Active learning

using the Platt approximation of the conditional probability to the SVMs

Jonathan Grizou – AE ESE

BioCircuits Institute, University of California, San Diego

Ramon Huerta, rhuerta@ucsd.edu

Abstract

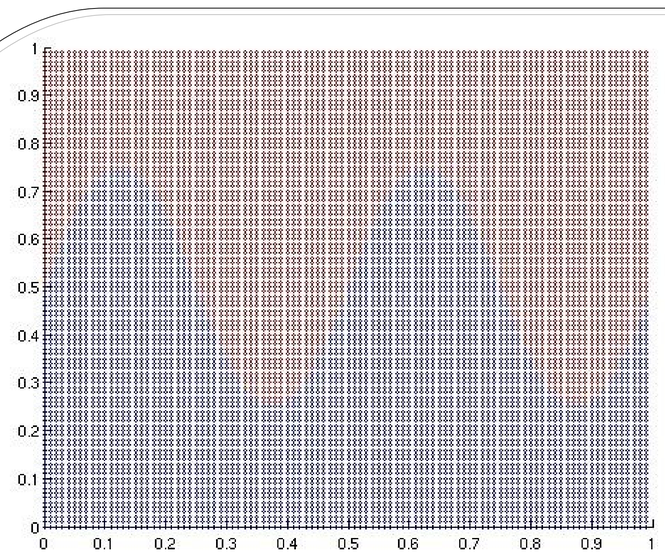
Nowadays the limits of model development are not imposed by the computer training time but by the available data and by the total time required to obtain sufficient data. We propose to develop a tailored strategy of active learning. Active learning is the process by which the selection of the next sample to learn is an integral part of the learning process itself. There are many applications like web searching, email filtering and ranking algorithms that do not have labeled data. Even though the data point is available, the class it belongs is expensive to obtain. For example, a piece of financial news is a data point, yet somebody have to read the news itself and determine what type of news it belong too. Thus, there are some applications of machine learning, where active learning is absolutely needed.

Brief introduction to the SVMs

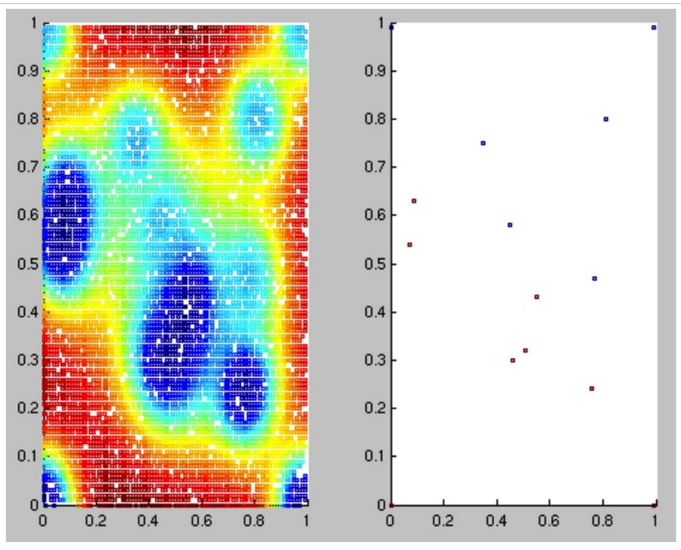
The Support Vector Machines (SVMs) is a **classification algorithm** well known to work with real database, and particularly with odor detection (main Ramon Huerta project topic).

Vocabulary : In the classification world, one data point is compose of a label and some features.
The label is the class belonging to the data, his identity.
The features represent the characteristic of this label, that could be a temperature, a concentration, a coordinate, a word,etc.

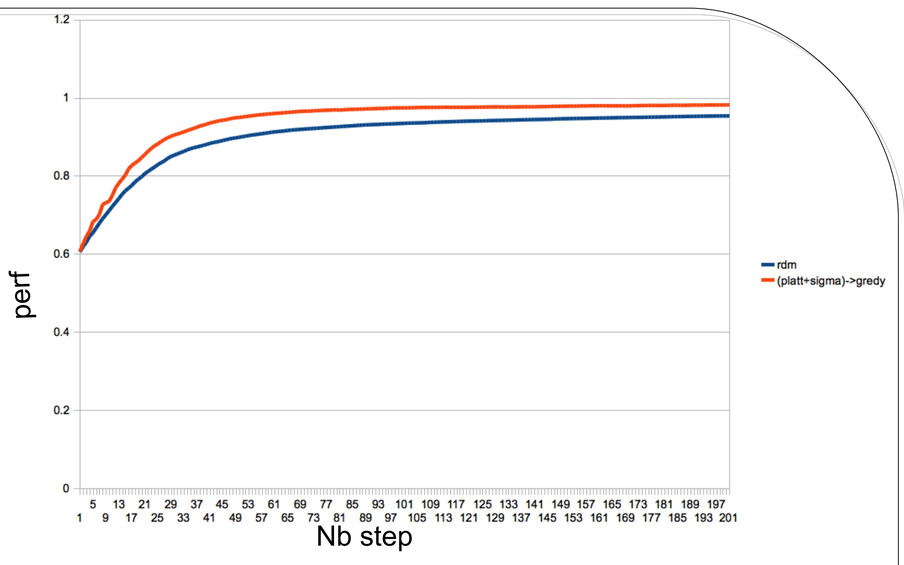
The feature space is the same for all the class and data in a classifier. Whatever the number of feature, the key is to find the correct features who permit to the SVMs to dissociate the different classes.

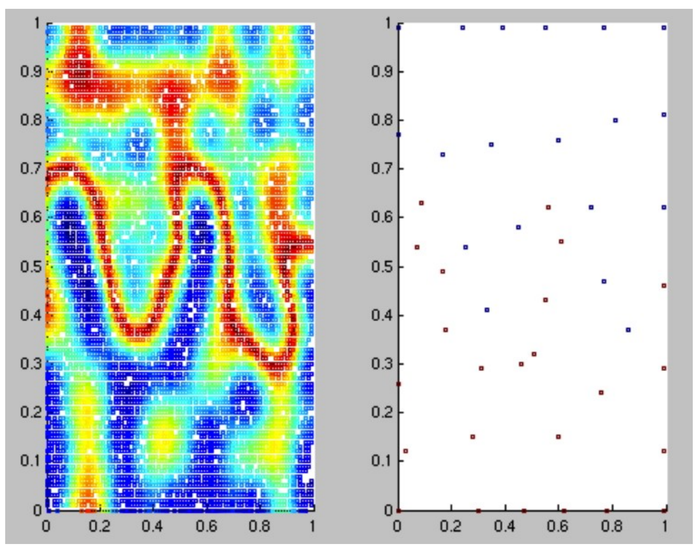


sinus-map

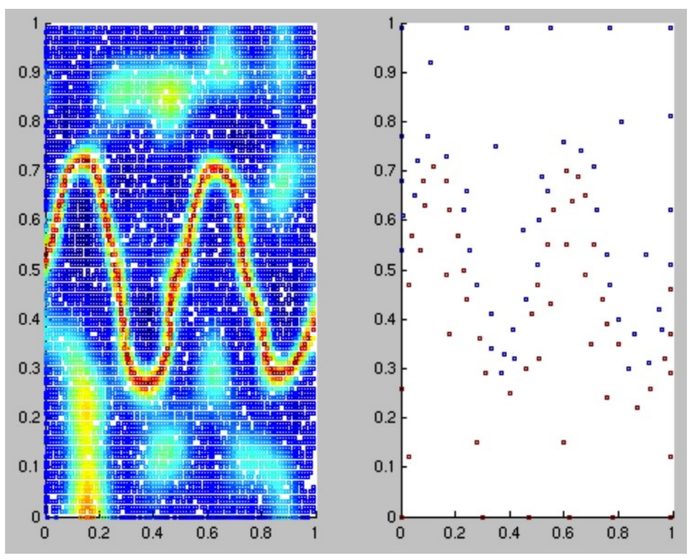


12 steps

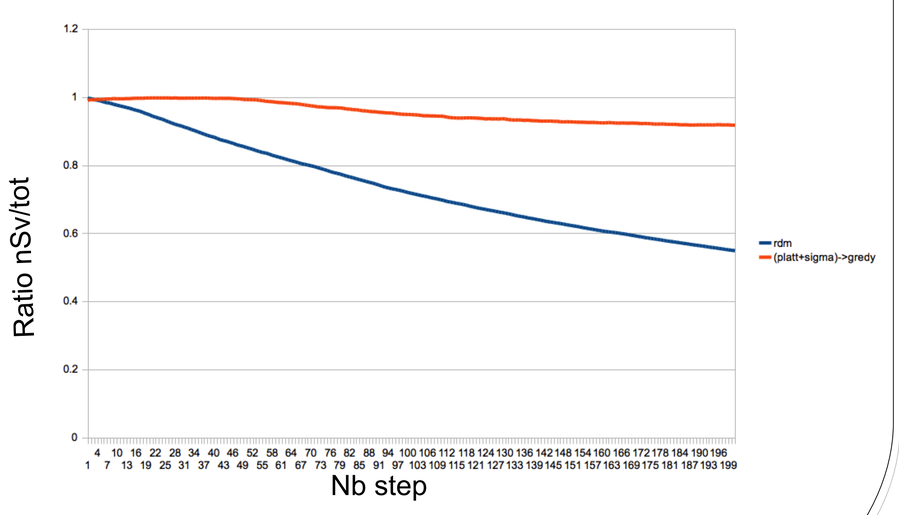




40 steps



100 steps



Conclusion and future

The algorithms developed show relevant performance improvement. I now have to use them on real data and to create a ranking between the different algorithm I developed (platt & sigma , greedy/montecarlo) in quantitative terms.

The next step is to reverse the process, instead of selecting the next point depending on his feature we will chose it depending on his label. In some real case, like odor sensing classification we know which gas or mixture of gases and which concentration we can test (label) but not the reaction of the sensor (features). So the algorithm will predict the next experiment to do in order to improve the “knowledge” of the SVM and consequently minimize the numbers of experiment needed.

Personal benefit

Skills developed:
- programing C++ / Matlab
- mathematics, probability

General overview:
- odor detection / odor sensor
- data collection / learning algorithms

Opportunity to visit other lab :
- neuromorphics engineering
- evolvable hardware / neuron on silicon
- signal processing / wavelet function