ROB311 - TD4 - Apprentissage pour la Robotique

Araujo Belén, Victor
De Carvalho Ferreira Sula, Julia

October 2019

SVM Digit Recognition

Support Vectors Machine have many applications, among them is recognition.For example, the algorithm can be used to recognized a digit basing itself on a set of images of hand written digits, and that is the main challenge will be solving throughout this TD.

As any machine learning algorithm there is the need of a dataset, thankfully, the MNIST dataset contains grayscale (8-bit), 28x28 pixels images of hand written digits which will be used for the training phase.

Finally, the algorithm efficacy will be verified by the detection accuracy and the confusion matrix.

Support Vector Machine (SVM)

The SVM is a supervised machine learning algorithm which can be used for classification purposes. Putting it very simply , the SVM aims to define a hyperplane in a N-dimensional space, N being the number of features available, that divides the data in different categories.

To find this hyperplane, this algorithm bases itself in support vectors. Support vectors the points closest to the boundary, that is to say, the points harder to classify. These vector have a direct influence in the margin size of a decision boundary.The Support Vector Machine uses these vectors to maximize the margin, therefore ensuring that the classes are distinct. Furthermore, the decision function is only specified by (usually small) subset of training samples, consequently this algorithm does not have a high computational cost.

Since the SVM is based in margins, it is possible to relax or tighten margin constraints, which allows to comport outliers without being strictly bounded to a perfect division.

Python: Sklearn library

Sklearn is a python's library for machine learning application.Besides all tools for data mining and data analysis that it offers, it contains also a SVM method, which allows a simple and clean implementation of this algorithm.

Firstly, this library offers a function to define the support Vector Machine in the form:

svc = svm.SVC(C=1.0, kernel='rbf', shrinking=True, probability=False, tol=0.001, max_iter=-1).

The main parameters that can be chosen are: c, penalty for the error term, kernel, type of decision boundary, shrinking, the use of shrinking heuristics, tol, stop criterion tolerance, max_iter, maximum number of interaction.

As the majority of machine learning methods, SVM has a training and testing phase that are also defined in the sklearn library the functions *svc.fit(trainX, trainY)* and *predY = svc.predict(testX)* respectively.

Furthermore, this library also offers metrics functions that allowed to measure the efficacy of a boundary decision. The main methods to do so are the svc.score which percentage of data points was well classify and the confusion matrix calculus.

Machine learning in big data sets

One of the main problems of machine learning is that normally there is a need to analyze large data sets, which can be rather computationally costly. Therefore, there are methods to synthesize data sets without loosing its main information,for example Principal component analysis and the Histogram of Oriented Gradients, which could also be applied to this problem if there was an performance constraint.