Kernel PCA and Spectral Clustering

Christophe Ambroise

Exercice 1: Kernel Principal Component Analysis

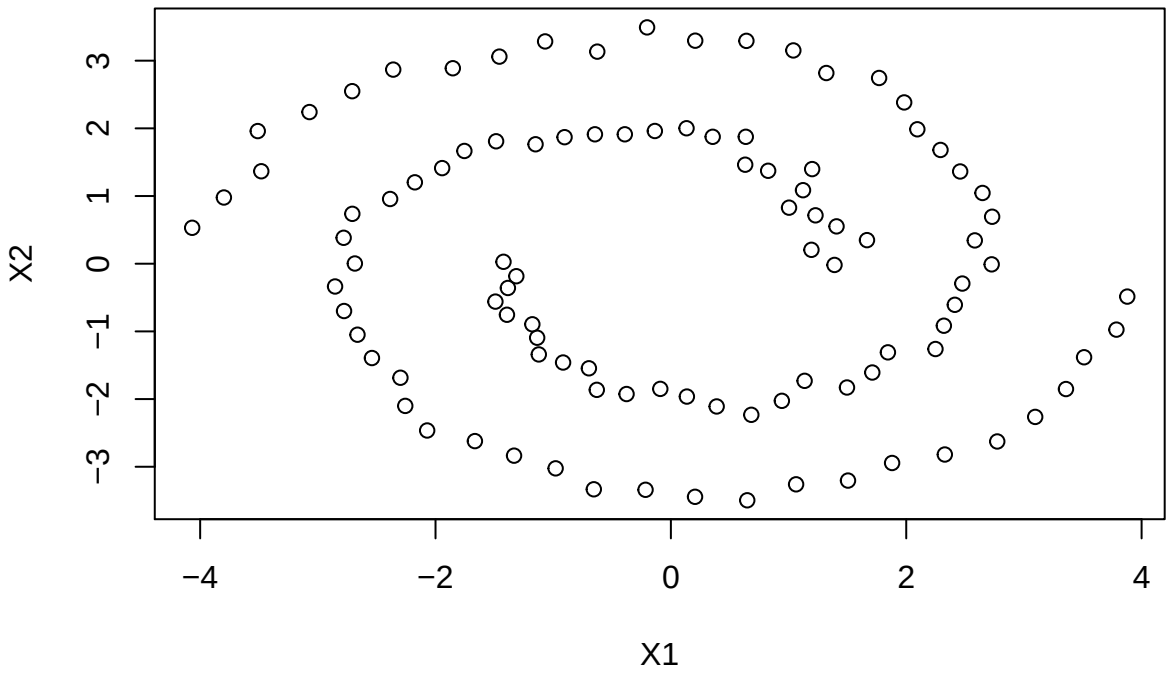
1. Install the `kernlab` package and load the dataset `spam`.
2. Perform a principal component analysis of the spam dataset.
3. Display the emails in the plane of the first two principal components with a color corresponding to their status (spam or not).
4. Do the same using various kernel and using KPCA.
5. Program your own `kpca` function.

Exercice 2: Spectral Clustering

1. Install package `mlbench` to access the `mlbench.spirals` and run the following code:

```
library(mlbench)

set.seed(111)
obj <- mlbench.spirals(100,1,0.025)
my.data <-  data.frame(4 * obj$x)
names(my.data)<-c("X1","X2")
plot(my.data)
```




```
my.data<-as.matrix(my.data)
```

2. Compute K the matrix of similarities for this dataset using the gaussian kernel
3. The next step consists in computing an affinity matrix A based on K. A must be made of positive values and be symmetric. This is usually done by applying a k-nearest neighboor filter to build a

representation of a graph connecting just the closest dataset points. However, to be symmetric, if A_{ij} is selected as a nearest neighboor, so will A_{ji}.

4. With this affinity matrix, clustering is replaced by a graph-partition problem, where connected graph components are interpreted as clusters. The graph must be partitioned such that edges connecting different clusters should have low weigths, and edges within the same cluster must have high values. Spectral clustering tries to construct this type of graph. Compute D the diagonal matrix of degrees (D_{ii} is the degree of node i).
5. Next compute the unnormalized graph Laplacian $U = D - A$ and/or a normalized version L.
6. Assuming we want k clusters, the next step is to find the k smallest eigenvectors (ignoring the trivial constant eigenvector).
7. Use standard k-means clustering to find the appropriate clusters
8. Perform spectral clustering using `kernlab` function `specc`
9. Perform Kernel PCA and then kmeans using the principal components. Compare the results and comment.