Machine Learning Project 2: Unsupervised Analysis of a dataset made of leaves

1 Goal of the Project

In the "unsupervised learning" part of the course, you learned what unsupervised learning is. In particular, you studied two types of unsupervised problems: clustering and visualization. In this project, you will use these two approaches to explore a particular dataset made of plant leaf images, and extract some insights about it. More particularly, you will provide visualizations of the data using *t*-SNE, and a clustering using the method of your choice. You are expected to handle out a report of max. 4 pages, including plots. This report can be written in English or in French. **Additional pages will not be read.** On the implementation side, you will use Python 3 with the `scikit-learn` framework (`http://scikit-learn.org`), which provides many tools for machine learning and has a detailed documentation. Please check the instructions in `https://scikit-learn.org/stable/install.html` to install scikit-learn. **Your code should be submitted with your report on Webcampus (by providing a notebook or a .py file).**.

2 The dataset

The dataset is made of a little bit less than 200 images of plant leaves. The images are coloured and have all the same size (72x128px). All these images are contained in the "dataset" folder. However, no further information are provided (i.e., no labels). As for the first project, a companion library (`utils.py`) is provided. The dataset can easily be loaded using the `load_data()` function in it. The `load_data()` function takes care of the preprocessing of the images. Use it and do not try to load the images manually.

3 Playing with *t*-SNE

We provide an implementation of *t*-SNE in the `tsne.py` file. For the first part of this project, you will have to visualize the data using this implementation of *t*-SNE. It requires two parameters : `X` the data to be processed (the images) and `perplexity`. Concerning the metaparameter `perplexity`, you will need to find a good choice for it. Once you obtain an embedding for the data, you can use the `imscatter()` function of the companion library to generate a nice scatter plot with the corresponding images embedded in it. The `x` and `y` arguments correspond to the two coordinates obtain with t-SNE, and the `images` argument correspond to the dataset, as retrieved by the `load_data()` function.

Your first task is to obtain visualizations of the data. Find a good choice of `perplexity` and show the visualization you obtained in your report. Again, justify your choice. What kind of information can you get thanks to this visualization? On another side, should you beware of particular things?

4 Clustering

In the last part of this project, you will use a clustering algorithm in order to obtain the different species of plants present in the data. Again, `Scikit-learn` provides implementations for different clustering algorithms. You are free to choose the one you want to use. Once you obtain a clustering for the data, you can use the embedding you obtained with t-SNE in task 1 to visualize it (this time, you can directly use the `scatter()` function of `matplotlib` to visualize the clusters with points of different colors. The companion library has a function `get_colors()` that can be used to generate the argument `c` that you will need when calling `scatter`)

Your second task is to choose a clustering algorithm. Why did you choose this one? Then, use this algorithm to cluster the data. How did you set the different metaparameters? What is the correct number of cluster for those data ? Justify. Visualize your clustering with the embedding you obtained in task 1 with t-SNE. Put this visualization in your report, and comment it.