**FST course - TP exam**

# 1 Aim of this exam and instructions

In this TP, you are given a dataset with examples labeled by a categorical variable. You are asked to design some classifiers to predict the class of nex examples (for which you don't know the true class). You will have to select classifiers that seem good for this task on the given dataset, following the procedure explained during the course and that you applied during the TPs.

The dataset is given in the file `arabic_dataset.csv` on teams. In the file `script_etu.ipynb` you will find some useful instructions to start with this work. You have to fill this file and send it to me after the end of this session.

Examples in this dataset are black and white images representing handwritten characters from the arabic alphabet. Below is an example of such an image :

This image corresponds to the character 'Taa' of the following picture (even if it is not exactly similar)


In the provided dataset, you have 1700 images of 32*32 pixels. There are only 7 different characters in this dataset :

- label 1 : alif
- label 2 : ba
- label 6 : hha
- label 12 : sin

— label 16 : taa
— label 18 : ayin
— label 22 : kaf

The aim of this TP is to build different kinds of classifiers that seem relevant for this task (in terms of accuracy of predictions). Then, you will apply the different classifiers that you have built to a new dataset of 500 images for which you don't know the label, and you will check the performance of your predictions on the Kaggle website. This 500 new images are in the file `competition.csv`. You will use this dataset only to make predictions using the different classifiers that you designed previously.

This exam is composed of 3 parts. In the first part, you have to make a quick description and analysis of the dataset you are given. In the second part, you will use the raw images to design different classifiers that you have seen during the course. In the third part, you will apply the HOG representation to raw images in order to try to improve the performance of the classifiers.

For part 2 and 3, you have to study the following methods in this given order :

1. Decision trees
2. SVM
3. k-NN
4. Random forest
5. Logistic regression

# 2 Expected work today (13-14 points)

You will be evaluated today on the first part and on the 3 first methods of classifiers for the second part (nothing mandatory with HOG today)

You are asked to select :
— 1 decision tree
— 3 different SVM (corresponding to the 3 kinds of SVM seen in the course)
— 1 k-NN classifier

Then, you will use these models to predict the labels of the competition dataset, and submitt your predictions on Kaggle (explainations below).

At the end of today's session, you have to send me a `.ipynb` file containing your code and some explainations including :
— Description and analysis of the dataset for this classification task
— A detailed description of the methodology that you use
— The results of the evaluation of the different methods **given in a table**, and if needed an analysis of the obtained results
— Your global conclusion

Mail adress to use : `simon.malinowski@irisa.fr`.

The file should be **cleaned, sequentially executable, and commented**.

**The quality of the methodology used and of the report will be more taken into account than the quatity of methods tried**

# 3 Submission of your predictions on Kaggle

The link to the Kaggle competition is in the file `kaggle-link.txt` on teams. You'll have first to create an account on the website `https://www.kaggle.com` (it is free).

After studying one of the methods above, you can predict the class of the images of the competition dataset, save them to a file and drag this file on Kaggle (button `Submit predictions`). More explanations are given on the `.ipynb` file. You can add a comment to the submission (which method, which parameter, etc...). An example to show how to submitt predictions is given in the file `script_etu.ipynb`.

The score that will be displayed on the leaderboard is computed on only 60% of the examples of `competition.csv`. The remaining 40% will be used for the final score after the end of the competition.

# 4 Post-exam work (6-7 points)

You have then until the 15th of October (included) to finish the 2nd part and do the third part.

You can send me an update of your work by mail by that date.

Kaggle competition will also be opened until that date. You'll be allowed to make only 5 submissions a day.