SECOND ASSIGNMENT. PART II. 2.5 POINTS

SCIKIT-LEARN APPLIED TO DIGIT RECOGNITION (SEMEION DATASET)

1. INTRODUCTION

We are going to work on automatic digit recognition. The idea is to classify automatically handwritten digits. We will use a dataset call SEMEION with digits from 0 to 9, represented with a 16x16 black and white image matrix. Instances in the dataset contain 16x16 = 256 binary input attributes plus the class. Binary attributes are 0 if the image was white for that pixel and 1 if the image was black for that pixel. The class contains numbers from 1 to 10, representing digits from zero ('0'), one ('1'), …, to nine ('9').

Figure 1: written in normal way

Figure 2: written in fast way

You can find more information about the Semeion dataset here: https://archive.ics.uci.edu/ml/datasets/Semeion+Handwritten+Digit

2. WHAT TO DO:

First of all, load the data, which you can do with:

semeion = np.load("semeion_bin.npy")

1) **(0.5 points)** In this case we are not going to use crossvalidation for estimating future performance, but train/test (this is in order for experiments to be faster). Therefore, the first thing to do is to decompose the original data in a train / test partition, as it was already explained in one of python notebooks
2) **(0.5 points)** Train a decision tree and a knn model with the training data, test them with testing data, and see which one does better
3) **(0.5 points)** Change some of the hyper-parameters by hand, and see if you can get better results. Report the success rate on the test dataset.
4) **(0.5 points)** Use hyper-parameter tuning, **on the training dataset**, in order to find the best hyper-parameters for knn and decision trees. Report the success rate of the best models on the test dataset. Do they improve on the results you got on steps 3 and 2?
5) **(0.5 points)** Use at least one attribute selection or attribute transformation method on the training data, then build a model with decision trees with the hyper-parameters you got on step 4, and then report success rate on the test dataset.

3. WHAT TO HAND IN:

A notebook containing the python code and comments that answer the previous steps. In the last section of the notebook, please explain if you have done something special or noteworthy, so that I can pay attention to it. For this assignment, it is allowed to work in **groups of two students**. Please, write both names at the beginning of the notebook.

Note: if experiments are too slow because the dataset is too large, instead of using data for digits ‘0’, ‘1’, …, ‘9’, use data for digits ‘0’ and ‘9’ only. If you do not know how to do this, please ask me.