Machine Learning partie 1

HEPIA Laroche David

Objectif

Ce projet a pour but de faire pratiquer les méthodologies caractéristiques de l'approche "Data Science" pour la classification de données. Il se déroulera en trois parties spécifiques associées aux modèles suivants :

1. modèles de base vus au cours ;
2. ensembles ;
3. réseaux convolutifs.

Fonctionnement du script

1. L'exécution du script prend en paramètres optionnels les données à traiter wisc pour `wisconsin.dat` et dj pour `dow_jones_index.data`, pour les données à choix.

2. Il prend aussi les méthode de classification PPV pour le plus proche voisin, AD pour les arbres de décision, PMC pour le perceptron multi-couches et finalement SVM pour la machine à vecteurs de support.

3. Le script va donc calculer pour chaque ensemble de données et de classificateurs indiqués avec ses paramètres (définis en durs dans le programme pour des raisons de praticité) puis produire le fichier .csv correspondant.

Ceci permet de comparer les résultats des différents modèles mais aussi quels paramètres influencent le mieux leur taux de classification (TCC).

Exemple de commande :

```
$ python3 dj wisc PMC AD
```

Produit les 4 fichier: `output/dj_pmc.csv`, `output/dj_ad.csv`, `output/wisc_pmc.csv`, `output/wisc_ad.csv` contenant les résultats de wisconsin et du dow jones pour les méthodes du perceptron et des arbres de décision.

Choix des données

J'ai choisi les données de la bourse `dow_jones_index.data`. Presque tous les attributs sont continus. Afin de travailler avec des classificateurs, la classe à prédire a été définie à partir de l'attribut `percent_return_next_dividend`. Ainsi une valeur inférieure à 1 indique une perte et donc appartenant à la classe 0 si elle est supérieure à 1 ceci indique un gain et donc appartenant à la classe 1. Les données manquantes ont été comblées avec la valeur moyenne des exemples. Finalement tous les attributs continus ont été normalisés.

Mesure des performances

Temps de calcul

Les temps d'entraînement par modèle sont plutôt négligeables. Par ailleurs, plus un estimateur possède de paramètres, plus il sera coûteux de trouver le meilleur ajustement, car cela implique de calculer un modèle supplémentaire pour chaque combinaison de paramètres. Il est donc judicieux de les limiter et de les choisir de manière intelligente.

Table 1. Temps de Calcul sur les données wisconsin.dat

Classificateur	Nombre de modèles entraînés	Temps médian pour l'entraînement d'un modèle [msec]	Temps total [secondes]
AD	32076	1.2	247
PPV	50	0.9	5.11
PMC	432	731	774

Table 2. Temps de Calcul sur les données `dow_jones_index.data`

Classificateur	Nombre de modèles entraînés	Temps médian pour l'entraînement d'un modèle [msec]	Temps total [secondes]
AD	32076	3.2	525
PPV	50	0.9	5.8
PMC	432	837	868

Nous pouvons observer que l'AD et le PPV sont peu coûteux en calculs (temps médian par modèle) contrairement au PMC qui est assez lourd en comparaison.

Validation croisée

Le nombre de sous échantillons que l'on choisis pour la validation croisée augmente grandement les temps de calculs car nous multiplions le nombre de résultats par le nombre d'échantillons à entraîner. J'ai donc commencé avec une division en 3 sous-échantillons afin de trouver les paramètres et ai terminé avec 10 segments comme demandé pour obtenir des résultats plus réalistes en les moyennant.

Taux de classification

Voici le meilleur taux de classification correct (TCC) moyen en calculant la moyenne des résultats obtenus de la validation croisée avec dix sous-échantillons (méthode : k-fold cross-validation).

Table 3. Meilleur TCC par Classificateur sur les données wisconsin.dat

Classificateur	TCC moyen test	TCC moyen entraînement
AD	0.88	0.88
PPV	0.97	0.98
PMC	0.97	0.98

Table 4. Meilleur TCC par Classificateur sur les données `dow_jones_index.data`

Classificateur	TCC moyen test	TCC moyen entraînement
AD	0.92	0.94
PPV	0.88	1
PMC	0.97	0.98

Les performances entre classificateurs sont assez proches. Les TCC des sets d'entraînement sont légèrement supérieurs aux TCC des tests ce qui est attendu car le modèle est entraîné avec le set d'entraînement par définition. Mais la différence de valeurs entre TCC test et TCC entraînement n'est pas assez grande pour indiquer avec assurance un overfitting. Seule la valeur du TCC au-delà de 95 % indiquerai une performance trop grande et donc un overfitting sur nos données malgré l'utilisation de la validation croisée.

Plus Proche Voisin

Paramètres modifiés:

1. Nombre de vosins pour la requête [`n_neighbours`]: nombre impair de 1 à 50
2. Poids pour la requête: [uniform, distance]

Ce classificateur ne possédant que deux paramètres nous intéressant dont un binaire nous pouvons représenter les résultats sur un graphe.

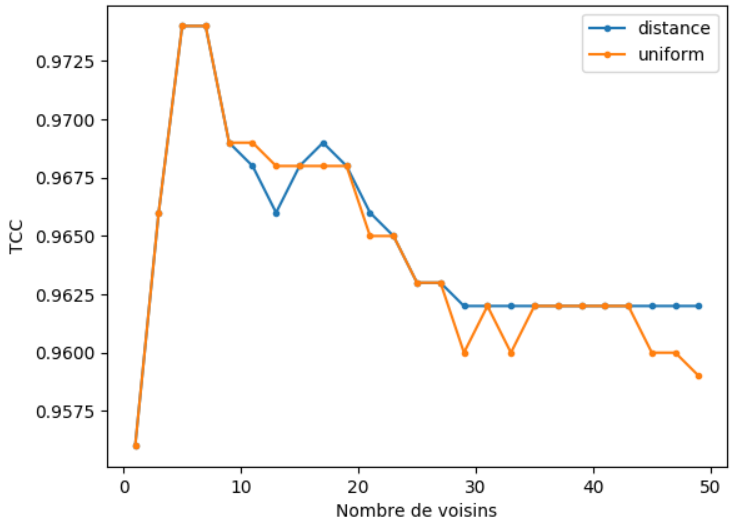

Fig 1. Plot on wisconsin.data

Nous avons un maximum à 5 et 7 voisins, ensuite l'augmentation de ce nombre n'améliore pas la performance. Les dents de scies sur le même nombre de voisins est causé par le mode de détermination de la classe (par distance ou par majorité). Avec ces données, le changement des paramètres influence peu la performance (différence de 1,8 point entre le maximum et le minimum).

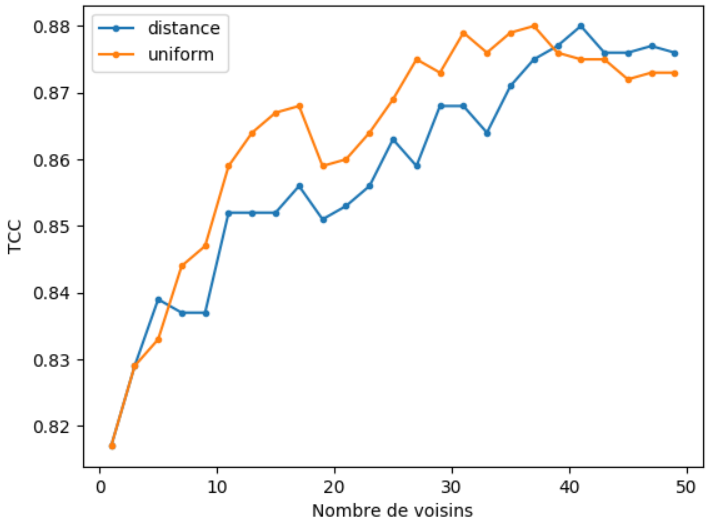

Fig 2. Plot on `dow_jones_index.data`

Cette courbe fait penser à la fonction logarithme, le TCC s'améliore avec l'augmentation de [`n_neighbours`] pour définir la classe mais devient de plus en plus plate.

Arbre de Décision

Paramètres modifiés:

1. Profondeur maximale (`max_depth`) : 1 à 99
2. Nombre d'exemples minimum par feuille (`min_samples_leaf`) : 2 à 20
3. Nombre minimal pour le découpage d'une node (`min_samples_split`) : 2 à 20

Le découpage d'une node se fait avec le meilleur gini par défaut.

Table 5. Meilleurs scores sur wisconsin.dat

TCC Test Moyen	STD	TCC Train Moyen	Fit Time Moyen	Max Depth	Min Samples Leaf	Min Samples Split
0.875	0.03	0.88	0.002	1	2	2
0.875	0.03	0.88	0.002	99	19	19

Par soucis de représentation j'ai omis les 100 modèles (ou lignes) possédant le même TCC Test Moyen (0.875), [`max_depth`] augmentant de 1 à chaque ligne

Nous pouvons observer (ref. Table 5) que [`max_depth`] n'influence pas le TCC test moyen (`mean_test_score`) pour une certaine combinaison des paramètres [`min_samples_leaf`] et [`min_samples_split`].

Il est tout a fait possible que certains paramètres comme [`min_samples_leaf`] arrêtent la croissance de l'arbre avant d'arriver à la profondeur maximale, ainsi [`max_depth`] ne produit aucune amélioration du TCC à partir d'un certain seuil. Nous pouvons avoir le même phénomène avec le nombre d'exemple minimal pour découper une node. Ces paramètres s'influence donc entre eux mais sont utiles pour se prémunir de l'overfitting, le pire des cas étant d'avoir une feuille pour chaque exemple.

Table 6. Meilleurs scores sur `dow_jones_index.data`

TCC Test Moyen	STD	TCC Train Moyen	Fit Time Moyen	Max Depth	Min Samples Leaf	Min Samples Split
0.923	0.09	0.94	0.002	3	2	7
0.923	0.09	0.94	0.002	3	19	19

Par soucis de représentation j'ai omis les 170 modèles (ou lignes) possédant le même TCC Test Moyen (0.923), tous ces modèles possèdent le même [`max_depth`] de 3.

Sur ces données Table 6 c'est la profondeur de l'arbre qui influence le plus les résultats, la profondeur restant à 3 malgré le changement des deux derniers paramètres pour les meilleurs TCC Test Moyen. On peut supposer que peu d'attributs suffisent à trouver la meilleure prédiction, le reste étant du bruit, il serait utile des les négliger dans nos modèles.

Perceptron Multi Couches

Paramètres modifiés:

1. Fonction d'activation (activation) : Identité, Tangente hyperbolique (tanh), Logistique, Unité de rectification linéaire (relu)
2. Alpha : 0.1, 1, 10
3. couches cachées et nombre de neurones y figurant (`hidden_layer_sizes`) : (5, 50), (5, 100), (10, 50), (10,100)
4. Taux d'apprentissage (`learning_rate`) : constant, invscaling, adaptif
5. solveur : lbfgs, sgd, adam

mean_test_score	std_test_score	mean_train_score	mean_fit_time	activation	alpha	hidden_layer_sizes	learning_rate	solver
0.974	0.038	0.982	0.7918	logistic	1	(5, 100)	constant	lbfgs
0.972	0.038	0.978	0.9364	tanh	10	(10, 100)	invscaling	lbfgs
0.972	0.038	0.978	0.8705	tanh	10	(10, 100)	adaptive	lbfgs
0.972	0.038	0.979	0.8291	tanh	10	(5, 100)	invscaling	lbfgs
0.972	0.038	0.979	0.8074	tanh	10	(5, 100)	constant	lbfgs
0.972	0.038	0.979	0.8057	logistic	1	(10, 100)	adaptive	lbfgs
0.972	0.038	0.978	0.5517	tanh	10	(10, 50)	constant	lbfgs
0.972	0.038	0.979	0.4857	tanh	10	(5, 50)	invscaling	lbfgs
0.972	0.038	0.979	0.4714	tanh	10	(5, 50)	constant	lbfgs
0.972	0.036	0.978	0.5675	tanh	10	(10, 50)	invscaling	lbfgs
0.972	0.036	0.982	0.5308	logistic	1	(10, 50)	invscaling	lbfgs

Fig 3. Meilleurs scores sur wisconsin.dat

Pour le PMC nous obtenons les meilleurs résultats avec un alpha de 1 ou 10 et la fonction d'activation tangente hyperbolique le plus souvent et lbfgs comme solveur. Dans notre cas, nombre de de couches cachées et le nombres de neurones y figurant influencent peu les résultats. Je suppose que le alpha de 0.1 systématiquement plus mauvais dans notre cas est causé par la tendance à rester dans des minimas locaux moins bons lors de la descente de gradient.

Le TCC minimum obtenu est de 0.439 ce qui est mauvais. Nous pouvons déduire que la configuration des paramètres est d'autant plus important pour ce classificateur. Cependant, à part les trois paramètres cités avant, il est difficile de déterminer quelle est l'influence des autres paramètres car c'est sans doute leur combinaison qui explique cette variation de résultats.

Variation que nous pouvons observer si nous classons les modèles par ordre de TCC décroissant.

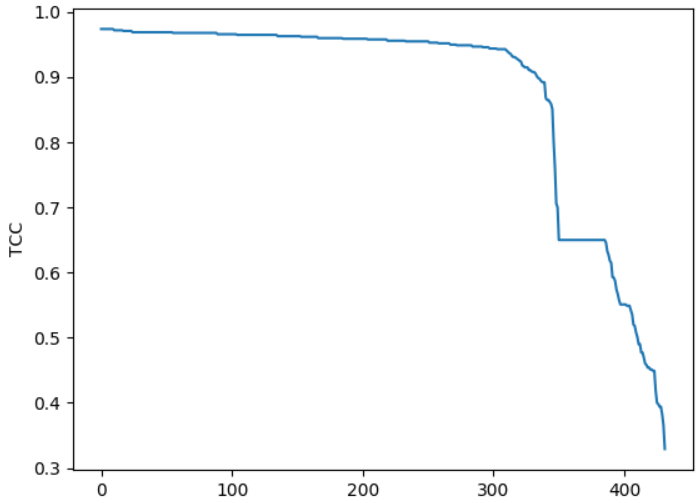

Fig 4. Décroissance du TCC sur wisconsin.data

Nous pouvons remarquer un saut de 0,9 à 0,65 environ de TCC, en étudiant les données, la combinaison des autres paramètres avec la fonction logistique donne systématiquement des résultats moins bon, ainsi qu'une combinaison entre certains paramètres et le solveur invscaling.

mean_test_score	std_test_score	mean_train_score	mean_fit_time	activation	alpha	hidden_layer_sizes	learning_rate	solver
0.974	0.038	0.982	0.8254	logistic	1	(5, 100)	invscaling	lbfgs
0.974	0.034	0.98	0.5005	tanh	10	(5, 50)	constant	lbfgs
0.972	0.044	0.981	0.8786	logistic	1	(10, 100)	constant	lbfgs
0.972	0.042	0.982	0.9145	logistic	1	(10, 100)	invscaling	lbfgs
0.972	0.04	0.982	0.898	logistic	1	(10, 100)	adaptive	lbfgs
0.972	0.038	0.979	0.5043	tanh	10	(5, 50)	invscaling	lbfgs
0.972	0.038	0.979	0.8432	tanh	10	(5, 100)	invscaling	lbfgs
0.972	0.038	0.979	0.8692	tanh	10	(5, 100)	adaptive	lbfgs
0.972	0.038	0.978	0.9345	tanh	10	(10, 100)	constant	lbfgs
0.972	0.033	0.971	1.3111	relu	1	(5, 100)	adaptive	adam

Fig 5. Meilleurs scores sur `dow_jones_index.data`

Il intéressant de relever la similarité des scores et des meilleurs paramètres malgré deux ensembles de données très différents. On observe le même phénomène si l'on classe les modèles par TCC.

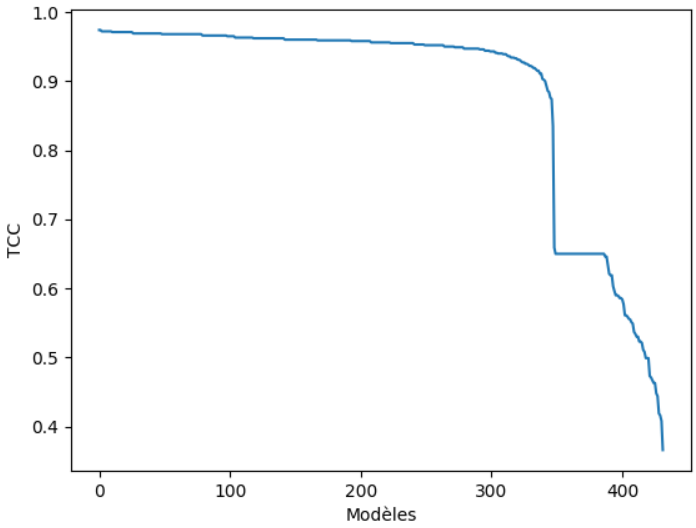

Fig 6. Décroissance du TCC sur`dow_jones_index.data`

Conclusion

Scikitlearn fournit beaucoup d'outils pour améliorer ses modèles de classification ou de régression. Cependant il n'est pas toujours évident de comprendre l'influence des paramètres utilisés, une idée serait d'utiliser une régression sur ces données (donc du machine learning sur du machine learning...) afin de déterminer les paramètres les plus influents en comparant leurs poids. Il serait aussi judicieux d'étudier le domaine de la statistique qui possède des liens forts avec le machine learning, mais le champ est vaste...

References

1. `https://scikit-learn.org`