Classement de chiffres manuscrits - Algorithme des K-moyennes et Classification Ascendante Hiérarchique

Timothé Rios

mai 2021

Table des matières

1 Algorithme des K-moyennes

1.1 Apprentissage

Afin de faire un premier clustering, on se propose d'effectuer dix fois l'algorithme des K-moyennes et de retenir le clustering ayant la plus petite erreur quadratique.

1.1.1 Erreur de quantification

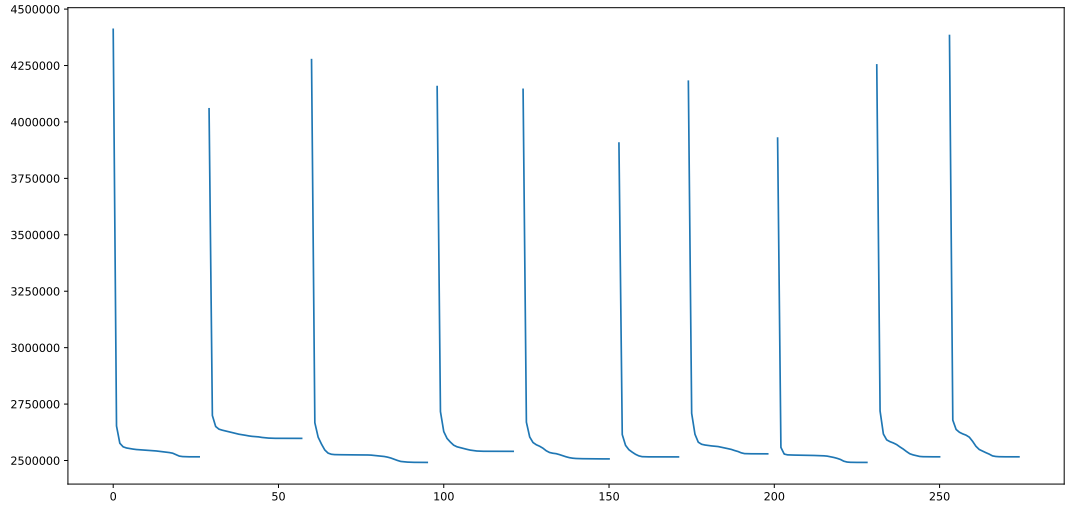

FIGURE 1 – erreur de quantification au fil des itérations des dix initialisations de K-moyennes

On remarque que les erreurs quadratiques initiales et finales de l'algorithme des K-moyennes peuvent varier significativement en fonction de son initialisation.

1.1.2 Histogramme des classes

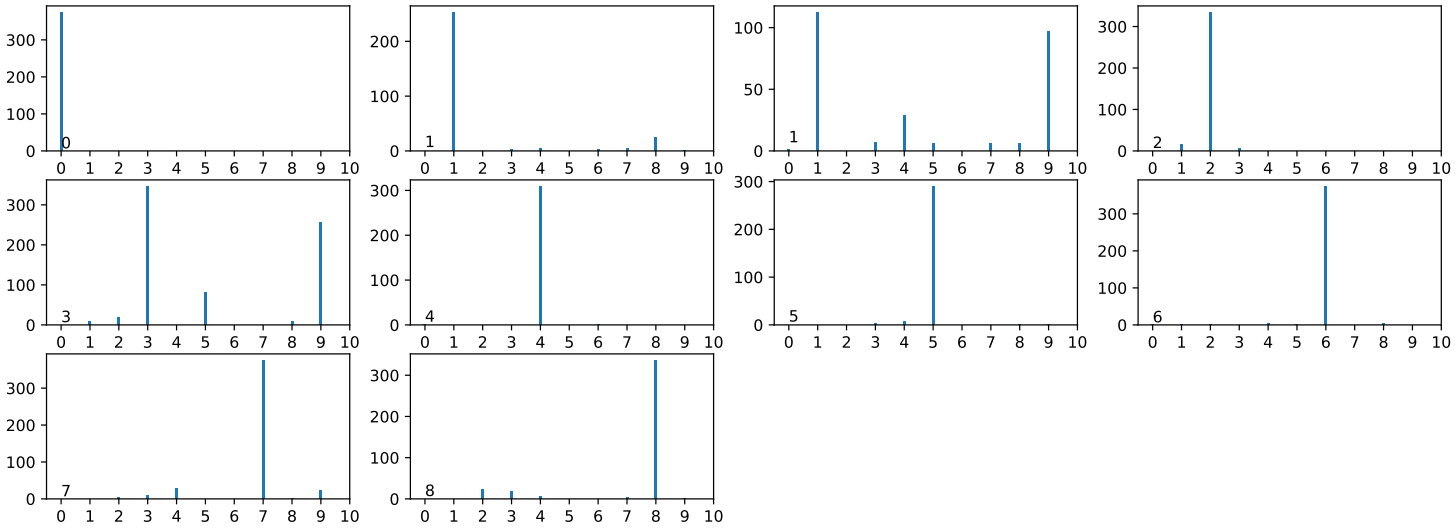

FIGURE 2 – histogramme du nombre de chiffres de chaque classe

On remarque que la plupart des classes ne contiennent qu'un seul chiffre ou presque : 0,2,4,5,6,7 et 8 sont bien différenciés par l'algorithme. Cependant, les trois chiffres restants, 1, 3 et 9, sont confondus par l'algorithme à tel point qu'il n'y a même pas de classe labellisée 9, à cause de la deuxième classe labellisée 1. L'algorithme a donc eu du mal a différencier 1 et 3 et 3 et 9.

1.1.3 Indice de la Silhouette

```
>>> silhouette(10)
0.19434538999602233
```

FIGURE 3 – Indice de la Silhouette pour K = 10

L'indice de la Silhouette de ce clustering est positif, ce qui signifie qu'en moyenne, chaque point est plus proche des points de sa classe que de ceux des autres classes. Le clustering est donc efficace.

1.1.4 Variation du nombre de clusters

Afin de considérer le meilleur clustering possible pour chaque valeur de K, nous effectuons pour chaque K mille fois l'algorithme des K-moyennes et ne gardons que celui ayant la plus petite erreur de quantification.

```
>>> for i in range(10,16):
...     silhouette(i,inits = 1000)
...
...
0.19436174510664778
0.19192231685805639
0.1928351025879894
0.1930903808773749
0.18810334992017724
0.18572189283994542
```

FIGURE 4 – Indices de la Silhouette de K=10 à K=15

La valeur de K ayant le plus grand indice de la Silhouette est 10. Cependant, il est important de noter que les problèmes de confusion remarqués dans la figure 2 sont grandement réduits avec une valeur de K égale ou supérieure à 14.

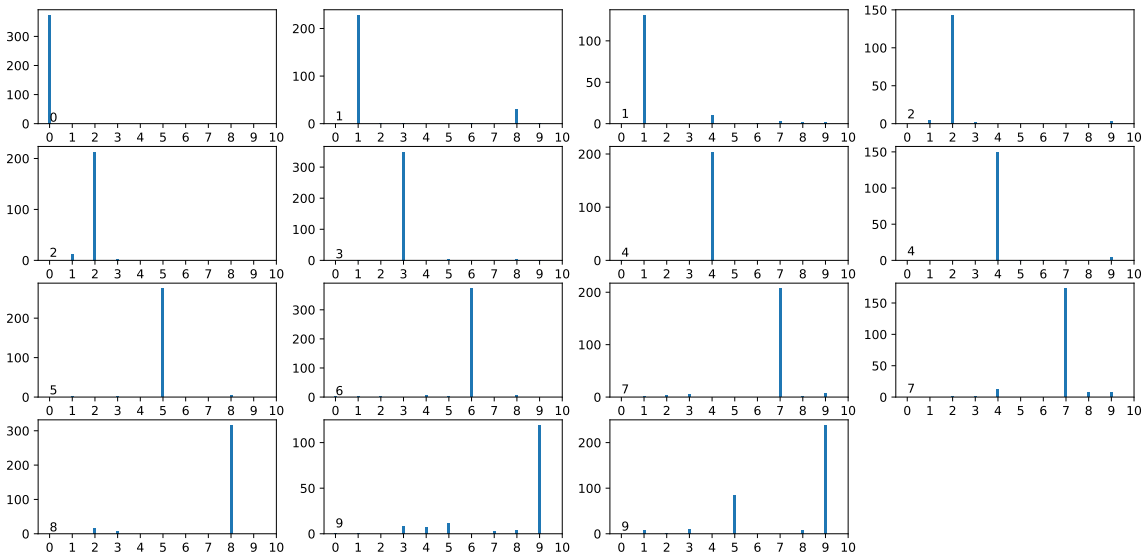

FIGURE 5 – histogramme du nombre de chiffres de chaque classe ($K = 15$)

Par exemple, pour K=15, il ne reste qu'une légère confusion entre 9 et 5 et tous les chiffres labellisent au moins une classe.

1.2 Classification de la base de test

Étant donné que le clustering pour K=10 a donné le meilleur indice de la Silhouette, il a été gardé pour la classification de la base de test.
Le clustering à K=15 a aussi été utilisé pour sa réduction des confusions.

1.2.1 Matrice de confusions

176	0	0	0	2	0	0	0	0	0
0	156	21	1	0	1	2	0	0	0
1	5	150	9	0	0	0	4	8	0
0	1	0	164	0	2	0	10	6	0
0	9	0	0	160	0	0	7	5	0
0	1	0	31	1	148	1	0	0	0
1	4	0	0	0	0	175	0	1	0
0	8	0	0	1	0	0	167	3	0
0	28	1	10	0	1	1	1	132	0
0	24	0	145	0	4	0	4	3	0

FIGURE 6 – Matrice de confusions pour K=10

Comme on s'y attendait, aucune donnée représentant le chiffre neuf n'a été classifiée comme neuf, étant donné que le clustering à K=10 ne produit pas de classe labellisée 9. On retrouve donc logiquement les confusions déjà observées à la dixième ligne de la matrice, où l'on s'aperçoit que les données représentant le chiffre 9 se sont vues attribuer en majorité le label 3, puis le label 1.

177	0	0	0	1	0	0	0	0	0
0	153	22	1	0	1	2	0	0	3
1	5	162	0	0	0	0	3	6	0
0	1	2	154	0	2	0	8	5	11
0	5	0	0	171	0	0	5	0	0
0	0	0	0	2	141	1	0	0	38
1	3	0	0	1	0	175	0	1	0
0	0	0	0	1	0	0	168	0	10
0	19	1	1	0	1	1	1	137	13
0	2	0	3	0	2	0	3	2	168

FIGURE 7 – Matrice de confusions pour K=15

Par contre, dans le cas du clustering à K=15, on remarque que la quasi-totalité des 9 ont bien été classés. Ainsi, malgré un indice de la Silhouette inférieur, l'algorithme des K-moyennes à K=15 présente de meilleurs résultats que celui à K=10. En contrepartie, le nombre de clusters dépasse le nombre de classes réel.

2 Classification Ascendante Hiérarchique

2.1 Apprentissage

Pour notre second clustering, nous faisons un clustering hiérarchique en classification ascendante en utilisant la distance de Ward.

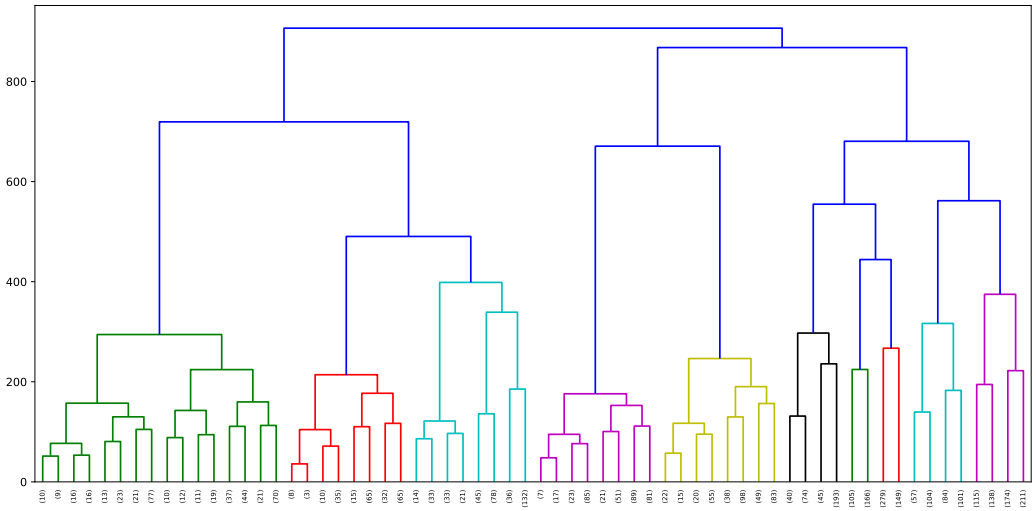

FIGURE 8 – Dendrogramme avec dix clusters visualisés, CAH Ward

2.1.1 Histogramme des classes

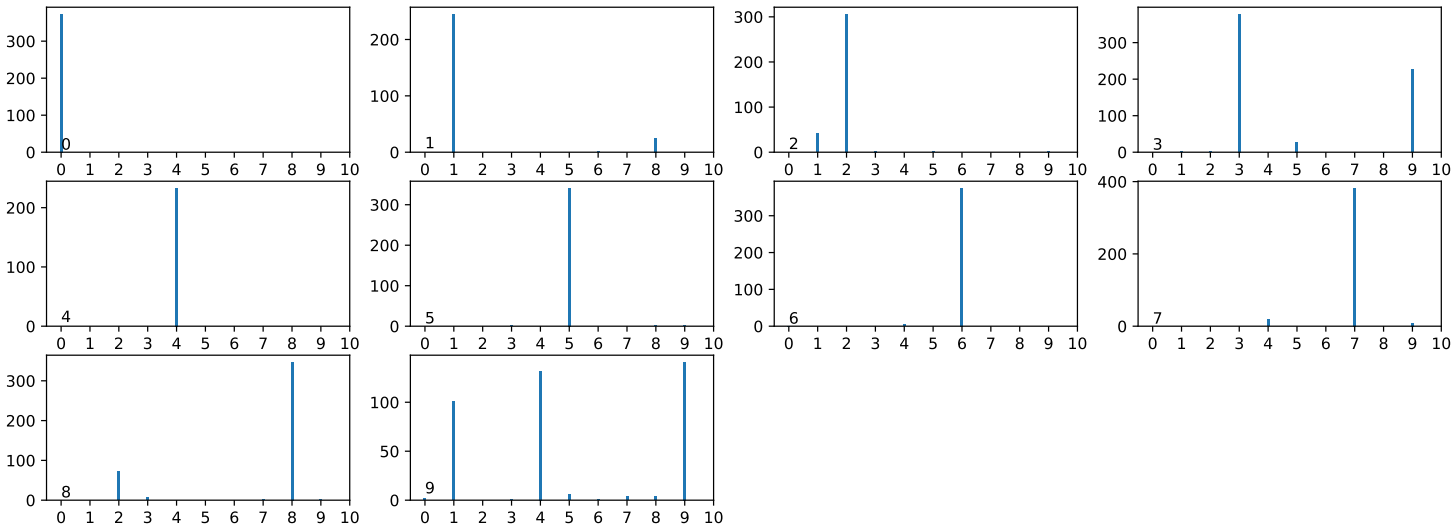

FIGURE 9 – histogramme du nombre de chiffres de chaque classe, CAH Ward

Contrairement à la méthode des K-moyennes pour K=10, chaque classe s'est vue ici attribuée un label différent et tous les chiffre sont donc représentés. Cependant, la CAH montre plus de confusions. On retrouve les confusions entre 1 et 3 et 3 et 9, comme avec l'algorithmes des K-moyennes, mais la CAH confond en plus 1, 4 et 9. Donc, même si tous les chiffres sont représentés dans les labels des classes, les confusions ne sont pas pour autant moins importantes.

2.1.2 Indice de la Silhouette

```
>>> silhouette_h(10)
0.1772647351102698
```

FIGURE 10 – Indice de la Silhouette pour 10 clusters, CAH Ward

L'indice de la Silhouette de ce clustering est positif, tout comme celui de la méthode des K-moyennes. Les deux valeurs sont d'ailleurs très proches. On note cependant que l'indice de la Silhouette de la méthode des K-moyennes est légèrement meilleur que celui de la CAH.

2.1.3 Variation du nombre de clusters

```
>>> for i in range(10,16):
...     silhouette_h(i)
...
...
0.1772647351102698
0.18167773853165362
0.17727560407509277
0.17879647870214888
0.17779971470592992
0.17269069425220068
```

FIGURE 11 – Indice de la Silhouette pour 10 à 15 clusters, CAH Ward

Contrairement à l'algorithme des K-moyennes, l'indice de la Silhouette de la CAH est maximum pour onze clusters. Néanmoins, ce maximum est toujours en dessous de celui obtenu avec la méthode des K-moyennes.

2.2 Classification de la base de test

2.2.1 Matrice de confusion

176	0	0	0	2	0	0	0	0	0
1	105	24	1	0	1	2	0	0	50
1	3	141	7	0	0	0	2	22	1
0	1	0	162	0	2	0	9	9	0
0	4	0	0	127	0	3	3	5	39
0	0	0	13	1	166	1	0	0	1
1	3	0	0	1	1	175	0	0	0
0	0	0	0	0	0	0	156	2	21
0	18	1	4	0	1	1	1	137	11
0	0	0	143	0	2	0	3	4	28

FIGURE 12 – Matrice de confusions pour dix clusters, CAH

Les confusions apparues dans l'histogramme des classes se retrouvent dans la matrice de confusion : comme pour l'algorithme des K-moyennes, peu de chiffres 9 ont été reconnus comme tel. À la différence de ce dernier, le nombre de 9 reconnus n'est pas nul mais les nombres de 1 et de 4 non reconnus sont beaucoup plus importants.

176	0	0	0	2	0	0	0	0	0
0	153	23	1	0	1	2	0	0	2
1	5	142	3	0	0	0	2	21	3
0	1	1	153	0	2	0	6	6	14
0	4	0	0	173	0	1	0	3	0
0	0	0	0	1	163	1	0	0	17
1	3	0	0	1	1	175	0	0	0
0	0	0	0	1	0	0	160	2	16
0	21	1	1	0	1	1	1	136	12
0	2	0	4	0	3	0	0	2	166

FIGURE 13 – Matrice de confusions pour quinze clusters, CAH

Comme pour la méthode des K-moyennes, si l'on augmente le nombre de cluster jusqu'à quinze, les plus importantes confusions disparaissent : on voit ici que la grande majorité des chiffres sont maintenant bien classés, avec notamment un bond pour les 9 qui sont passés de 28 à 168 instances bien classées.

2.3 Comparaison Single Linkage et Linkage de Ward

Pour ce dernier type de clustering, nous faisons un clustering avec la méthode CAH en utilisant la distance minimale (*Single Linkage*) : contrairement à la distance de Ward qui définit la distance entre deux clusters comme la distance au carré de leurs barycentres, pondérée par leurs effectifs, la distance minimale définit la distance entre deux clusters comme la plus petite distance mesurée entre deux données appartenant chacune à un cluster différent.

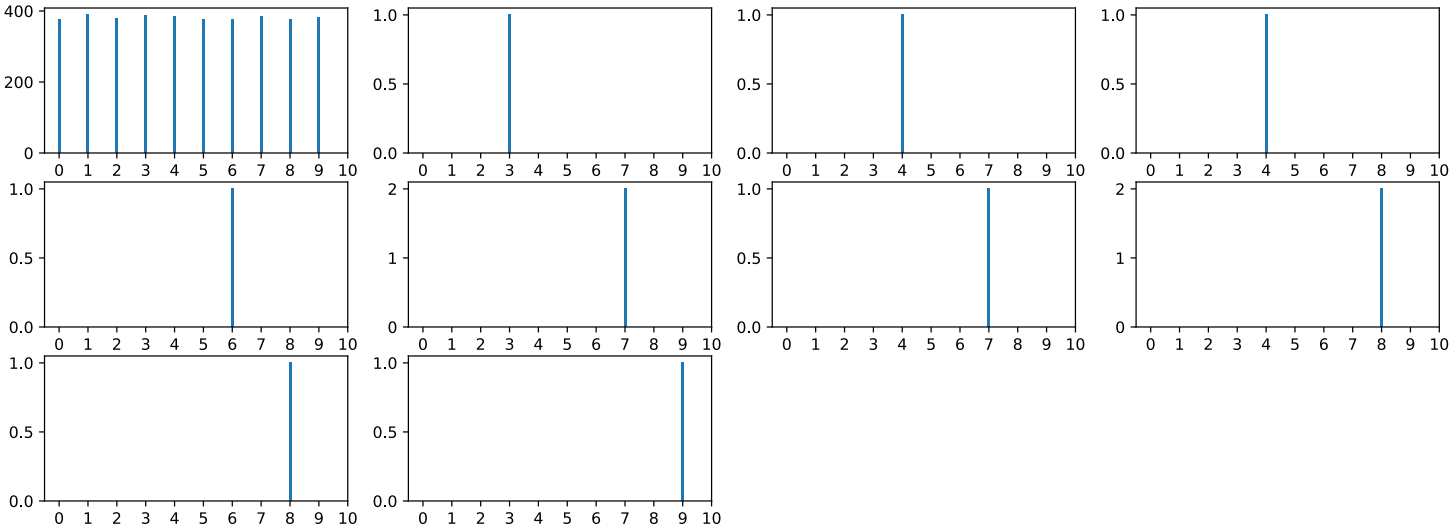

FIGURE 14 – histogramme du nombre de chiffres de chaque classe, CAH Single Linkage

Cette méthode a créé un unique mega groupe contenant la quasi totalité des observations. les neuf groupes restant ne contiennent seulement qu'une ou deux observations. Cette méthode n'est donc pas du tout adaptée à notre base de données.

3 Conclusion : Comparaison des confusions

nombre de clusters / Méthode	**10**	**11**	**12**	**13**	**14**	**15**
K-moyennes	369	233	249	218	202	191
CAH	424	390	261	200	199	193

FIGURE 15 – Tableau des confusions

Ce tableau montre le nombre de chiffres mal classés pour chaque algorithme en fonction du nombre de clusters créés. Il apparaît qu'avec un faible nombre de clusters, l'algorithme des K-moyennes est nettement supérieur à la CAH. Cependant, à mesure que l'on augmente le nombre de clusters, l'écart se ressert jusqu'à être insignifiant par rapport aux variations dans les résultats de l'algorithme des K-moyennes (dues à sa nature stochastique). L'algorithme des K-moyennes est donc meilleur avec un faible nombre de clusters mais la CAH est meilleur avec un grand nombre, notamment parce qu'il est déterministe : il suffit d'une seule itération de l'algorithme pour trouver le clustering optimum.