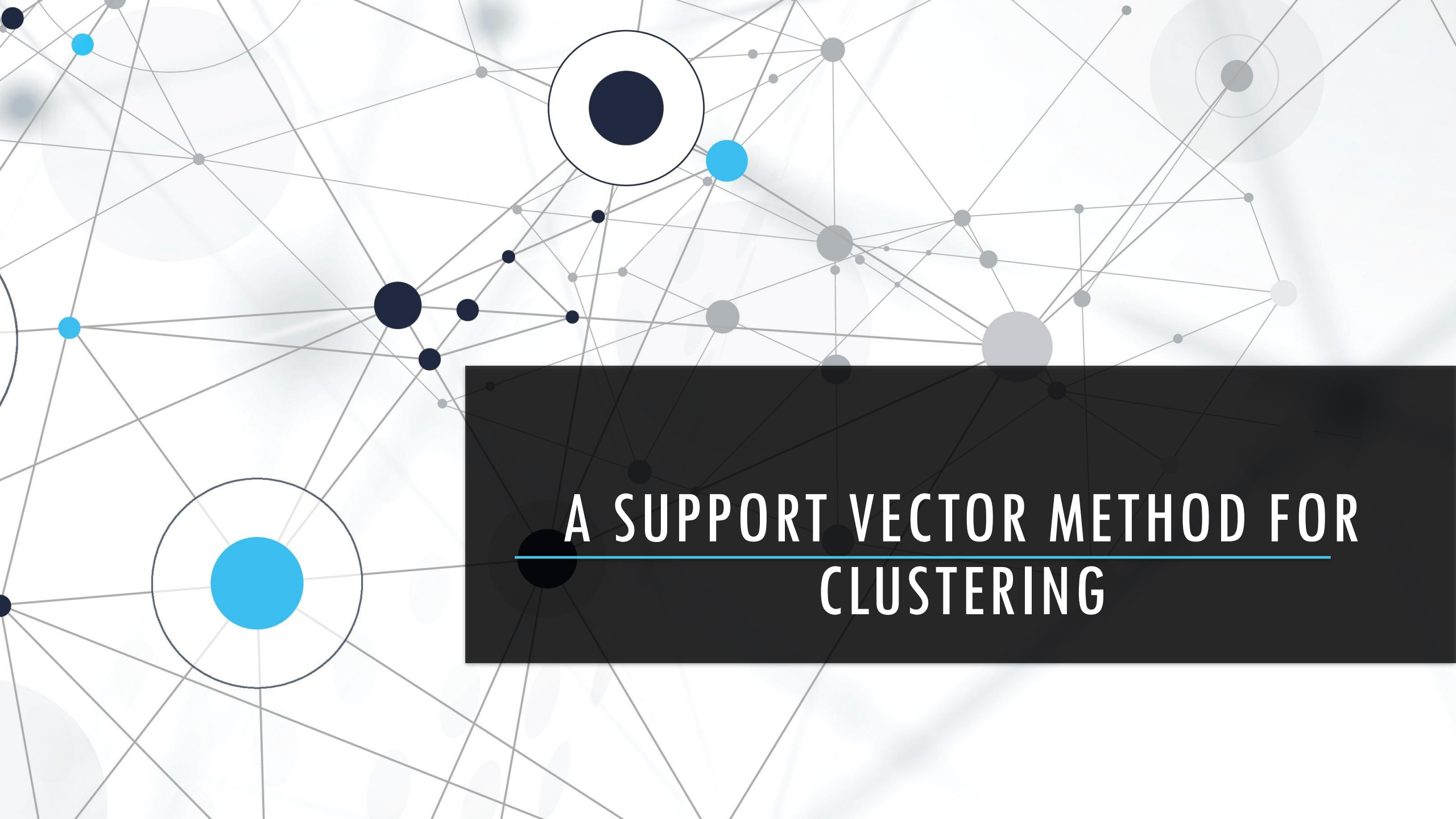


A SUPPORT VECTOR METHOD FOR CLUSTERING

INTRODUCTION

1. The objective of clustering is to partition dataset into groups according to some criterion to organize the data.
2. In this paper the clustering method is motivated by the ubiquitous Support Vector Machine.
3. The Support vector machine is one of the most powerful machine learning algorithm because of its convex optimization and its ability to separate non linear data points efficiently.
4. In this text ,the focus is on the kernel trick of the Support vector Machine, which maps the non linearly separable datapoints from the input space to the higher dimensional feature ,where a hyperplane can be obtained which linearly separates the data.

APPROACHING THE MAIN ALGORITHM

1. Clustering boundaries are made in regions where data density is low in input space.
2. In this algorithm, Gaussian kernel is used to map the input space to feature space.
3. In feature space a sphere with minimal radius is computed consisting of the datapoints, this sphere is then mapped back on the input space as result a set of contours are obtained in input space which form the cluster boundaries.
4. SVC uses SVDD(Support vector domain description) which provides a decision function that tells whether a given input is inside the feature sphere or not.
5. There are two main parameters q and C. q is width parameter of Gaussian Kernel whereas C is regularization parameter used to penalize misclassifications /outliers.

CLUSTER ASSIGNMENT

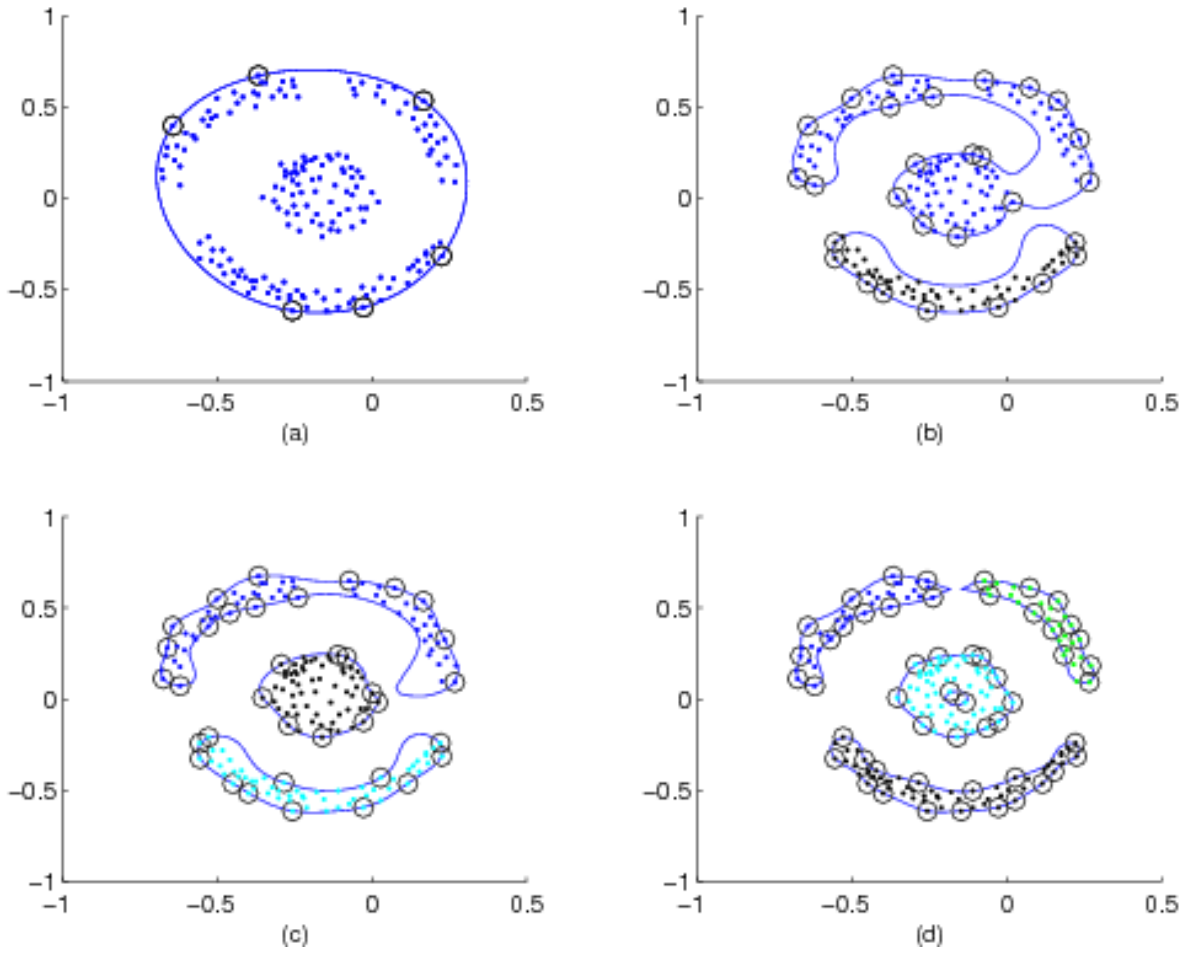
1. After mapping the datapoints back to the input space clusters are assigned to each contour.
2. An adjacency matrix A is generated. This matrix has either values 0 or 1
3. The value is 0 when the line segment between x_i and x_j crosses the hypersphere. The value is 1when the line segment between x_i and x_j is always in the hypersphere.

WIDTH PARAMETER

Initially we keep the value of C=1(without outliers)

It can be seen through the graphs from a to d that as we increase the value of q the contours start to break leading to increase in number of clusters.

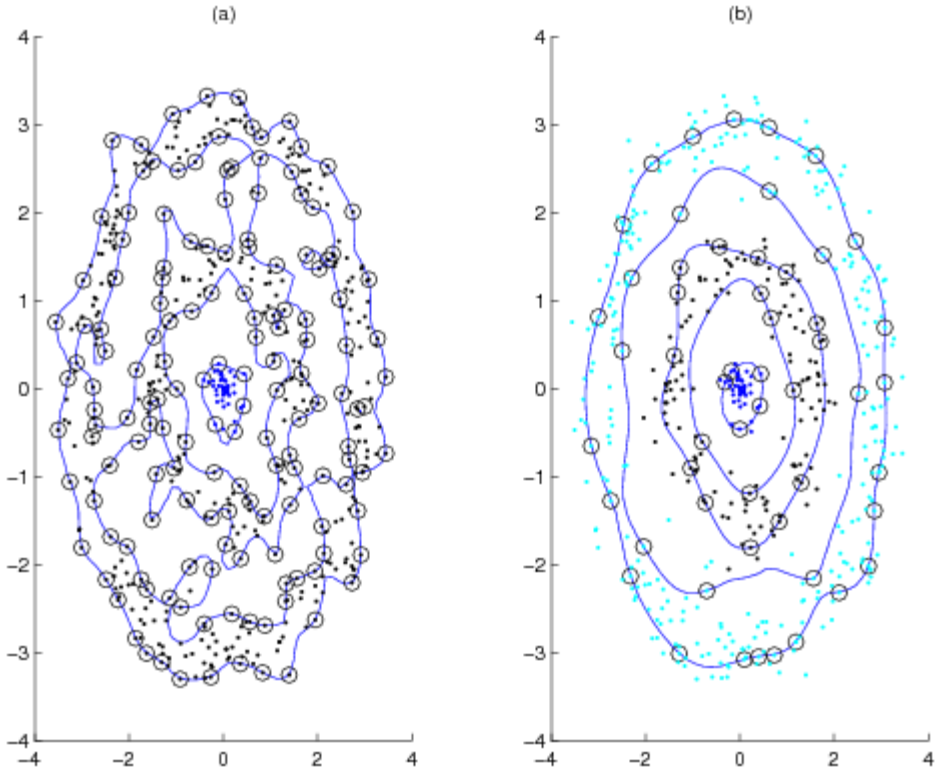
But continuous increase in the value of q may lead to overfitting of data.



REGULARIZATION PARAMETER

It controls the trade off between smooth decision boundary and proper number of clusters. If C is large contour splitting does not occur for the two outer ring for any value of q. Outliers are basically unclassified SVs , since they lie outside the enclosing sphere.

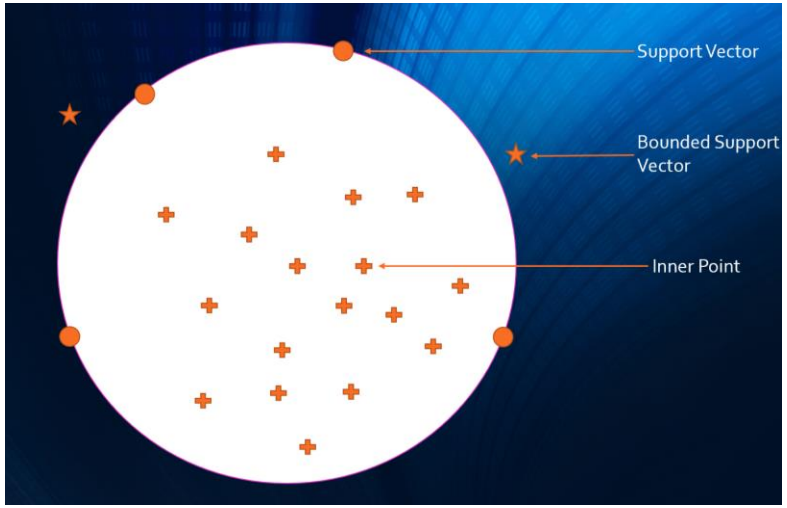
If outliers are allowed i.e., if C is considerably small then at some value of q , proper number of clusters are obtained.



SUPPORT AND BOUNDED SUPPORT VECTORS

Some bounded support vectors should be allowed so that the model does not overfit the training data. Allowing BSVs also makes the separation between two cluster clear.

As the number of bounded support vectors increase the number of support vectors gradually decreases which makes the cluster boundaries smooth and good.Number of SVs and BSVs are function of q,C and N(no. of samples)



CONCLUSION

1. Hence it can be said that overall SVC is powerful clustering algorithm with carful hyperparameter selection as it makes no assumptions of the clusters in the input space.
2. The quadratic programming problem introduced in SVC is a convex function and on optimization convergence is guaranteed .
3. On tuning the hyperparameters SVC algorithm can deal with outliers and noise in the data , which many other clustering algorithms cannot deal with.

THANK YOU