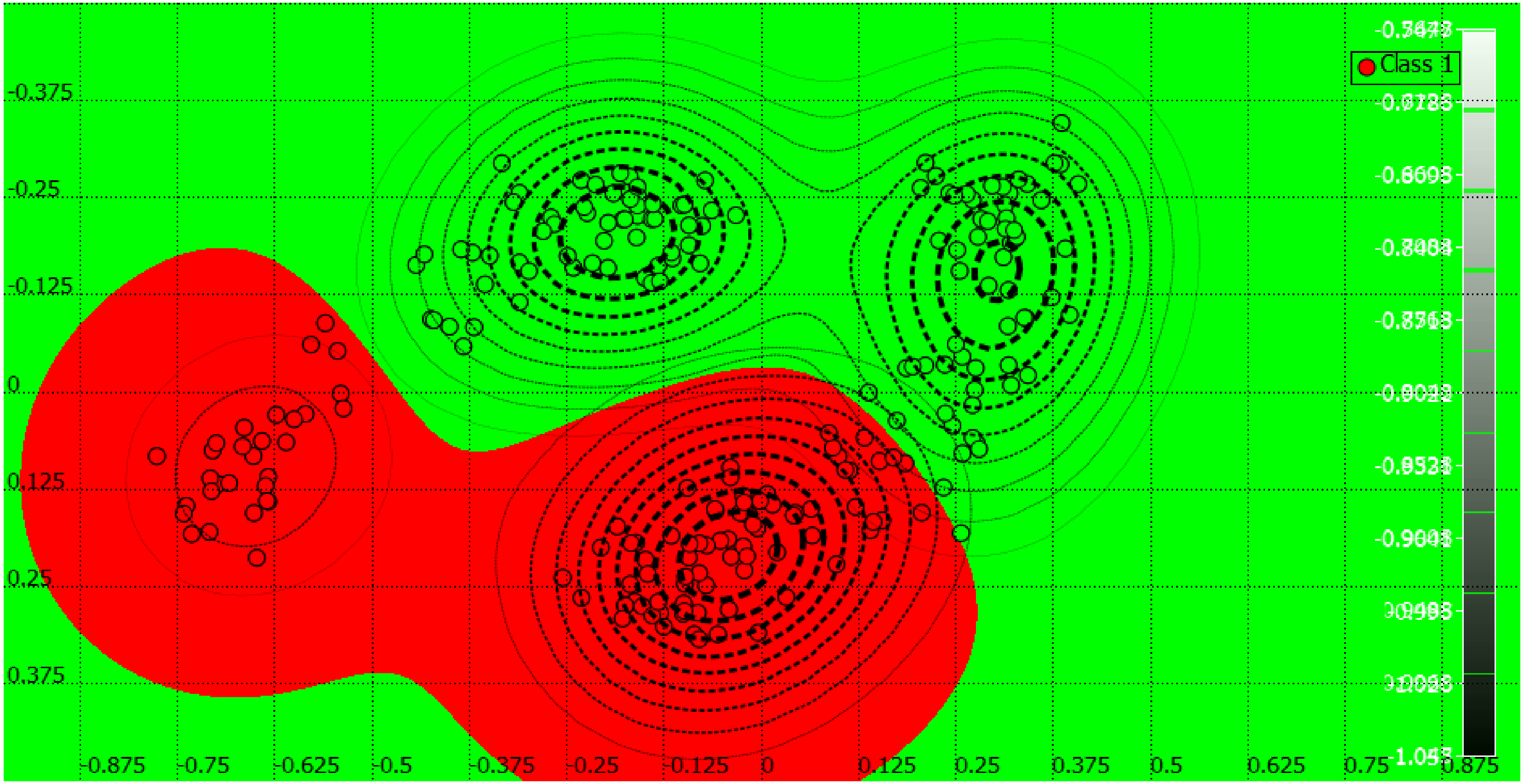


Limitations of kernel K-means



Raw Data

Limitations of kernel K-means



kernel K-means with K=2, RBF kernel