Analyse d'un jeu de données et évaluation d'algorithmes

Par Tifenn FABRICI-LOURENCO et Kyllian DELAYE-MAILLOT (groupe 2)

SCIENCES SORBONNE UNIVERSITÉ

Présentation du jeu de données

Représente des groupes de discussions

- Environ 18 000 messages
- Plus de 12 000 mots différents
- 20 classes différentes

Nettoyage des données

198 mots inutiles enlevés

Correspondance des labels

Réalisé à partir du jeu de données originel

Nous avons réussi à faire correspondre chaque label avec le nom officiel de son groupe de discussion, en faisant des recherches sur le jeu de données à notre disposition.

Label	Catégorie réelle
0	alt.atheism
1	comp.graphics
2	comp.os.ms-windows.misc
3	comp.sys.ibm.pc.hardware
4	comp.sys.mac.hardware
5	comp.windows.x
6	misc.forsale
7	rec.autos
8	rec.motorcycles
9	rec.sport.baseball
10	rec.sport.hockey
11	sci.crypt
12	sci.electronics
13	sci.med
14	sci.space
15	soc.religion.christian
16	talk.politics.guns
17	talk.politic.mideast
18	talk.politics.misc
19	talk.religion.misc

Questions

- Y a-t-il des similarités entre certains sujets (classes) ?
- Peut-on identifier le sujet d'un message appartient via un classifieur ?

Problématiques rencontrées

- Faible quantité de données
- Absence de métadonnées
- Machines non-adaptées à certains calculs
  - Usage de **500 exemples** (contre 18.000 au total)
- Temps de calcul important
  - Parallélisation des calculs effectués

Précision des Classifieurs Supervisés (en %)

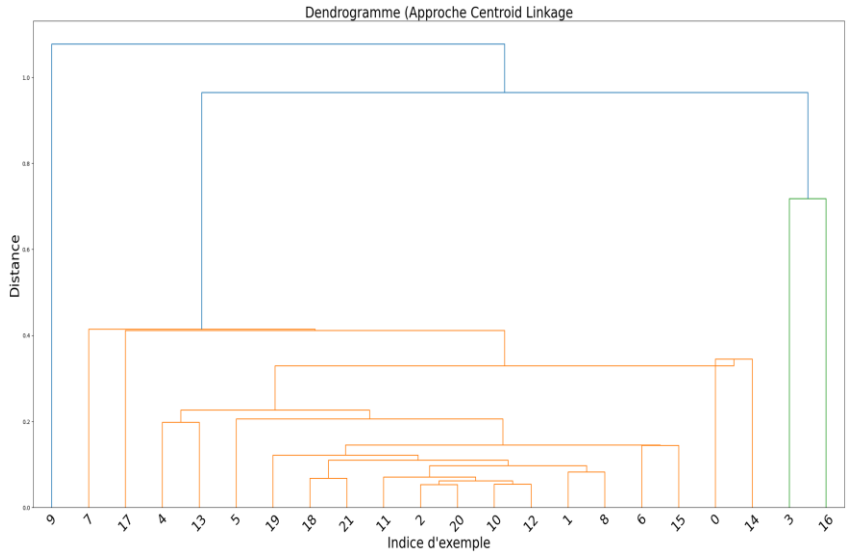
Classifieur	Sur 2 classes	Sur 4 classes
Arbre de décisions	50%	68%
Naïve Bayes	56%	23%
Perceptron	60%	68%

Conclusion : Impossible de faire une classification précise avec un jeu de données restreint. Toutefois, le Perceptron et l'Arbre de décisions semblent être les meilleurs classifieurs.

Dendogramme (Centroid Linkage)

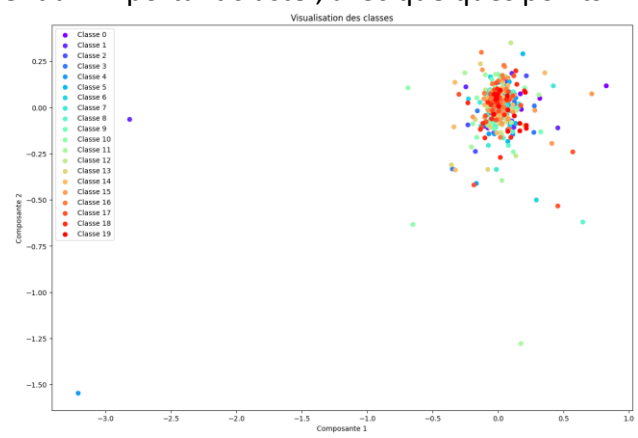
Basé sur les centroids d'un K-moyenne

Avec un seuil de distance de 0.8, un important cluster ainsi qu'un petit se démarquent.



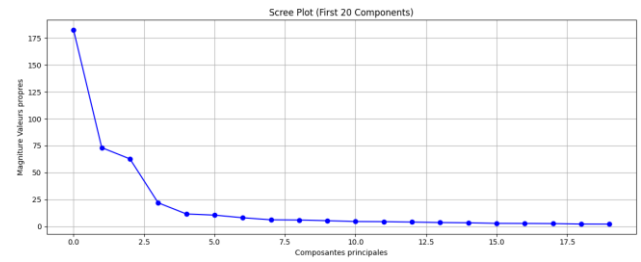
Projection en 2D des données avec la PCA

Les données forment un important cluster, avec quelques points gravitant autour.



Perte d'information avec la PCA

La première composante explique 50% des données tandis que les 19 autres principales ont des gains moindres de 5% à 20%.



Précision des Classifieurs Supervisée – Avec PCA

- **KNN & dist. Euclidienne** : Entre 7% (1NN) et 4% (4NN)
- **KNN & dist. cosinus** : Entre 38% (1NN) et 30% (4NN)
- **Arbre de décisions** : 2%

Précision des Classifieurs Supervisée – Sans PCA

- **KNN & dist. Euclidienne** : Entre 38% (1NN) et 30% (4NN)
- **KNN & dist. cosinus**: Entre 65% (1NN) et 66% (4NN)
- **KNN & dist. de hamming** : Entre 30% (1NN) et 25% (4NN)
- **Arbre de décisions** : 65%
- **Naive Bayes :** 23%

Conclusion : Le **KNN avec dist. cosinus** est le plus efficace.
Remarque : Le manque de données semble avoir un fort impact sur la PCA et conduit à un **sur-apprentissage**, donnant lieu à des performances moindres sur le jeu de test.

Conclusion

Au vu du cluster ci-dessous, nous constatons qu'il y a des classes similaires.

On constate également qu'un nombre de classes réduits fournit une meilleure précision. La similarité des classes et l'absence de données complémentaire empêchent donc une classification précise, ce qui s'atténue quand on restreint les classes à classifier.

L'identification précise devrait toutefois être possible avec un jeu de données plus détaillé et plus rempli.