ISTIC 2024/2025
M2 MIAGE October 2024

Module FST - Évaluation TP

1 Objectifs et instructions

Etant donnée une tâche de classification supervisée définie par la donnée d'un ensemble d'exemples étiquetés, votre objectif est de déterminer le meilleur classifieur pour cette tâche, selon la méthodologie présentée durant le module. Vous pouvez (et devez) utiliser pour cela des fonctions vues en TP.

Les données sont dans le fichier `arabic_dataset.csv` sur le Teams. Le fichier `script_etu.ipynb` contient des indications utiles. C'est ce fichier que vous devez compléter.

Les données sont ici des images en noir et blanc représentant des caractères manuscrits de l'alphabet arabe. Voici un exemple d'image de ce jeu de données :

Cet image correspond au caractère 'Taa' de l'image suivante (même si ce n'est pas forcément tout à fait ressemblant !)

Dans le jeu de données que vous avez, il y a 1700 images de taille 32*32 pixels. Il n'y a que 7 caractères différents représentés (pas tout l'alphabet) :

- label 1 : alif
- label 2 : ba
- label 6 : hha
- label 12 : sin
- label 16 : taa

— label 18 : ayin
— label 22 : kaf

L'objectif de ce TP est de construire des classifieurs qui semblent performants pour la tâche de reconnaissance de ces caractères. Ensuite, vous appliquerez les classifieurs que vous avez choisis à un autre jeu de données de 500 nouvelles images pour lesquelles vous ne connaissez pas les labels, et vous pourrez aller regarder la performance de vos prédictions sur Kaggle. Ces 500 nouvelles images sont dans le fichier `competition.csv`. **Vous ne vous en servirez que pour prédire la classe de ces images avec les différents classifieurs que vous aurez sélectionnés auparavant.**

Cet examen comporte trois parties. Dans la première partie, vous ferez une description et une analyse rapide du jeu de données que vous avez. Dans la deuxième partie, vous devrez utiliser les données brutes fournies dans le fichier `arabic_dataset.csv` pour créer des classifieurs en utilisant les différentes familles de classifieurs vus en cours. Dans la troisième partie, vous appliquerez la représentation HOG aux images afin de créer des classifieurs plus performants.

Dans les deux dernières parties, les méthodes suivantes doivent être étudiées selon l'ordre fourni ci-dessous :

1. Arbres de décision
2. SVM
3. k-NN
4. Forêts d'arbres aléatoires
5. Régression logistique

2 Rendu à l'issue de la séance d'aujourd'hui (13-14 points)

Vous serez évalués aujourd'hui sur la première partie, ainsi que les 3 premières familles de classifieurs appliquées aux données brutes (pas de HOG aujourd'hui). Vous devrez sélectionner :
— 1 arbre adapté à ces données brutes
— 3 différents SVM (car 3 types de SVM vus en cours)
— 1 modèle de plus-proches-voisins
Et vous devrez utiliser ces modèles choisis pour prédire le jeu de compétition et soumettre ces différentes prédictions sur Kaggle
A l'issue de la séance, vous devrez rendre la synthèse des résultats de votre étude sur laquelle devra figurer :
— La description et l'analyse du jeu de données et donc de la tâche de classification
— La description précise de la méthodologie mise en oeuvre
— Les résultats des évaluations des différentes méthodes restituées **obligatoirement sous forme de tableau**, ainsi (si besoin) qu'une analyse rapide des résultats obtenus.
— Votre conclusion globale

La synthèse de vos résultats pourra être faite directement dans votre script notebook ou bien sur un fichier pdf séparé. Vous enverrez également le code source utilisé par mail à : `simon.malinowski@irisa.fr`.

Le code doit impérativement être **dans un fichier unique, nettoyé, exécutable séquentiellement et commenté**. En l'absence d'une de ces conditions, le code ne sera pas regardé.

La qualité de la méthodologie et des rendus seront évalués, et non la quantité des méthodes testées.

3 Soumission de vos prédictions sur Kaggle

Le lien vers la compétition sur Kaggle est dans le fichier `kaggle-link.txt` sur le teams. Cela nécessite de créer un compte sur Kaggle : `https://www.kaggle.com`.

Après avoir étudié une méthode de la liste ci-dessus, vous pouvez prédire les classes des individus du fichier `competition.csv`, et soumettre ces prédictions sur Kaggle (bouton `Submit predictions`). N'oubliez pas de mettre en commentaire à quoi correspondent ces prédictions (méthode, paramètre, ...). Un exemple de soumission de prédictions sur Kaggle est donné dans le fichier `script_etu.ipynb`.

Le score affiché dans le leaderboard est calculé en utilisant seulement 60% des individus de `competition.csv`. Les 40% restant seront utilisé pour calculer votre score final à l'issue de la compétition. Vous avez le droit de choisir 2 soumissions parmi toutes celles que vous allez faire (sur la page de la compétition). La meilleure de ces 2 soumissions correspondra à votre score final.

4 Travail post-épreuve (6-7 points)

Pour finir la deuxième partie et la troisième partie de ce TP, vous avez jusque dimanche 13 Octobre soir.

Vous pourrez m'envoyer la suite de votre travail (jupyter notebook, rendu pdf).

Vous pourrez soumettre pour chaque méthode vos meilleurs résultats sur Kaggle jusqu'au Dimanche 13 Octobre soir. Vous avez droit à 5 soumissions par jour.