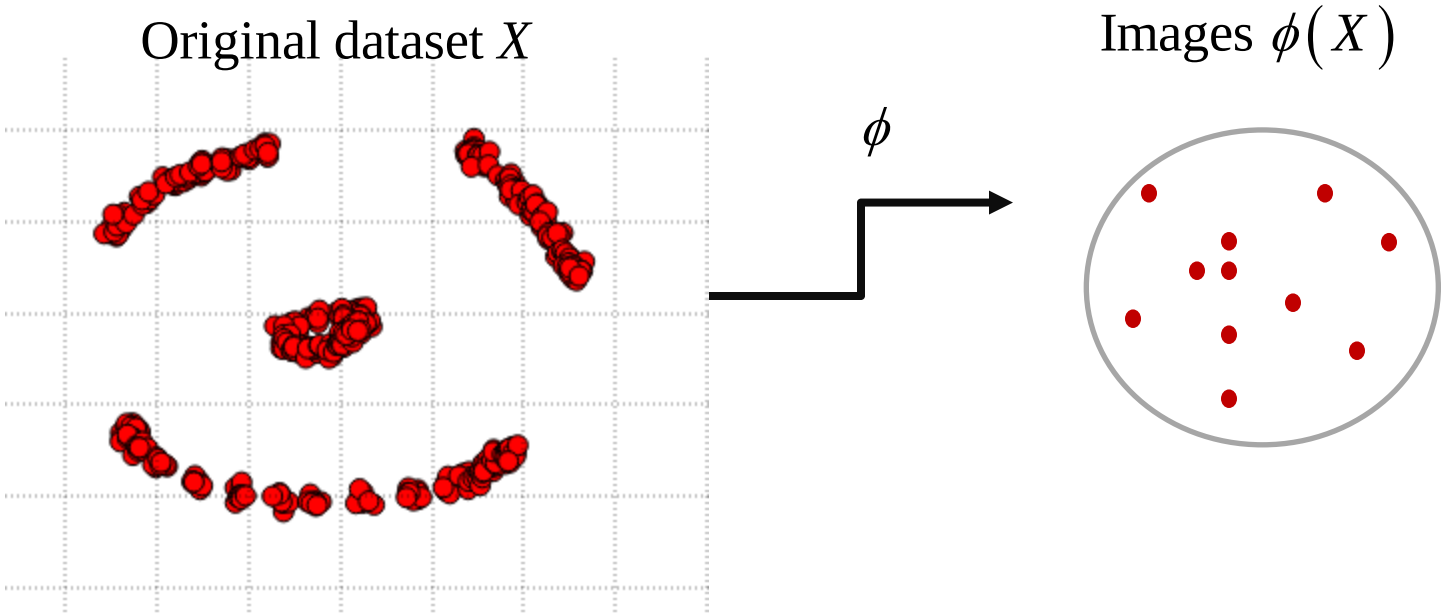
Support Vector Clustering (SVC)

See supplement (Ben-hur et al, IJML 2001)

Support Vector Clustering

Idea:

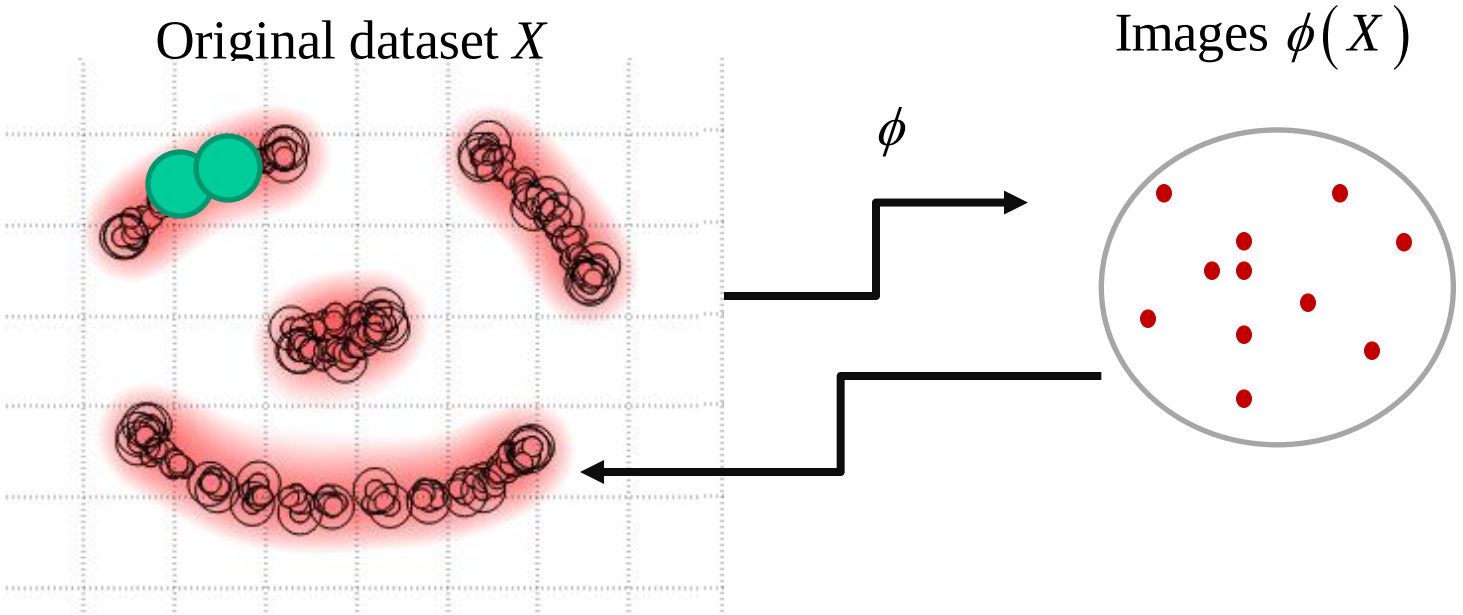
- Data points are mapped from data space to a high dimensional feature space using a Gaussian kernel (RBF kernel)
- In feature space, we look for the smallest sphere that encloses the image of the data.



Support Vector Clustering

Idea:

- Data points are mapped from data space to a high dimensional feature space using a Gaussian kernel (RBF kernel)
- In feature space, we look for the smallest sphere that encloses the image of the data.
- This sphere is mapped back to data space, where it forms a set of contours which enclose the data points. These contours are interpreted as cluster boundaries.



Support Vector Clustering

Requesting that all points x^j, $j = 1...M$, be in a sphere in feature space gives the following constraint:

$$\left\|\phi\left(x^j\right) - \mu\right\|^2 \leq \mathrm{R}^2, \quad \forall\, x^j, j = 1...M$$

μ: center of the sphere, R: radius of the sphere

To soften the constraint, one can add some slack variable ξ^i:

$$\left\|\phi\left(x^j\right) - \mu\right\|^2 \leq R^2 + \xi^i, \qquad \text{with } \xi^i \geq 0$$

To find the tightest sphere enveloping the points means that one wants the smallest R.

$\rightarrow$ This can be formulated as a constrained optimization problem.

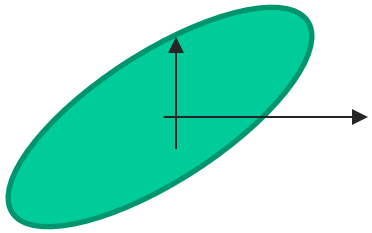
Support Vector Clustering

$$\min_{R,\xi} \left(\underbrace{R^2}_{\text{Sphere with minimal radius}} + \underbrace{C\sum_{j=1}^{M}\xi^i}_{\text{Penalty for slacks}} \right) \qquad (C \geq 0 : \text{ hyperparameter})$$

under the constraints

$$\left\|\phi\left(x^j\right)-\mu\right\|^2 \leq R^2+\xi^i, \quad \forall x^j, j = 1...M$$

$$\xi^i \geq 0, \quad , j = 1...M$$



The Lagrangian is:

$$L\left(R,\xi,\beta,\eta\right) = R^2 + C\sum_{j=1}^{M}\xi^i - \beta_j\left(R^2+\xi^i-\left\|\phi\left(x^j\right)-\mu\right\|^2\right) - \sum_{j=1}^{M}\eta_j\xi^i$$

Lagrange multipliers for all the constraints

Support Vector Clustering

This problem can be solved as for the standard SVM problem.
It has one global optimum. The quadratic form for the constraint yields a series of inner product of the form $\langle \phi(x^i), \phi(x^j) \rangle$, which we can replace by the kernel form: $k(x^i, x^j) = \langle \phi(x^i), \phi(x^j) \rangle$.
(see Ben-Hur et al, IJML 2001 - supplementary material for the derivation)

The Lagrangian is:

$$L(\mathrm{R}, \xi, \beta, \eta) = \mathrm{R}^2 + C\sum_{j=1}^{M} \xi^i - \beta_j \left(\mathrm{R}^2 + \xi^i - \left\| \phi(x^j) - \mu \right\|^2 \right) - \sum_{j=1}^{M} \eta_j \xi^i$$

Lagrange multipliers for all the constraints

Support Vector Clustering

For a query point x, we can write the distance of its image in feature space from the center of the sphere as:

$$R^2(x) = k(x,x) - 2\sum_{j=1}^{M} \beta_j k(x, x^j) + \sum_{i,j}^{M} \beta_i \beta_j k(x^i, x^j)$$

The coefficients β_i are non zero for the support vectors.

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

Kernel K-means metric!

In Kernel K-means, the cluster boundaries are determined by considering all the datapoints. Each datapoint has "equal weight" within one cluster. Each cluster's influence is weighted by the number of datapoints in the cluster.

In SVC, only a selected subset of datapoints are used to compute the cluster boundaries.
The influence of each datapoint is weighted by its β_i.
The optimization process of SVC determines the influence of the datapoints (the β_i are variables in the optimization).

→ SVC is a sparse version of Kernel K-means

Support Vector Clustering

For a query point x, we can write the distance of its image in feature space from the center of the sphere as:

$$R^2(x) = k(x,x) - 2\sum_{j=1}^{M} \beta_j k(x, x^j) + \sum_{i,j}^{M} \beta_i \beta_j k(x^i, x^j)$$

The coefficients β_i are non zero for the support vectors.

Cluster assignment is then determined:

1) either by looking at the isolines solution of $R(x) = cst$.

2) or one can build a binary similarity matrix S for each pair of datapoints

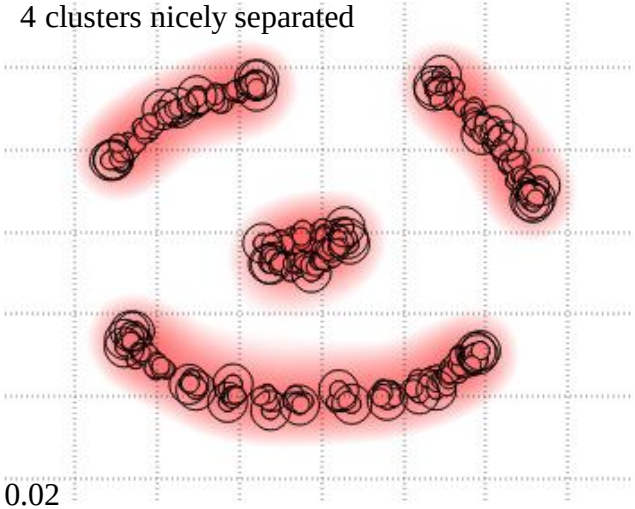
$$\begin{cases} S_{ij} = 1 \text{ if } R^2(x^i, x^j) < cst \\ 0 \text{ otherwise} \end{cases}$$

$\rightarrow$ we are back to spectral clustering. The clusters are found through a spectral decomposition of the Laplacian.

Support Vector Clustering

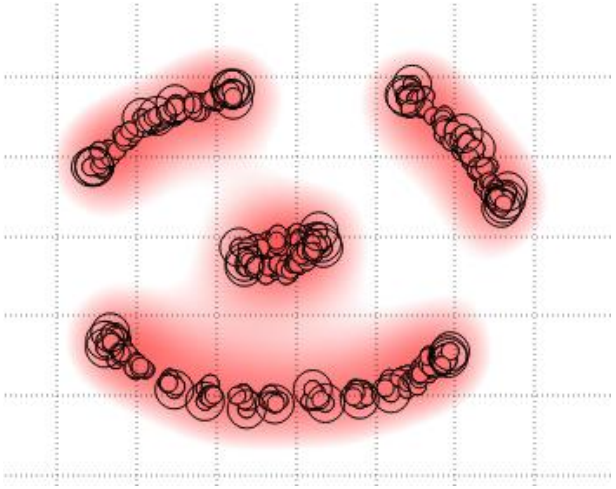
$\sigma = 0.01$

4 clusters nicely separated



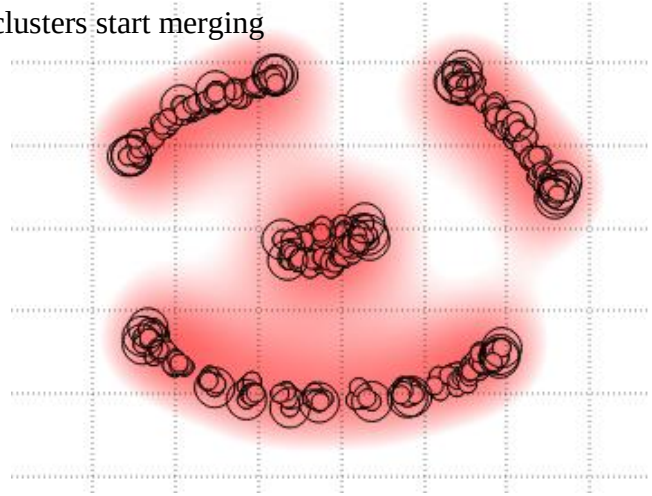
$\sigma = 0.015$

2 clusters start merging



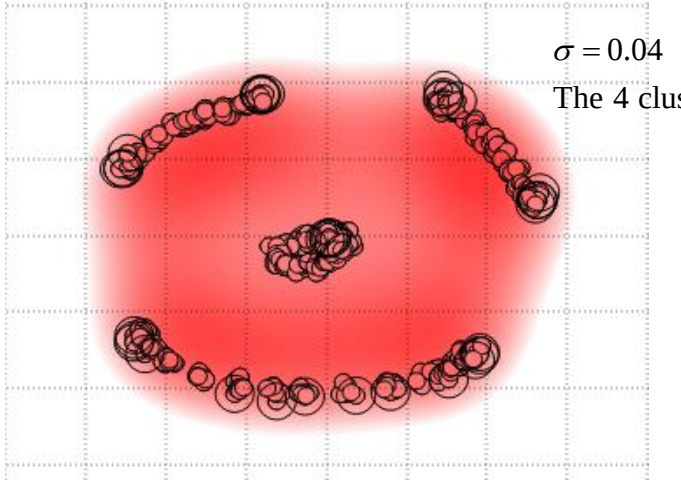
$\sigma = 0.02$

3 clusters start merging



$\sigma = 0.04$

The 4 clusters have merged



As the kernel width is decreased, the number of disconnected contours in data space increases, leading to an increasing number of clusters.