Classification de locuteurs

Introduction

Ce rapport présente le travail que nous avons réalisé sur une tâche de classification de locuteurs. Les données textuelles proposées correspondent à un ensemble de phrases étiquetées selon l'identité de leur locuteur ; François Mitterrand ou Jacques Chirac.

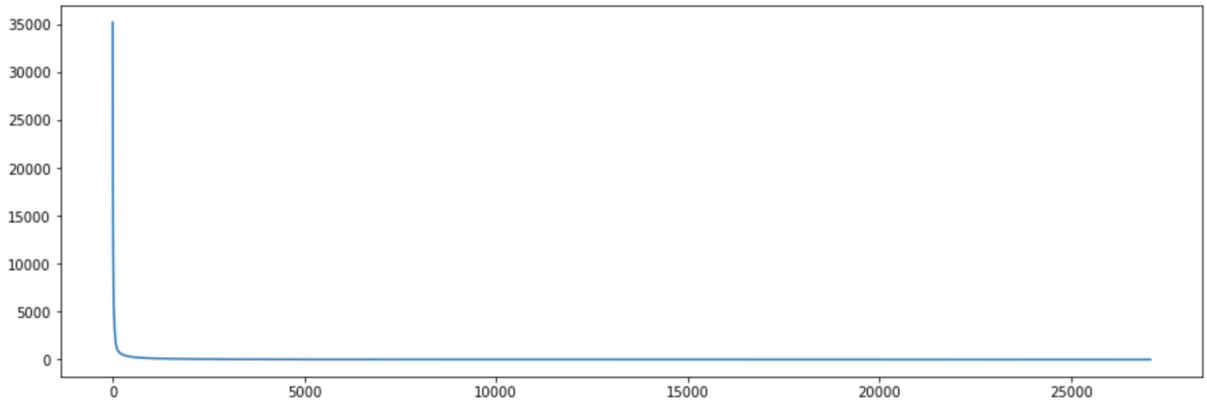
Le jeu de données est composé de 57413 phrases avec ponctuation et accents. Parmi ces phrases, 49890 (**87%**) sont de la classe 1 (classe majoritaire) et 7523 (**13%)** de la classe -1 (classe minoritaire), nous sommes donc dans un cas déséquilibré.

Le vocabulaire, après suppression de la ponctuation, des accents et majuscules, est composé de **27054 mots**.

Si l'on regarde les fréquences documentaires de ces mots, on observe que très peu de mots ont une fréquence documentaire importante :

Fréquence documentaire des mots du vocabulaire

(Triés par fréquence décroissante)



Méthode

Nous avons traité le cas déséquilibré de deux manières différentes :

- Par **data augmentation** en ajoutant des exemples de classe minoritaire (-1) jusqu'à avoir des classes à peu près équilibrées.
- En **pénalisant l'erreur de la classe minoritaire** : les poids des exemples sont ajustés de manière inversement proportionnelle à la taille de la classe à laquelle l'exemple appartient.

Pour ces deux méthodes, nous avons testé trois prétraitements :

1. Prétraitements basiques (ponctuation, accents, majuscules), suppression des stop-words et stemming
2. Prétraitement 1 + uni/bi/trigramme au niveau mot
3. Prétraitement 1 + bi/tri/quadrigramme au niveau caractères

Pour chaque méthode de traitement du cas déséquilibré et chaque prétraitement, nous avons entrainé et optimisé (selon les paramètres de régularisation) trois classifieurs : un SVM linéaire, une régression logistique et un classifieur bayésien naïf. L'optimisation a été réalisée via une validation croisée à cinq blocs.

Nous avons testé deux métriques d'optimisation :

- Score f1 : c'est la métrique classique. Nous avons réorganisé les classes pour que la classe minoritaire corresponde à la classe 1 et la classe majoritaire à la classe -1. De cette façon, l'optimisation du score f1 revient à optimiser la précision et le rappel de la classe minoritaire.
- AUC de la courbe précision-rappel : cette métrique est utilisée dans les cas déséquilibrés d'après nos recherches. Comme le score f1, on réorganise les classes pour que la classe « positive » considérée soit la classe minoritaire.

Il est à noter que l'objectif n'est pas clairement connu ici, nous ne savons pas si nous devons viser un modèle à bonne précision ou bon rappel. Nous avons considéré que l'objectif était de bien prédire les deux classes, c'est-à-dire avoir un taux de bonne classification élevé pour les deux classes.

Dans le cas où les données ont été traitées par data augmentation, seul le score f1 a été testé étant donné que cette méthode rétablit l'équilibre des classes.

Pour chaque métrique, nous avons optimisé les classifieurs selon celle-ci et calculé (sur les données d'entrainement) leurs performances de prédiction globale et pour chaque classe via la matrice de confusion.

Résultats

Sur les données d'entrainement (70%)

Les tables *Table 1* et *Table 2* présentent les résultats de performance sur le jeu de données d'entrainement pour les deux métriques d'optimisation testées.

*Table 1 : Résultats de performance sur les données d'entrainement avec la métrique d'optimisation **score f1**, validation croisée à cinq blocs*

Traitement du cas déséquilibré	Prétraitements effectués	SVM			Bayésien naïf			Régression logistique		
		F1	Maj	Min	F1	Maj	Min	F1	Maj	Min
Pénalisation de l'erreur de la classe minoritaire	Ponctuation/accents/majuscule + stop-words + stemming	0.69	0.89	0.92				0.82	0.89	0.92
	Prétraitements précédents + bigramme/trigramme niveau mots	0.95	0.98	0.99				0.97	0.98	0.99
	Prétraitements précédents + bigramme/trigramme niveau caractères	0.85	0.95	0.99				0.88	0.90	0.94
Data augmentation : oversampling de la classe minoritaire	Ponctuation/accents/majuscule + stop-words + stemming	0.96	0.96	0.94	0.88	0.86	0.88	0.95	0.97	0.95
	Prétraitements précédents + bigramme/trigramme niveau mots	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	Prétraitements précédents + bigramme/trigramme niveau caractères	0.99	0.99	0.99	0.82	0.8	0.82	0.99	0.99	0.99

F1 : score f1 global
Maj/Min : taux de bonne classification de la classe majoritaire/minoritaire

*Table 2 : Résultats de performance sur les données d'entrainement avec la métrique d'optimisation **AUC de la courbe précision-rappel** (notée AP), validation croisée à cinq blocs Traitement du cas déséquilibré : pénalisation de l'erreur*

Prétraitements effectués	SVM			Régression logistique		
	AP	Maj	Min	AP	Maj	Min
Ponctuation/accents/majuscule + stop-words + stemming	0.61	0.89	0.92	0.61	0.89	0.92
Prétraitements précédents + bigramme/trigramme niveau mots	0.91	0.98	0.99	0.90	0.98	0.99
Prétraitements précédents + bigramme/trigramme niveau caractères	0.78	0.95	0.99	0.64	0.90	0.94

AP : AUC de la courbe précision-rappel
Maj/Min : taux de bonne classification de la classe majoritaire/minoritaire

Les résultats précédents montrent de très bonnes performances en entrainement pour la plupart des prétraitements et classifieurs testés. Ces bonnes performances peuvent laisser présager un risque de sur-apprentissage des classifieurs entrainés même si le paramètre de régularisation a été optimisé. Nous avons donc choisi de construire un modèle final qui consiste à faire voter plusieurs classifieurs pour une classe, et le classifieur final prédit la classe qui apparait le plus dans les votes.

Concernant la sélection des classifieurs à considérer pour le vote, nous avons conservé les classifieurs entrainés après le prétraitement par unigramme/bigramme/trigramme au niveau mots, avec les deux types de de traitement du cas déséquilibré et optimisés avec la métrique f1. Au total, cinq modèles répondent à ces critères.

Sur les données test (30%)

Le modèle final a été appliqué au jeu de données test, les performances obtenues sont présentées dans la table *Table 3*.

Table 3 : Résultats de performance sur les données test du modèle final par vote, sous forme de matrice de confusion

		Classe prédite	
		-1	**1**
Classe réelle	**-1**	14585 (**0.91**)	1369 (0.09)
	1	359 (0.28)	911 (**0.72**)

Le modèle final présente donc une performance de 91% sur la classe majoritaire et 72% sur la classe minoritaire.

Conclusion

Nous avons réussi à entrainer plusieurs classifieurs sur des données issues de différents prétraitements et à obtenir de très bonnes performances en entrainement.

Le traitement du cas déséquilibré a été approché par deux méthodes dont les résultats sont satisfaisants en entrainement, et par l'utilisation d'une métrique d'optimisation spécifique qui n'a pas semblé donner de meilleures performances que le score f1 classique.

La combinaison de plusieurs des modèles par un système de vote a permis de construire le modèle final qui présente de bonnes performances en test sur la classe majoritaire, mais pas tout à fait satisfaites pour la classe minoritaire.

Parmi les difficultés rencontrées, nous avons notamment eu des difficultés à trouver une méthode rigoureuse pour passer de l'étape d'entrainement au test du modèle final sur les données test. En effet, étant donné que tous les classifieurs, quels que soient les prétraitements associés, présentaient de très bonnes performances, il était difficile de choisir un modèle à conserver pour la généralisation. Nous ne pouvions pas tous les tester sur les données test et garder le meilleur sinon cela reviendrait à continuer d'apprendre sur l'ensemble de test.