Advanced Machine Learning
Kernel methods
Practical Session

Master MLDM

Assignment: Due date 10/11/2017 22:00 on Claroline
The work can be done by groups of 2 students

Note: the file uploaded on claroline must contain your report and your implementations. The report must be written with care, write it as a data analyst that want to provide information on the interest of methods for a company or a lab. For the experimental parts, you must define the experimental setup (do not forget the problem of parameter tuning). Preferably, you can use scikit learn for this assignment.

1 Kernel-PCA

The objective of this part is to compare the classic PCA and the k-PCA on existing benchmarks. Your work:

- Implement the PCA and k-PCA approaches (note that for k-PCA the Gram matrix must be centered)
- You must illustrate the behavior of the two methods and you can generate your own datasets from existing benchmarks: moons, classification, circles, swiss roll.
- Your comparison must include different kernels (and parameters) and can include objective measures (variance, time, quality of pre-processing for classification, running time, ...)

2 Kernel K-means

In this part, the objective is to do a similar study for comparing the classic k-means algorithm and the Kernel k-means approach.

3 Logistic Regression

Logistic regression can be kernelized in a similar way as PCA or k-means algorithm and the objective of this section is to study Kernel Logistic Regression.

Work to do.

- Recall the principles of logistic regression.
- Find how to kernelize it, you must write all the justifications.
- Compare the performance of the 2 versions of logistic regression on synthetic data and existing benchmarks (different kernels must be tested).
  Note that in this part, it is not required to provide your own implementation (but you can if you want), you are allowed to use existing implementation, the focus must be on the justification of the kernelization and on the experimental study.

4 One class SVM and Maximum Enclosing Ball

In this part, the objective is to study the *One class SVM* or equivalently the *Maximum Enclosing Ball* problems which are particular cases of *Support Vector Data Description* (SVDD). Work to do:

- Propose an implementation of the SVDD problem under the form of an optimization problem. You should be able to compare the exact form, the one with errors (*i.e.* with slack variables) and using kernels.
- To evaluate the performance of this model, you can generate your own dataset and use existing classification benchmarks. One setting that you test is to learn the model with only one class and with only one class accompanied with few instances (if you have an implementation of the two class SVDD) of the other class and compare to models that use all data (such as with a classic SVM).
  Ideally your study can provide graphical illustrations and measures related to classification accuracy.