Graphs in Machine Learning
TP n°2

Marc Szafraniec

27/11/2016

1 Semi Supervised Learning and Harmonic Function Solution

Q1.1: I chose a eps graph, with parameters threshold = .2 and $\sigma^2 = .2$.

Even with only four points we can get a perfect labeling.

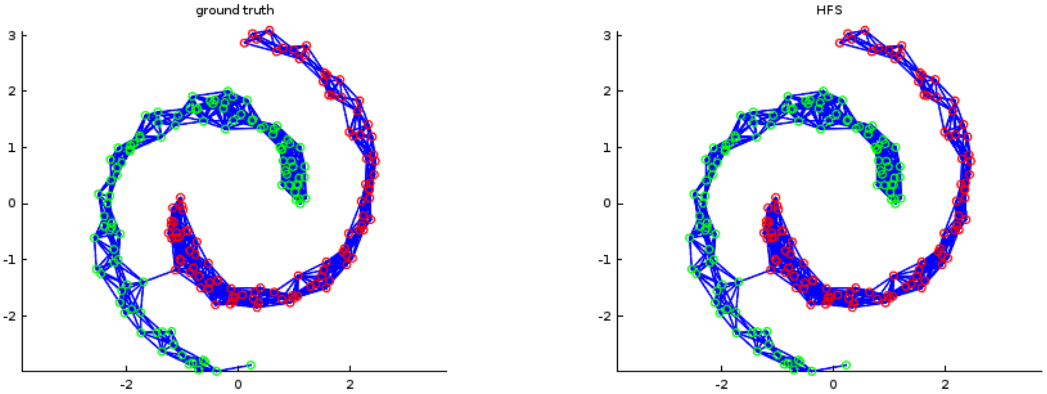

Figure 1: One try of 4 random labels (S) and Hard-HFS, with accuracy = 1

Q1.2: For this second experiment, with the same distribution but a bigger sample sets, we got exactly the same results. The main difference is the time consumption. A problem can appear here if we are unlucky because there is a not-so-small chance to only pick 4 labels that are the same (class 1 or class 2). If this is the case, the program considers that there is only one class, and the accuracy is very low.

Q1.3: For the Soft HFS, we must choose parameters c_l and c_u. As no indications were given and that I found nothing else about it, after some attempts I took $c_l = 0.95$ and $c_u = 0.1$ which give perfect results, the same as Hard HFS.

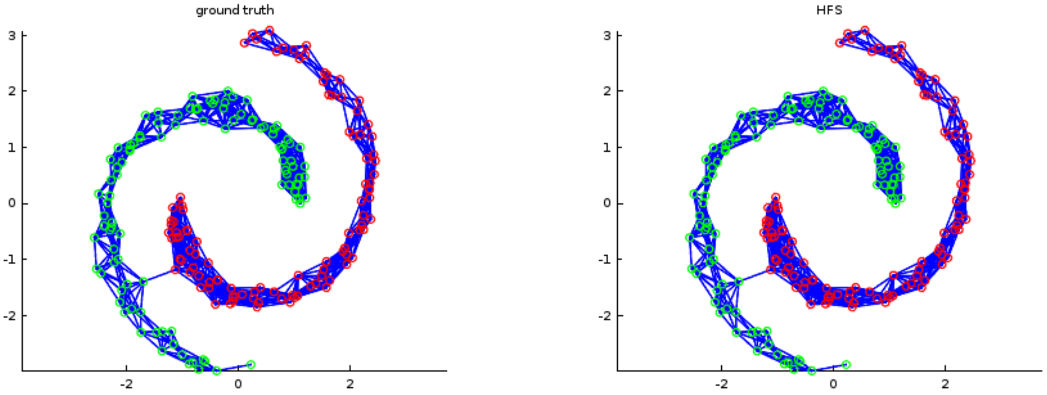

Figure 2: One try of 4 random labels (S) and Hard-HFS, with accuracy = 1

When comparing the two models, we have 20 labeled points, and we see that the performance is very similar for the two models (~ 0.9 accuracy).

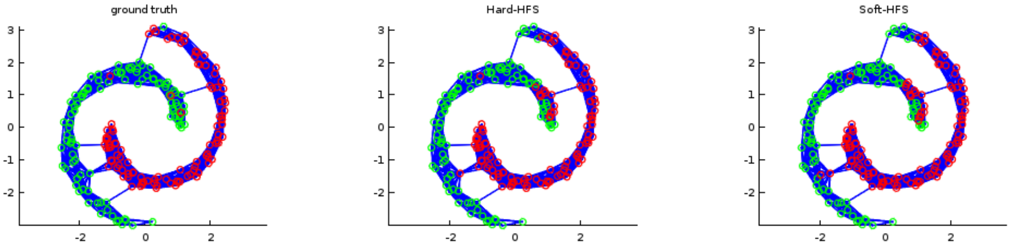

Figure 3: Comparison Hard vs. Soft, accuracy 0.905 vs. 0.880

2 Face recognition with HFS

Q2.1: In order to generalize from 2 class (like in part 1) to 10 classes (for these 10 people), I encoded the classification problem with a target vector $Y \in \{-1, 1\}^K$ that allows to have as many classes as we want.

Q2.2: I applied both *GaussianBlur* and *equalizeHist*, but it doesn't improve the performance for me. The performance goes from 84% to 78% with *GaussianBlur* or both or even 62% with *equalizeHist* alone.

Q2.3: HFS reaches a good performance on this problem, with an accuracy of 84%. Soft or Hard HFS doesn't make a big difference, but the choice of parameters does. Here I used a knn graph with $k = 100$, $\sigma^2 = 80$ and a rw Laplacian with a 0.001 regularization parameter.

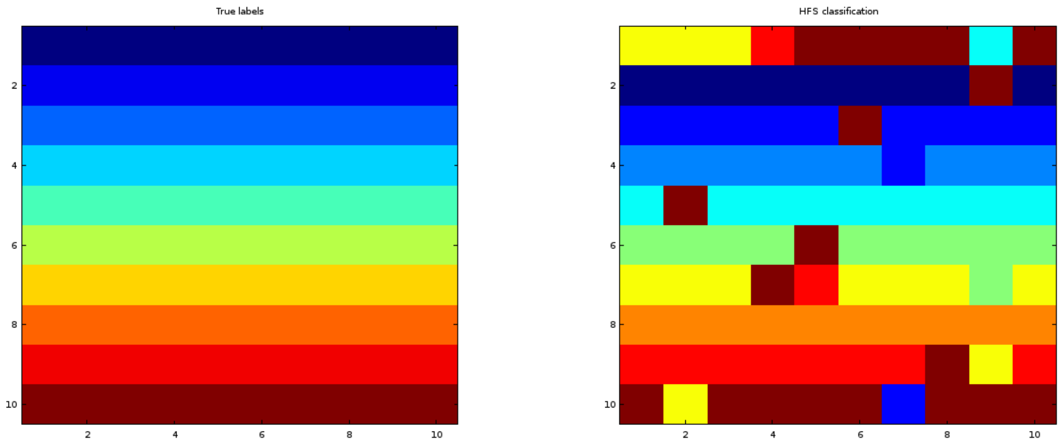

Figure 4: Hard HFS Face Classification, with 84% accuracy

Q2.4: Adding 100 new images, randomly selected in the extended set, sometimes improves the performance by a few percents (around 5% in general), and sometimes reduces it by a similar amount. The images that could improve the performance are images that are easy to classify, that fits the class defined by the already labeled images.

Q2.5: The data that could degrade performance would probably be data that is very different from the existing data (other persons in this case), or if the new data totally unbalances the different classes. Generally, new data should have a similar distribution to the one in the labeled data.