Université Lumière Lyon 2 - ICOM

Projet Model Based Learning

Comparaison entre l'algorithme de Classification EM et d'autres méthodes de classification sur MNIST

Etudiants :
Mouhamadou Mansour LO
Mame Libasse MBOUP

Encadrant :
Julien JACQUES

2 avril 2022

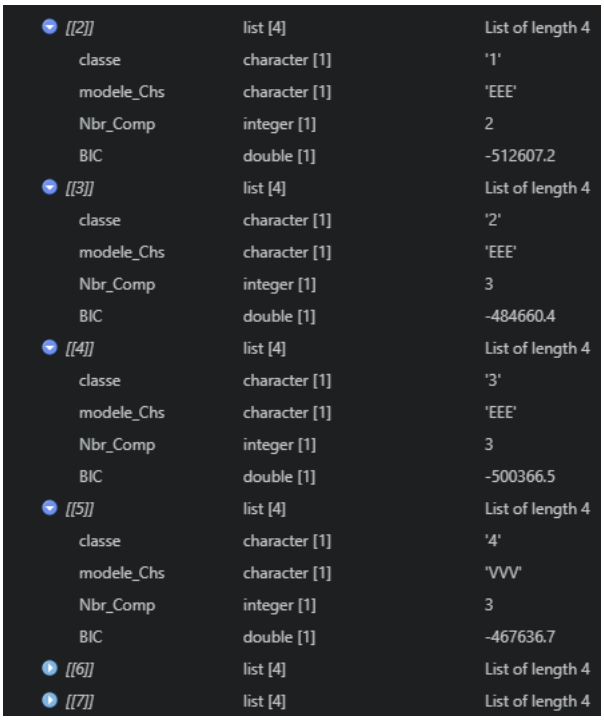
Le Jeu de données **MNIST** est un jeu de données de chiffre écrits a la main utilisé principalement pour la détection de chiffre (1 a 9) en apprentissage supervisé.

Ce jeu de données contient 60000 observations et 785 variables pour le fichier de Train et 10000 observations et 785 variables pour le fichier Test. Il est disponible sur le site de Yann Lecun

Dans le cadre du Projet de Model Based Learning , un modèle de mélange ou chaque classe peut être modélisée lui même par un mélange de Gaussiennes a été développé .

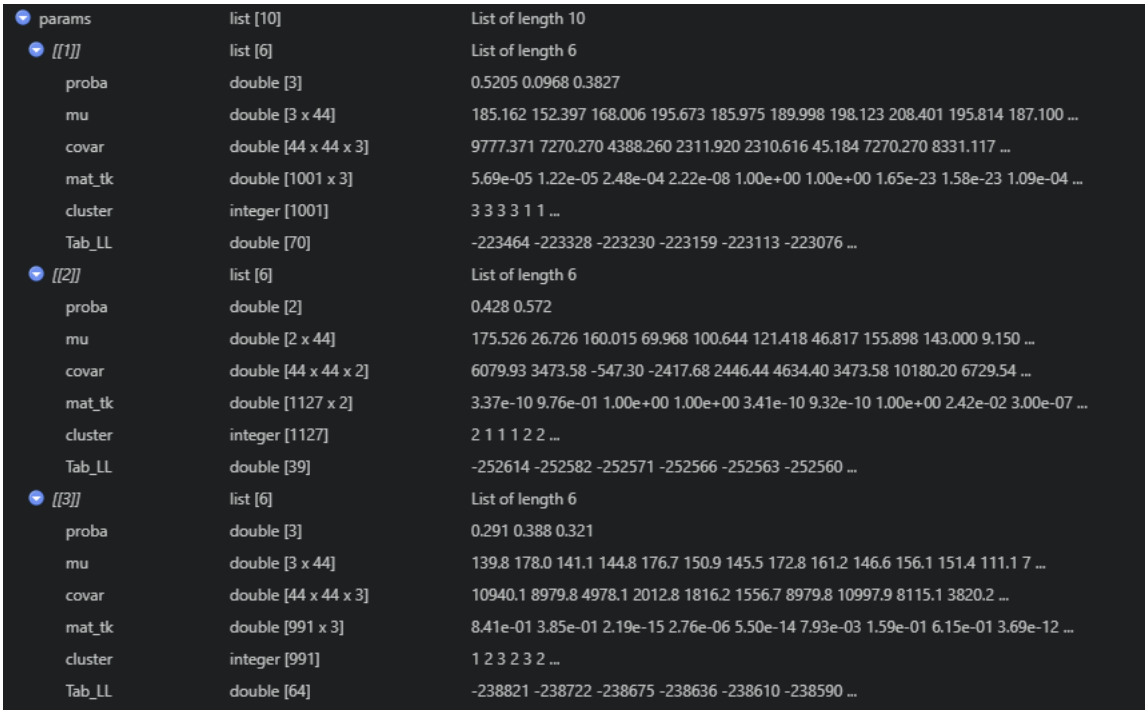
Nous allons dans ce qui suit d'abord présenter les paramètres obtenus après inférence puis comparer avec d'autres modèles.

Les Infos du modèle obtenus pour chaque classe



Entre autres : la valeur de la classe , le modele choisi par le **critere BIC**, le nombre de composants par **classe** , le BIC **minimal**.

Les Paramètres du modèle obtenus pour chaque classe



params	list [10]	List of length 10
[[1]]	list [6]	List of length 6
proba	double [3]	0.5205 0.0968 0.3827
mu	double [3 x 44]	185.162 152.397 168.006 195.673 185.975 189.998 198.123 208.401 195.814 187.100 ...
covar	double [44 x 44 x 3]	9777.371 7270.270 4388.260 2311.920 2310.616 45.184 7270.270 8331.117 ...
mat_tk	double [1001 x 3]	5.69e-05 1.22e-05 2.48e-04 2.22e-08 1.00e+00 1.00e+00 1.65e-23 1.58e-23 1.09e-04 ...
cluster	integer [1001]	3 3 3 3 1 1 ...
Tab_LL	double [70]	-223464 -223328 -223230 -223159 -223113 -223076 ...
[[2]]	list [6]	List of length 6
proba	double [2]	0.428 0.572
mu	double [2 x 44]	175.526 26.726 160.015 69.968 100.644 121.418 46.817 155.898 143.000 9.150 ...
covar	double [44 x 44 x 2]	6079.93 3473.58 -547.30 -2417.68 2446.44 4634.40 3473.58 10180.20 6729.54 ...
mat_tk	double [1127 x 2]	3.37e-10 9.76e-01 1.00e+00 1.00e+00 3.41e-10 9.32e-10 1.00e+00 2.42e-02 3.00e-07 ...
cluster	integer [1127]	2 1 1 1 2 2 ...
Tab_LL	double [39]	-252614 -252582 -252571 -252566 -252563 -252560 ...
[[3]]	list [6]	List of length 6
proba	double [3]	0.291 0.388 0.321
mu	double [3 x 44]	139.8 178.0 141.1 144.8 176.7 150.9 145.5 172.8 161.2 146.6 156.1 151.4 111.1 7 ...
covar	double [44 x 44 x 3]	10940.1 8979.8 4978.1 2012.8 1816.2 1556.7 8979.8 10997.9 8115.1 3820.2 ...
mat_tk	double [991 x 3]	8.41e-01 3.85e-01 2.19e-15 2.76e-06 5.50e-14 7.93e-03 1.59e-01 6.15e-01 3.69e-12 ...
cluster	integer [991]	1 2 3 2 3 2 ...
Tab_LL	double [64]	-238821 -238722 -238675 -238636 -238610 -238590 ...

Entre autres : le vecteur de **probabilité** , la matrice **mu**, la matrice de **covariance**, le tableau de **Log Likelihood** etc.

Pour pouvoir tester **MNIST** convenablement, nous l'avons samplé .

Pour réduire le nombre de variables du train , Nous avons choisi que les variables avec une variance $>$ 12350 donc , il ne restait que **45 variables** / 785

Pour les individus , nous avons samplé **10000 individus** sur le dataset de train.

Pour le dataset de Test , nous avons maintenus les mêmes individus en ne prenant que les variables prises pour le dataset de Train.

Résultats de notre modele sur MNIST

```
    0    1    2    3    4    5    6    7    8    9
0 897    0   20   42    0   62   25    9    3    9
1   0 1019    3    0    1    0    2    9    2    4
2  38   36  834   33   14    7   33   34   39   26
3  10    2   30  775    0  129   20    1   95   24
4   3   30   48   18  927   19   16  150    9  265
5  17    2   18   57    0  601   28    1   25    6
6   5    5   16   13    1   12  807    2   13    3
7   3    0    6    0    1   10    1  754    4   13
8   7   38   45   63   20   48   26   30  773   10
9   0    3   12    9   18    4    0   38   11  649

Overall Statistics

               Accuracy : 0.8036
                 95% CI : (0.7957, 0.8113)
    No Information Rate : 0.1135
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7817
```

Résultats de Random Forest sur MNIST

```
> test_pred = predict(fit, test1_reduit)
> mean(test_pred == test1_reduit$y)
[1] 0.8506
> table(predicted = test_pred, actual = test1_reduit$y)
         actual
predicted    0    1    2    3    4    5    6    7    8    9
        0  935    0   20   35    0   34   16    4    4    6
        1    0 1110   14    0    9    4    5   19    8    8
        2   10    6  823   27   20    6   21   37   29   11
        3    4    2   25  771    3   90    9    2   93   16
        4    2    6   24    3  827    6   21   30   15   31
        5   12    4   18   71    1  706   31    2   30    7
        6    2    2   32   25    4   18  837    2   20    3
        7   11    1   12    3   28    5    5  866    4   32
        8    3    4   42   66    8   15   10   20  751   15
        9    1    0   22    9   82    8    3   46   20  880
```

Résultats de SVM sur MNIST

```
> test_pred = predict(svm_model, test1_reduit)
> mean(test_pred == test1_reduit$y)
[1] 0.8645
> table(predicted = test_pred, actual = test1_reduit$y)
         actual
predicted    0    1    2    3    4    5    6    7    8    9
        0  925    1   14   29    0   27   14    4    5    4
        1    1 1108    6    0    5    2    1   16    2    9
        2   10    5  854   28   14    2   13   23   23   11
        3    3    3   25  787    1   77    3    2   88   14
        4    2    5   21    4  855    6   17   42   17   39
        5   23    1   14   72    4  738   31    1   37   21
        6    8    5   29   28    2   19  862    2   17    2
        7    5    3    9    3   33    4    2  892    7   38
        8    1    4   43   53    8   12   11   15  759    6
        9    2    0   17    6   60    5    4   31   19  865
```

Les résultats montre que $EMmodele : 80,36\% < RandomForest : 85,06\% < SVM : 86,45\%$

L'entraînement de notre modèle sur les 10000 individus du dataset de train a pris environ **45 minutes**.

NB : Le nombre de composants max est fixé a 3 et il n'existe que 3 modèles (parcimonieux ou non) disponibles : **VVV**, **EEE**, **VII**