Résultats de recherche sur la classification supervisée

M. Deker Sylvain[1] Mme. Courdy-Bahsoun Clémence[2]

[1] M1 IGAI Université Paul Sabatier
[2] M1 IGAI Université Paul Sabatier

sylvain.deker@univ-tlse3.fr
clemence.courdy-bahsoun@univ-tlse3.fr

Résumé

Cet article vise à comparer les avantages et les désavantages des classifications supervisées par bayésienne avec gaussienne et par la méthode des k-plus proches voisins, en s'appuyant sur les résultats obtenus sur un même jeu de données.

Mots Clef

Apprentissage, classification bayésienne, classification k-ppv.

1 Introduction

La classification bayésienne avec gaussienne et l'approche par la méthode des k-ppv sont des classifications supervisées dont l'objectif est de répartir les données par classe avec un taux d'erreur minimisé. Les tests sont menés sur un jeu de données contenant une centaine d'occurences de la prononciation d'une dizaine de voyelle aau format csv.
A l'aide des travaux réalisés précédamment et en adaptant un peu le code nous avons pu implémenter le modèle de Bayes. L'implémentation pour une approche par les k-ppv, le sujet nous imposait l'utilisation de *Scikitlearn*.
Dans les tests menés les données fournies ont été repartie de telle sorte à avoir 80% de données d'apprentissage et 20% de données de tests. Dans les démarches d'apprentissage le choix d'un taux de données d'apprentissage élevé favorise l'obtention de meilleurs résultats, ici une meilleur prédiction de la répartition par classe.

2 Classification bayésienne avec Gaussienne

La classification bayéssienne avec Gaussienne est une minimisation de la vraisemblance qui vise à prédire la classe d'appartenance la plus probable d'une donnée. Néanmoins avec des données réelles il est difficile d'obtenir des résultats satisfaisants car les données ne sont pas simples. Ainsi pour améliorer le modèle on tient compte de la probabilité a priori en applicant la loi de Bayes.
Dans les tps précedant, nous avions déjà implémenté en python l'algorithme de prédiction permettant de classer les données fournies. Ici nous somme reparti de ce code qui était basé sur une simplification de la vraisemblance en log-vraisemblance. C'est-à-dire que la vraisemblance qui est une expression exponentielle a été simplifié par l'application du logarithme népérien afin d'obtenir un calcul un peu plus optimisé. De plus toutes les constantes ne dépendant pas des données ont été supprimé, n'apportant pas une précision indispensable dans la reconnaissance des données.

3 Classification k-NN

La méthode des K plus proches voisins (K-NN) a pour but de classifier des points cibles (classe méconnue) en fonction de leurs distances par rapport à des points constituant un échantillon dápprentissage (cést-à-dire dont la classe est connue a priori). Le calcul de la distance peut dépendre de l'expérience menée, Scikitlearn se base sur un calcul euclidien de la distance entre les données de l'échantillon, placées dans un repère après un traitement par une **A**nalyse en **C**omposantes **P**rincipales, et les centroïdes de la base d'apprentissage. Scikit learn est une bibliothèque libre pour python dont certaines fonctionnalités ont été utilisé pour implémenter la classification par k-NN. En effet toutes les fonctions appelé sur la variable n de la fonction k_NN sont des fonctions appelés pour la classe KNeighborsClassifier, une classe de la bibliothèque permettant d'effectuer les traitements nécéssaire à une classification par plus proche voisin.

Algorithm 1 k_NN

Require: train_data, train_labels, test_data, test_labels
Ensure: m : matrice de confusion
p : precision
n = neighbors.KNeighborsClassifier(n_neighbours,weights = 'distance')
n.fit(train_data,train_labels) %apprentissage
y = n.predict(test_data) mconf = matriceConf(test_labels,y) %calcul matrice de confusion, fonction rédigé
return (mconf[0], mconf[1])

4 Evaluation par validation croisée

La validation croisée est une méthode déstimation de la fiabilité dún modèle fondé sur la technique d'échantillonage. Les essembles de données d'apprentissages sont les fichiers de test data2.csv, data3.csv et data12.csv sur lequel on peut entrainer le modèle. Nous appelons "leanring ratio" la proportion des données d'un fichier dédié à lápprentissage, la partie restante sert déchantillons da validation. Cette méthode de test s'appelle "testset validation" ou "holdout method".

5 Résultats

Nous rappelons que le "leanring ratio" est la proportion des données d'un fichier dédié à l'apprentissage, la partie restante servant d'échantillons da validation. Pour evaluer la performance d'un algorithme, nous disposons de deux métriques, le premier étant le temps nécéssaire au processus d'apprentissage et de validation, et la second est la précision avec laquelle les résultats sont fiables. Une série de 50 tests ont été effectués sur les fichiers tests data2.csv, data3.csv et data12.csv pour les learning ratio suivant : 0.2 , 0.4, 0.6 et 0.8. Les résultats ci-dessous présentent les moyennes sur ces 50 tests.

Pour reproduire les resultats observés il suffit de faire varier les paramètres dans le programme (cad le learning rate, nombre de test, affichage etc...). Par defaut la fonction testmain est configuré pour tester 50 fois l'opération suivante : calcul de la durée et de la précision des algorithmes KNN et Bayes pour un learning ratio fixé. En redirigeant la sortie standard vers un fichier (ex : python3 main.py > resultat.csv) il suffit ensuite de copier/coller correctement les résultats dans le tableur resultat_test.ods pour faire apparaitre les résultats. Les FIGURE 1, FIGURE 2, FIGURE 3 et FIGURE 4 ont été réalisé avec le logiciel GNUPlot avec les commandes suivantes :

5.1 Matrice de confusion

Afin d'analyser et de pouvoir comparer les résultats obtenus par les deux méthodes étudiées ici, on calcule pour chaque échantillon la matrice de confusion, permettant de déduire un taux d'erreur. Cette matrice est construite à l'aide d'une classe présentes dans la bibiliothèque Scikit-learn et en appliquant les calculs nécessaire pour calculer le taux d'erreur.

5.2 Jeu de test

Les jeux de test fournit montrent des classe distinctes les unes des autre pour les fichiers data2.csv (voir FIGURE 1) et pour le fichier data3.csv (voir FIGURE 2, FIGURE 3 , FIGURE 4)

5.3 Observations

La FIGURE 5 nous montre que plus la proportion des données liés à l'apprentissage est basse (cad plus la pro-

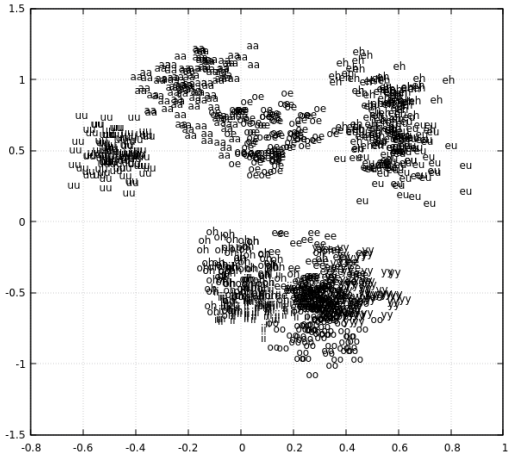

FIGURE 1 – Jeu de test data2.csv

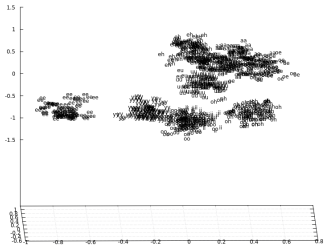

FIGURE 2 – Jeu de test data3.csv vue 1

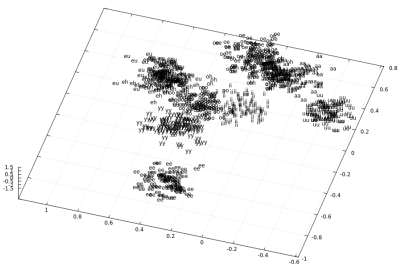

FIGURE 3 – Jeu de test data3.csv vue 2

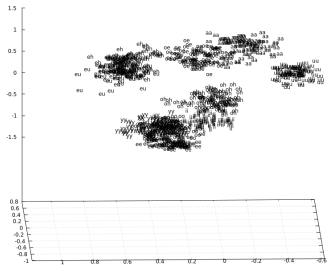

FIGURE 4 – Jeu de test data3.csv vue 3

portion lié à la validation est haute) plus le temps d'execution global de l'algorithme est long (cad temps d'apprentissage + temps de validation). La restitution des données pour la validation demande plus de temps de calcul que pour l'apprentissage. Cette difference de durée est observé aussi avec la méthode KNN (VOIR FIGURE 6) mais de manière moins drastique. Le pic correspondant à un learning ratio de 0.8 sur le jeu de test data12.csv de la FIGURE 6 reste inexpliqué.

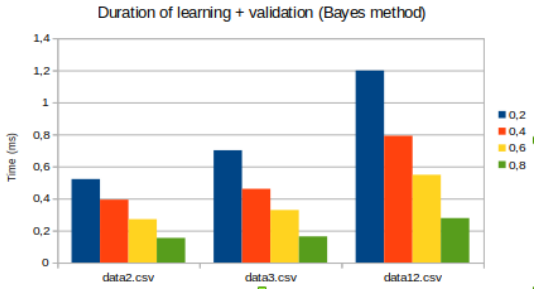

FIGURE 5 – Mesure du temps de l'algorithme Bayes sur les différents jeux de test pour chaque learning rate observé

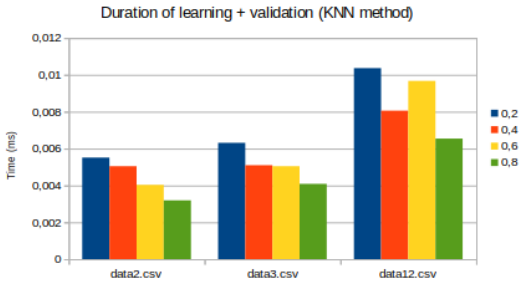

FIGURE 6 – Mesure du temps de l'algorithme KNN sur les différents jeux de test pour chaque learning rate observé

Les FIGURE 7 et FIGURE 8 montre que sur ces jeux de test les méthode KNN et Bayes sont équivalentes, en effet la différence de précision entre ces deux méthode ne sont pas significative. Des données chiffrés sont disponible dans les FIGURE 9, FIGURE 10, FIGURE 11 et FIGURE 12.

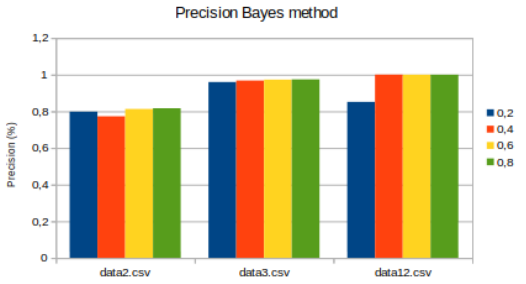

FIGURE 7 – Mesure du la précision de l'algorithme Bayes sur les différents jeux de test pour chaque learning rate observé

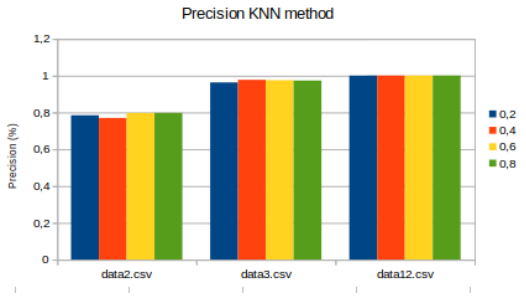

FIGURE 8 – Mesure du la précision de l'algorithme KNN sur les différents jeux de test pour chaque learning rate observé

Learning ratio = 0.2.

20 % des données ont été utilisé pour l'apprentissage de facons aléatoire pour chaque jeu de test (fichier.csv). Les données pour chaque jeu de test est une moyenne de 50 tests réalisés dans le mêmes conditions. Dans ce contexte nous observons que l'algorithme KNN est environ 40 fois plus rapide de celui de Bayes. La précision reste stable, KNN est très légerement moins précis que Bayes pour les deux premiers jeux de test (voir FIGURE 9).

	Bayes		KNN		Evolution from Bayes to KNN	
Files test	Times	Precision	Times	Precision	Time Acceleration	Delta Precision
data2.csv	0,1538571166992	0,8165	0,0031902837753	0,7976	48,22678918047	-0,0189
data3.csv	0,1633808898926	0,9739	0,0040844964981	0,9733	40,000252165294	-0,0006
data12.csv	0,2776164579391	0,9997	0,0065380048752	1	42,481953338842	0,0003

FIGURE 9 – Resultat pour 20% des données dédiées à l'apprentissage

Learning ratio = 0.4.

40 % des données ont été utilisé pour l'apprentissage de facons aléatoire pour chaque jeu de test (fichier.csv). Les données pour chaque jeu de test est une moyenne de 50

tests réalisés dans le mêmes conditions. Dans ce contexte nous observons que l'algorithme KNN est environ 60 fois plus rapide de celui de Bayes. La précision reste stable, KNN est très légerement plus précis que Bayes pour les deux derniers jeux de test (voir FIGURE 10).

	Bayes		KNN		Evolution from Bayes to KNN	
Files test	Times	Precision	Times	Precision	Time Acceleration	Delta Precision
data2.csv	0,270641374588	0,81255	0,00402904033366	0,7971	67,1726644560365	0,01545
data3.csv	0,3280623006821	0,9727	0,0050398731232	0,9749	65,093364984512	0,0022
data12.csv	0,547613954544	0,9995	0,0096665334702	1	56,650500019978	0,0005

FIGURE 10 – Resultat pour 40% des données dédiées à l'apprentissage

Learning ratio = 0.6.

60 % des données ont été utilisé pour l'apprentissage de facons aléatoire pour chaque jeu de test (fichier.csv). Les données pour chaque jeu de test est une moyenne de 50 tests réalisés dans le mêmes conditions. Dans ce context nous observons que l'algorithme KNN est environ 80 fois plus rapide de celui de Bayes. La précision reste stable, KNN est très légerement plus précis que Bayes pour les deux derniers jeux de test (voir FIGURE 11).

	Bayes		KNN		Evolution from Bayes to KNN	
Files test	Times	Precision	Times	Precision	Time Acceleration	Delta Precision
data2.csv	0,3923126220703	0,8080333333333	0,0050431013107	0,7954333333333	77,791937519797	-0,0126
data3.csv	0,4601988983154	0,9706666666667	0,005856089592	0,973	78,584675163727	0,0023333333333
data12.csv	0,790948343277	0,9994	0,0108628463745	1	72,812255278951	0,0006

FIGURE 11 – Resultat pour 60% des données dédiées à l'apprentissage

Learning ratio = 0.8.

80 % des données ont été utilisé pour l'apprentissage de facons aléatoire pour chaque jeu de test (fichier.csv). Les données pour chaque jeu de test est une moyenne de 50 tests réalisés dans le mêmes conditions. Dans ce contexte nous observons que l'algorithme KNN est environ 110 fois plus rapide de celui de Bayes. La précision reste stable, KNN est très légerement plus précis que Bayes pour les deux derniers jeux de test (voir FIGURE 12).

	Bayes		KNN		Evolution from Bayes to KNN	
Files test	Times	Precision	Times	Precision	Time Acceleration	Delta Precision
data2.csv	0,5208987569809	0,79775	0,005506310463	0,7845	94,600324570451	-0,01325
data3.csv	0,7010438108444	0,9596	0,0063081979752	0,96305	111,13218285239	0,00345
data12.csv	1,19847310199	0,8511	0,0103621292114	1	115,65896214433	0,1489

FIGURE 12 – Resultat pour 80% des données dédiées à l'apprentissage

6 Conclusion

Les tests menés sur trois jeux de données différents, et sur une cinquantaine d'éxécution nous permet de mettre en evidence la corrélation entre la taille de la base d'apprentissage et le temps d'éxécution des algorithmes de classification supervisées étudiés est important. En effet on remarque, et ce de manière plus probante sur l'application de la méthode de Bayes, que le temps d'éxécution tant à diminuer pour des bases d'apprentissage fournies généreusement. Les tests menés montre également que pour un même ratio de partage des données, l'algorithme K-NN est beaucoup plus performant en temps d'éxécution que l'algorithme de Bayes avec gaussienne. Une explication possible à ce résultat et que les fonctions utilisé par K-NN appartenant à une bibliothèque propre à python, sont des fonctions optimisés au maximum. A cela peu s'ajouter un manque d'optimisation lors de la rédaction des fonctions pour la méthode par bayésienne, en exploitant pas suffisament les optimisations prévu dans le langage python par un manque de connaissance. On peut seulement déduire que dans le contexte fournit il est plus interessant d'opter pour la méthode des k plus proches voisins . Ce raisonnement est encouragé par les résultats obtenus lors des tests de précision qui se valent pour les deux méthodes. D'après nos résultats on peut en conclure que pour de grands jeux de données il est plus interessant d'opter pour la méthode des k-NN, et eventuellement favoriser la méthode par approche bayésienne pour des petits jeux de données. Cependant il s'agit de résultats expérimentaux. De plus il existe de nombreuses autres méthodes de classification supervisées qui n'ont pas été étudiés ici et qui auront probablement elles aussi leurs propre avantage en fonction du contexte, comme par exemple les resaux de neurones.