

RENDU TP EXAMEN

Objectif :

Déterminer le meilleur modèle pour ensuite prédire le label du jeu de données compétition.

Méthode :

étape 1 : pour chaque famille d'algorithme (Arbres de décision, SVM, k-NN, Forêt d'arbres aléatoires, Régression logistique, Réseaux de neurones) trouver le meilleur modèle.

étape 2 : déterminer les performances de ce modèle sur les données de compétition.

NB : seuls les modèles jugés susceptibles de battre mon meilleur score kaggle antérieur seront testés sur le jeu de données compétition.

description et analyse du jeu de données

taille des données :

Nous avons **1700** lignes (caractères) et **1025** colonnes (1024 variables descriptives et la variable cible). Il y a beaucoup trop de variables dans ce jeu de données. Il faudra peut-être faire une ACP ou une transformation HOG pour essayer de mieux synthétiser ces données. Pour l'instant utilisons toutes les variables pour notre analyse

répartition des labels

```
> table(dataset$label)
```

1	2	6	12	16	18	22
238	238	250	255	233	241	245

Nous avons donc pour chaque valeur de label(1ère ligne) le nombre d'occurrence (seconde lignes). d'après la capture, chaque classe est plutôt bien représentée.

Partie 1 : utilisation des images brutes

séparation Apprentissage/Validation/test

```
# taille des ensembles|
dim(train) #1190 lignes et 1025 colonnes
dim(valid) #255 lignes et 1025 colonnes
dim(test) #255 lignes et 1025 colonnes
```

nous avons bien un total de 1700 lignes.

Arbres de décision :

Récapitulatif

modèle	bonne prédiction apprentissage	bonne prédiction en validation	bonne prédiction en test
sans elagage	1190/1190 soit 100%	130/255 soit 50,98%	non testé
avec elagage (cp = 0.0005)	1185/1190 soit 99,57%	130/255 soit 50.98%	137/255 soit 53,72 %

Conclusion : après avoir essayé plusieurs valeurs de cp, c'est avec la valeur cp = 0.0005 que j'obtiens le meilleur compromis apprentissage/validation. j'obtiens un score kaggle de 0.54 soit taux de bonnes réponse de 54% sur les données de la compétition

SVM :

Récapitulatif

modèle	bonne prédiction en apprentissage	bonne prédiction en validation	bonne prédiction test
SVM linéaire	1190/1190 soit 100%	254/255 soit 99%	255/255 soit 100%
SVM gaussien	1092/1190 soit 91.76%	188/255 soit 73.73%	190/255 soit 74.5%
SVM poly	1189/1190 soit 99.91%	176/255 soit 69.01%	182/255 soit 71.37%

Conclusion :

Mauvaise surprise. En testant ce modèle sur les données compétition j'obtiens un score de moins de 10% de bonne réponse. Cela s'explique peut-être par le fait que les SVM apprennent des frontières et non des classes ? (a creuser plus tard)

K-NN

Récapitulatif

modèle	bonne prédiction validation	bonne prédiction test
k = 1	177/255 soit 69,41%	pas testé
k = 2	173/255 soit 67,84%	182/255 soit 71,37%
k = 4	173/255 soit 67,84%	170/255 soit 66,66%

conclusion : à ce stade de mon analyse ce modèle (k=2) est le modèle le plus performant de tous ceux déjà étudié. utilisons le pour prédire les données de compétition.

Forêts Aléatoire :

récapitulatif :

modèle	bonne prédiction en apprentissage	bonne prédiction validation	bonne prédiction test
Forêt 2 arbres	956/1190 soit 80,33%	117/255 soit 45,88%	non testé
Forêt 500 arbres	1190/1190 soit 100%	190/255 soit 74.5%	201/255 soit 78.82%

Conclusion :

Après plusieurs essais, le modèle forêt de 500 arbres a donné le meilleur score en validation, plus de 74%.

Pour l'instant c'est le modèle avec lequel j'obtiens le meilleur score sur le jeu de données compétition avec un score de 0.68 soit un taux de bonnes prédictions de 68%.

Régression logistique

Récapitulatif :

modèle	bonne prédiction apprentissage	bonne prédiction validation	bonne prédiction test
modèle avec 10 itérations et 200 individus.	65/200 soit 32,5%	65/255 soit 25,49%	non testé
modèle avec 100 itération et tout l'ensemble d'apprentissage	452/1190 soit 37.98%	85/255 soit 33.33%	90/255 soit 35.29%

Conclusion :

Les performances de ce modèle sont largement inférieures à celles du modèle de forêt d'arbres aléatoire vu précédemment. Alors inutile de l'utiliser pour essayer de prédire les données du fichier compétition.

Réseaux de neurones

NB : séparation Apprentissage/Test uniquement. pas de validation car une partie des données d'apprentissage utilisés pour calculer les performances en validation

récapitulatif

modèle	accuracy apprentissage	accuracy validation	accuracy test
modèle 1 (600 epochs, taille des lots : 10)	1 soit taux de bonne réponse = 100%	0.68 soit taux de bonne réponse de 68%	0.67 soit taux de bonne réponse de 67%
modèle 2 (100 epochs, taille mini lot : 10)	1 soit taux de bonne réponse = 100%	0.7206 soit taux de bonne réponse de 72.06%	0.7205 soit taux de bonne réponse de 72.5%

conclusion : j'obtiens un score kaggle de 0.64666 soit 64,66% de bonnes prédictions avec le meilleur modèle de réseaux de neurones.

Jusqu'ici le modèle de forêt aléatoire semble avoir la meilleure performance sur le jeu de données compétition.

Partie 2 : Utilisation de la représentation HOG

Forêt aléatoire :







Récapitulatif :

modèle	bonnes prédiction apprentissage	bonne prédiction validation	bonne prédiction test
ntree = 600 (cell = 2, ori = 6)	1190/1190 soit 100%	219/255 soit 85,88%	233/255 soit 91,37%
ntree = 1000 (cell = 2, ori = 6)	1190/1190 soit 100%	218/255 soit 85,49%	232/255 soit 90,98%
ntree = 100 (cell = 3, ori = 8)	1190/1190 soit 100%	230/255 soit 91,76%	238/255 soit 93,33%
ntree = 400 (cell = 3, ori = 8)	1190/1190 soit 100%	232/255 soit 90,98%	238/255 soit 93,33%
ntree = 800 (cell = 3, ori = 8)	1190/1190 soit 100%	232/255 soit 90,98%	241/255 soit 94,50%

Conclusion :

l'augmentation de la valeur de la variable orientation (ori) a permis de mieux synthétiser les données. Après plusieurs essais sur cette nouvelle transformation, il semble que le modèle avec 800 arbres donne le meilleur score. voyons ses performances sur les données de compétition.

je me trouve à ce stade en tête de la compétition kaggle avec un score de 89.66% grâce à ce nouveau modèle (**ntree = 800 ,cell = 3, ori = 8**).

#	Team Name	Notebook	Team Members	Score ?	Entries	Last
1	N'dri MENAN			0.89666	7	5m
Your Best Entry ↑						
Your submission scored 0.89666, which is an improvement of your previous score of 0.84000. Great job!					Tweet this!	
2	Pierre Delaunay			0.88333	20	2h
3	William ZOUNON			0.87000	17	3h
4	Kirti SWEENARAIN			0.85333	16	3h
5	Yannick ZOHOU			0.84666	19	4h

Arbres de décision :

Récapitulatif

modèle	bonnes prédiction apprentissage	bonne prédiction validation	bonne prédiction test
sans elagage	1190/1190 soit 100%	183/255 soit 71,76%	non testé
avec elagage (cp = 0.0005)	1184/1190 soit 99,49%	189/255 soit 74,11%	193/255 soit 75,68%
avec elagage (cp = 0.0015)	1108/1190 soit 93,10	189/255 soit 74,11%	197/255 soit 77,25%

conclusion :

Le modèle arbre de décision avec transformation HOG permet d'obtenir des performances supérieures à celui sans transformation HOG. Cependant, inutile de l'utiliser pour prédire nos données de compétition car sa performance est largement inférieure au modèle de forêt aléatoire (avec transformation HOG). A ce stade de mon analyse, le modèle de forêt d'arbres aléatoire (avec transformation HOG) est donc toujours le meilleur modèle.

K-NN

Récapitulatif :

modèle	bonne prédictions validation	bonne prédiction test
k = 1, cell = 3 et ori = 8	223/255 soit 87,45%	non testé
k = 4, cell = 3 et ori = 8	224/255 soit 87,84%	220/255 soit 86,27%
k = 8, cell = 3 et ori = 8	224/255 soit 87,84%	222/255 soit 87,05%

Conclusion : modèle intéressant mais performance inférieure au modèle forêt (avec transformation HOG) pas donc nécessaire de l'utiliser pour prédire les données de compétition.

SVM

Récapitulatif :

modèle	bonne prédiction apprentissage	bonne prédictions validation	bonnes prédiction test
SVM linéaire avec cost = 5	1169/1190 98,23%	214//255 soit 83,92%	non testé
SVM gaussien cost = 1, gamma =0.02	1158/1190 97,31%	227/255 soit 89,01%	non testé
SVM logistique cost = 1, degree = 4, coef = 1	1186/1190 soit 99,66%	231/255 90,58%	230/255 90,19%

Conclusion : les modèles SVM avec transformation HOG sont meilleurs que ceux sans transformation. cependant très mauvaise performance obtenue sur les données de compétition.

Régression logistique

Récapitulatif

modèle	bonne prédiction apprentissage	bonne prédiction validation	bonne prédiction test
avec 100 itérations	1016/1190 soit 85,37%	206/255 soit 80,73%	209/255 81,96%

Conclusion : le modèle de régression logistique avec transformation HOG permet d'obtenir une performance meilleure comparée à celle sans transformation HOG. Cependant pas assez puissant pour battre mon score kaggle à ce stade de la compétition. Inutile donc de l'utiliser pour faire des prédictions sur les données de compétition.

Réseaux de neurones

récapitulatif

modèle	accuracy apprentissage	accuracy validation	accuracy test
modèle 1 (600 epochs, taille des lots : 10)	1 soit taux de bonne réponse = 100%	0.8860 soit taux de bonne réponse de 88,60%	0.8765 soit taux de bonne réponse de 87,65%
modèle 2 (1000 epochs, taille mini lot : 256)	1 soit taux de bonne réponse de 100%	0.8860 soit taux de bonne réponse de 88.06%	0.8676 soit taux de bonne réponse de 86.67%

conclusion : les modèles neuronaux avec transformation HOG permettent d'obtenir de meilleures performances comparé à ceux sans transformation HOG. On obtient jusqu'à 87% de bonnes réponses. On aurait peut-être pu faire mieux avec plus de données. Malheureusement nous n'en avons pas assez de ce cas de figure. Alors on reste toujours en dessous de la performance du modèle forêt avec transformation HOG pour lequel j'obtiens un taux de bonne prédiction de plus de 94%

Conclusion générale

Au terme de mon analyse, il en ressort que pour prédire efficacement des caractères, il faut d'abord faire une transformation HOG du jeu de données initial. Ensuite, en fonction de la taille du jeu de données, utiliser soit un modèle basé sur les réseaux de neurones (si taille des données conséquente). soit un modèle basé sur les forêts d'arbres aléatoires.