Graphs in Machine Learning
Homework 2 - Face Recognition

Mohammed Amine KHELDOUNI

27 November 2018

1 Harmonic Function Solution

Question 1

In this question, we implemented a hard Harmonic Function Solution (HFS) algorithm capable of predicting labels Y of our data X, given a set $\mathcal{L}$ of labeled data (Semi Supervised Learning). In this first section, we coded a model based on binary classification (two classes) where the prediction can be easily recovered by the sign of f_i (following the handout's notations). Later on, we adapted this hard HFS to work on multi-class problems. Therefore, the code proposes a multi-class implementation without the two classes special case.

Experimenting our hard HFS on a two moons dataset, we construct a masked labels vector where only 4 labels are randomly selected as known ($l = |\mathcal{L}| = 4$) and we construct a $k - NN$ Graph (with $k = 6$) with a variance of $\sigma^2 = 1$. The classification is illustrated in the following figure, and reaches a perfect accuracy of 1.0.

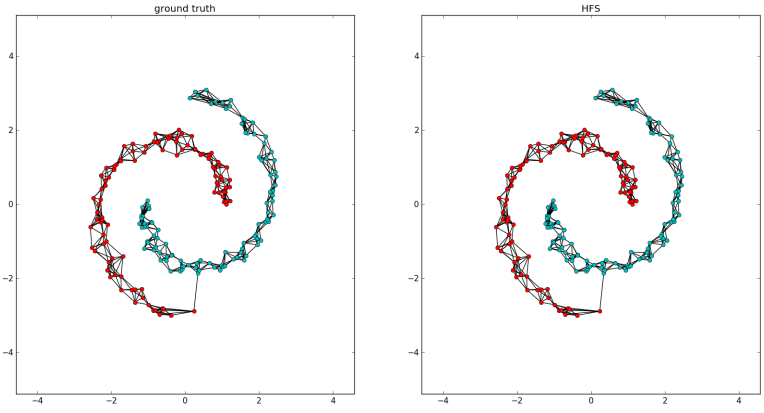

FIGURE 1 – SSL classification of two moons using a hard HFS with 4 known labels

Question 2

This time, we run the hard HFS on a large dataset of two moons clusters having 1000 samples without changing the number of known labels ($l = 4$). We chose to build the Laplacian matrix of an ϵ

graph rather than a $k - NN$ graph because of the compactness of each moon in the large sample set. We therefore tuned the value of ϵ (setting it at 0.72) for the graph to be fully connected. The results of the classification are displayed in figure (2) and performs really well.

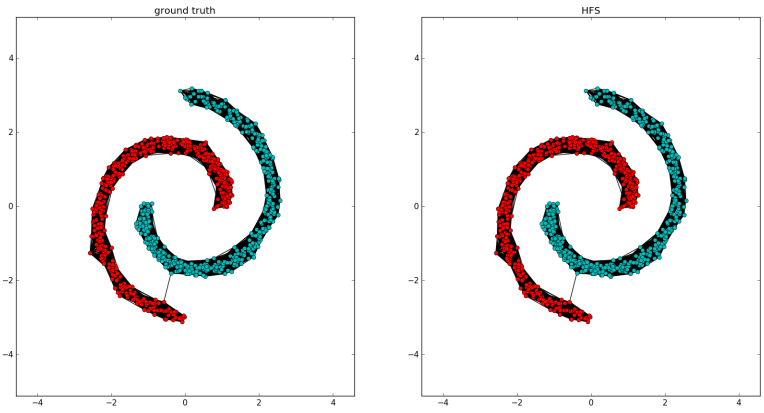

FIGURE 2 – Two moons classification using a hard HFS with $l = 4$ known labels on a large sample ($N = 1000$, $\epsilon = 0.72$)

However, with such a large sample of data, pulling randomly 4 known examples may result in a problem if labels from a certain class are not picked as known labels. In this case, the model fails the classification because it does not have the right number of labels and does not acquire enough information about the different labels' location in the graph (3). Hence a really bad accuracy when this issue happens. However, we had to run the implementation several times before this misclassification case occured, but we acknowledge the fact that with much larger samples and with multiple labels (non binary classification examples), this issue would occur even more frequently.

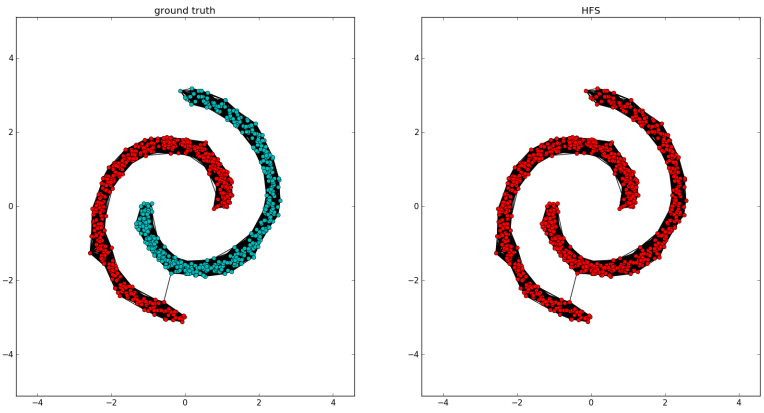

FIGURE 3 – Two moons misclassification using a hard HFS when there are only $l = 4$ known labels on a large sample ($N = 1000$, $\epsilon = 0.72$)

Question 3

We introduce some noise upon the known labels and implement a soft Harmonic Function Solution. This model introduces some uncertainty over the known $l = 4$ examples of labeling, by adding a regularization term. The classification in figure (4) shows that both algorithms still have a good accuracy and

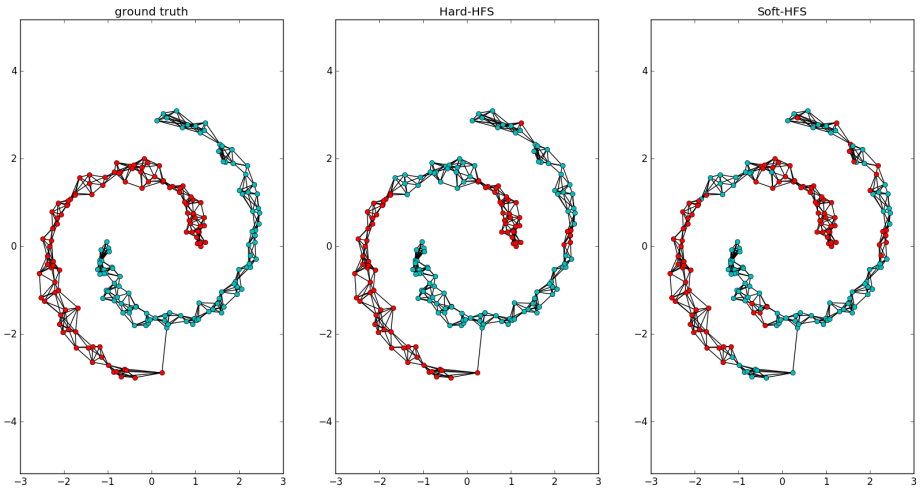

FIGURE 4 – Hard and Soft HFS classifications when the known labels are uncertain in two moons dataset

provide roughly the same accuracy. Indeed, the average accuracy of both hard and soft HFS are very similar. In the experiment above, the accuracy of hard HFS reaches $a_{hard} = 0.855$ and the accuracy of soft HFS is $a_{soft} = 0.785$.

The similarity between the two algorithms is still noticeable over the large two moons dataset of 1000 samples where the performances are ($a_{hard} = 91.6\%$, $a_{soft} = 81.6\%$) for our example. Note that there may be cases where soft HFS may perform better than HFS and restrict the impact of a bad labeling of some examples.

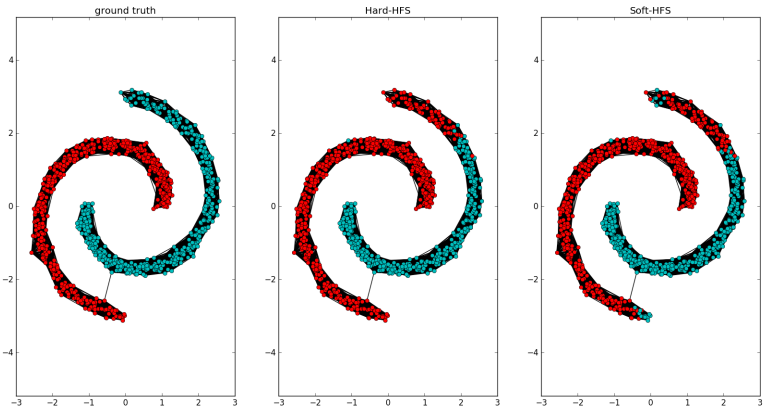

FIGURE 5 – Hard and Soft HFS classifications when the known labels are uncertain in large two moons dataset ($N = 1000$, $\epsilon = 0.72$)

Finally, we tuned the parameters of regularization to obtain good performances, and we set :
— The Laplacian regularization coefficient to $\gamma = 0.04$
— Regularization for labeled examples $c_l = 0.96$
— Regularization for unlabeled examples $c_u = 0.04$

2 Face recognition with HFS

Question 1

In this section, we generalized our HFS approach by adding the multi-class labeling aspect. In the first section, we used the property where we could classify by the sign (positive/negative) of our computed vector f. We managed to consider $K > 2$ classes by considering f as a matrix of 0 or 1 values. f has N rows and K columns and $f_{ij} = 1$ if the example i is predicted to be of label j and 0 otherwise. By construction, since each example can only belong to one of the K classes :

$$\sum_{j=0}^{K} f_{ij} = 1$$

Therefore, we recover our predicted labels by looking at the position of 1 for each example (ie. each row of f).

Question 2

We used two pre-processing steps and compared the results to the case where we do not perform any pre-processing. More precisely, we used the *equalizeHist* function to equalize the histogram of the image. This processing normalizes the brightness and increases the contrast of the image. Then we used a Gaussiand blur (*GaussianBlur* function) that convolves the image with a gaussian kernel to get a smoothed output image [6].

As expected, the normalization of brightness and the increase in contrast decreases the accuracy of the hard HFS model because those features could have been useful for the graph to learn from the brightness and the contrast in the picture, which face corresponds to the tested image. Therefore, the Gaussian blur performs better results with an accuracy of 88%.

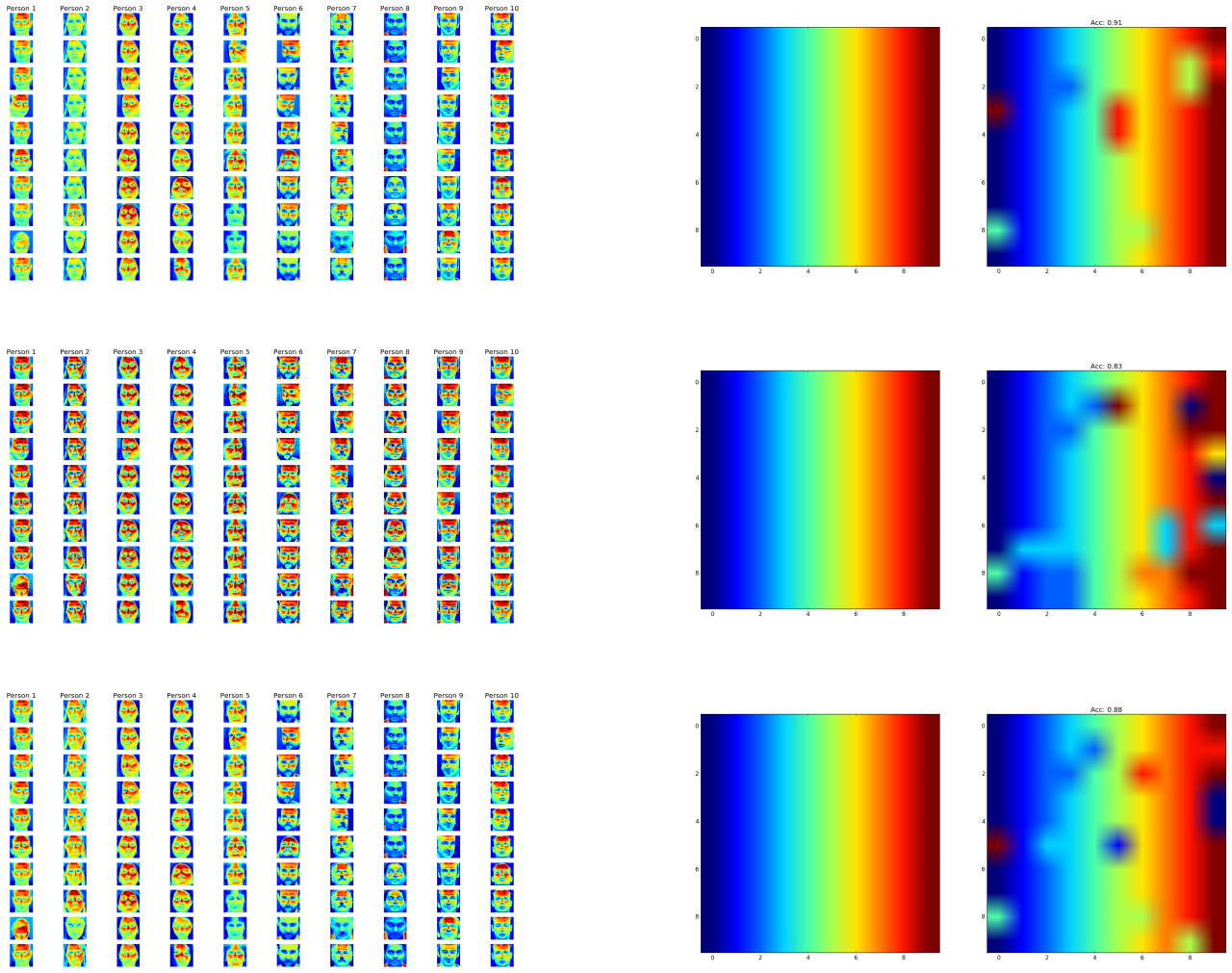

FIGURE 6 – Figure resuming the results for gray image without pre-processing (*top*) with equalization of histogram (*center*) and with a Gaussian blur processing (*bottom*). The hard HFS results in accuracy of 0.91, 0.83 and 0.88 respectively.

Question 3

As explained earlier, the hard HFS performs pretty well classifying the gray images blurred with a Gaussian kernel. The resulting accuracy reaches 88% which represents a good accuracy.

Question 4

Adding more additional data to the classification task does not improve performance. Indeed, our accuracy decreases from 88% to 68.4% for large dataset of images.

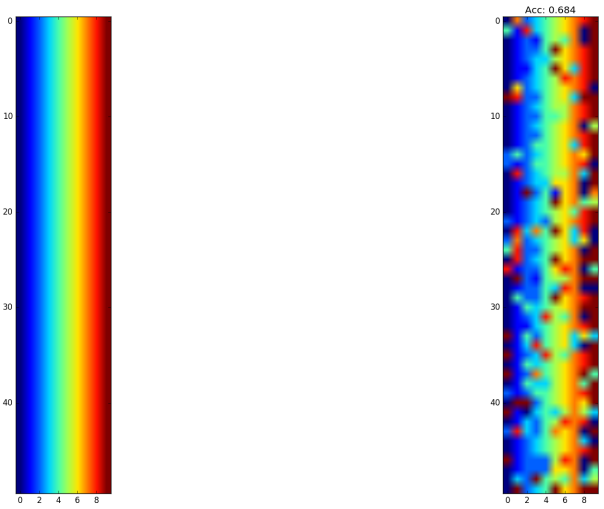

FIGURE 7 – Performance of the hard HFS for augmented dataset ($a_{augmented} = 68.4\%$)

Question 5

The performance did not improve when adding additional data. We have many assumptions about this decrease in the accuracy.

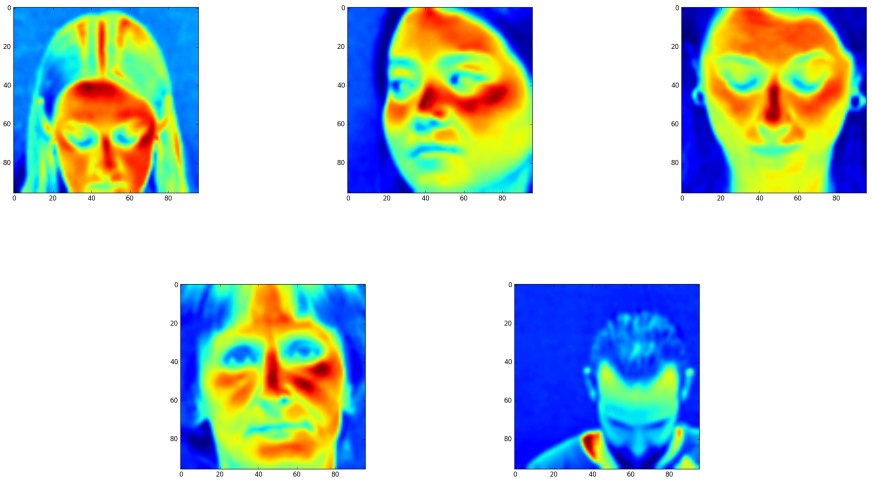

FIGURE 8 – Some examples where the model fails to recover the true label

First, we tend to believe that some added images can be really difficult for the model to recognize. Indeed, adding too many features or covering a part of the face may result in an error of classification.

Second, in Semi Supervised Learning, extending the data may lead to the same issue we discussed in the previous section. As we grow the number of examples, the known labels may not help the model to recover the face of a person. In other words, if the only known labels (4 images classification) for person

i are complex images where features of the person's face cannot be learnt, we end up misclassifying some images.

3 Conclusion

In this work, we managed to construct a working classification model based on Semi Supervised Learning and Harmonic Function Solutions. We compared Hard-HFS and Soft-HFS when noise is introduced on known labels.

Finally we applied this classification system to solve face recognition problem and discuss the capability of this solution with a high number of images (or data in general).