Machine Learning for Natural Lenguage Processing Sujet 1

Pablo BARRIO

April 30, 2024

1 Introduction

L'objectif de ce projet est de prédire le sexe d'individus en France à partir des différents données disponibles telles que leur prénom, nom, profession, état civil entre autres. Cependant, la base de données correspond à une transcription par reconnaissance automatique de l'écriture de listes nominatives manuscrites des recensements de 1836 à 1936 et nous n'avons accès qu'à un petit échantillon de ces transcriptions contenant 241 individus. La base de données est comportée d'une partie appelée "groundtruth" qui correspond à ce qui est vraiment écrit sur les listes de recensements, une partie "prediction" qui correspond à la transcription par reconnaissance automatique, ainsi qu'à un identifiant poutr chaque individu et leur sexe. L'objectif de projet est de prédire le sexe de chaque individu en n'utilisant comme données que la partie prédiction. Ceci, nous verrons plus tard comment, va difficulter la tâche puisque les transcriptions réalisées contiennent des erreurs. Ensuite, nous disposons aussi d'une base de données contenant des prénoms d'individus en France, ainsi que la fréquence du sexe des personnes portant ces prénoms. D'aurenavant , nous nous referrons à la preière base de données comme "transcription", et la deuxième comme "fréquence prénoms".

Commençons d'abord par faire des statistiques descriptives des parties "prediction" et "sex" de la base transcriptions.

1.1 Statistiques Descriptives

La base de donées transcriptions contient 241 lignes, une pour chaque individu. Ces individus sont formés par 107 femmes, 125 hommes et 9 personnes pour lesquelles on ne connait pas le sexe et donc labélisées comme "ambigu". On remplace "ambigu" par "femme" ou "homme" en regardant la fréquence d'hommes et femmes portant le nom de chaque individu labelisé de sexe "ambigu" sur la base "fréquence prénoms". On leur attribue le sexe dont la fréquence est la plus importante pour ce prénom.

Dans "prediction" nous trouvons les données personnelles de chaque individu: employeur, profession, état civil, relation, lieux de naissance, date de naissance, prénom

et nom. Nous commençons par étudier le nombre de valeurs manquantes pour chaque caractéristique en créant un dataframe où chaque caractéristique est sur une colonne. Voici les valeurs manquantes:

Catégorie	Nombre
Employeur	204
Profession	177
Relation	83
État Civil	203
Lieux de Naissance	60
Date de Naissance	11
Prénom	13
Nom	26
Sexe	0

Table 1: Nombre de valeurs manquantes par caractéristique

L'objectif étant de prédire le sexe de l'individu à partir de ses données personnelles, il semblerait que pas toutes les informations sont fiables. C'est le cas notamment des informations employeur et état_civil, pour lesquelles nous avons beaucoup de valeurs manquantes. De plus, lorsque nous regardons plus en détail les valeurs de état_civil et employeur qui ne sont pas manquantes, la plupart n'ont aucun sens. Ceci explique pourquoi nous avons presque le même nombre de valeurs non manquantes que de valeurs uniques pour ces caractéristiques. Les variables date_naissance et lieu_naissance présentent aussi beaucoup de valeurs sans sens, ce qui explique le nombre important de valeurs uniques pour ces variables. Ceci est intéressant et nous allons étudier quel est l'impact sur la prédiction en fonction du modèle.

Nous faisons un nuage de mots pour les femmes puis un autre pour les hommes, afin d'observer quels mots sont les plus importants pour différencier les deux sexes (1).

Figure 1: Word Cloud des données personnelles des femmes

L'autre word cloud pour les hommes est disponible dans le code. Nous pouvons observer des différences entre les mots qui apparaissent entre les deux nuages de mots. Notamment, les mots qui semblent être le plus fréquents dans les deux catégories sont des prénoms tel que Marie ou Jean puis aussi des relations tel que mère ou fils, et des professions tel que patron. Ces variables semblent être donc les plus utiles à prédire le sexe de l'individu. Ce sont aussi les variables qui contiennet le plus de mots ayant un sens.

2 Modèles

Afin de prédire le sexe des individus, nous allons tester différents modèles puis comparer leurs performances en utilisant différents types de données.

On commence par une vectorisation des textes, tout d'abord en comptant le nombre de fois qu'apparaît un mot dans le texte, puis en utilisant le TF-IDF. On utilise ces deux derniers pour entraîner un modèle bayésien naïf et donc prédire le sexe de l'individu. On ne s'attend pas à obtenir des résultats très différents entre les deux représentations (comptage et tf-idf), puisque le nombre de mots par individus est très limité et parce que, en général, un mot n'apparaît pas plus d'une fois. Par exemple, le prénom Marie, devrait que être présent une fois dans les données personnelles d'un individu qui s'appelle Marie, à moins que son nom soit aussi Marie (très rare).

Ensuite, puisque le nombre de données disponibles n'est pas très élevé (241), et que le nombre de mots non plus, la prédiction du sex de l'individu pourrait bénéficier de l'utilisation d'un modèle pré-entrainé tel que CamemBERT. L'utilisation du tokenizer de CamemBERT qui va découper les mots en sous-entités pourraît être utile à cause du faible nombre de mots dans le corpus, mais ceci ne reste qu'un hypothèse.

3 Expérimentation

3.1 Bag of Words: Count Vectorizer et TF-IDF

On utilise le vectorisation en comptage et en tf-idf dans deux cas différents. Dans le premier on n'utilise que les prénoms comme texte, dans le second on utilise l'ensemble des données disponibles. On fait la prédiction dans les deux cas avec un modèle bayésien naïf.

Lorsque l'on utilise que les prénoms, on trouve des résultats très proches entre le modèle qui utilise les vecteurs de comptage et ceux de TF-IDF. De plus, il semblerait que le modèle overfit dans les deux cas avec une accuracy bien supérieure sur le train set que sur les set de validation et test. Ceci semble logique, puisque les vecteurs dans les deux cas vont ressembler à des vecteurs contenant des variables binaires, résultant du fait qu'un prénom n'apparaît qu'une seule fois dans une ligne d'information d'un individu. Si les individus qui s'appellent Marie dans le train set sont majoritairement des femmes, alors un homme qui s'appelle Marie dans les test et validation sets sera

prédit comme femme. Les prénoms par eux mêmes, sans tokenization, ne sont donc pas une information suffisante pour prédire parfaitement le sexe d'un individu.

On a donc testé les mêmes vectorisations et modèle en utilisant l'ensemble du texte disponible avec toutes les caractéristiques de l'individu. On trouve des meilleurs résultats dans ce cas, comme attendu avec des meilleurs accuracy sur les test et validation sets. En effet, on pourrait faire une analogie avec une régression linéare. En présence d'une seule covariable, le modèle aurair du mal à prédire les cas particuliers, comme les hommes qui s'appellent Marie, alors qu'en rajoutant d'autres variables, cette tâche devient moins difficile.

A contination, on fait le même exercice en utilisant un CamemBERT pré-entraîné. En effet, l'utilisation du tokenizer de CamemBERT, pourrait nous donner des meilleurs résultats, puisque la décomposition en des entités plus petites pourrait porter plus d'informations qu'un mot par lui même. L'utilisation de toutes les données pourrait même être avantageuse, si le modèle arrive à "trouver" du sens pour des mots que nous avons pensé ne pas aider à la prédiction du sexe de l'individu. A continuation on essaye de trouver une réponse à cette question.

3.2 CamemBERT

Dans cette partie, on fait le même exercice que dans la partie précédente mais on rajoute une nouvelle catégorie de données sur laquelle on peut évaluer les performances du CamemBERT. On prend les prénoms, la relation et la profession. En effet, ces informations sont meilleurs que les autres parce qu'elles contiennent des mots qui ont du sens, contrairement à date_naissance par exemple, qui semble contenir des chiffres n'ayant pas de sens.

On entraîne donc un CamemBERT pré-entraîné pour trois textes différents: les prénoms (1), les prénoms, la relation et la profession (2), et finalement toutes les informations disponibles (3).

On fixe pour tous les modèles un batch size de 16, 3 epochs et un learning rate de 2e-5, et on utilise le CamemBERT pour la prédiction, aprés la tokenisation.

En n'utilisant que les prénoms comme source d'information, on obtient une accuracy de 0.9167 sur le set de validation et de 0.9565 sur le test set. Nous obtenons déjà des résultats bien meilleurs qu'avec les bag of words et le modèle naïf bayésien, pour lesquels les meilleurs résultats étaient obtenus sur l'ensemble des données avec 0.875 d'accuracy pour le set de validation et 0.95 pour le testset. Cependant on veut voir si on peut améliorer ces résultats en utilisant plus de textes.

Lorsque l'on prend en compte les information correspond au prénom, à la relation et à la profession, on obtient des accuracy de 0.9583 pour le set de validation et de 0.9565 pour le test set. Les résultats se sont un peu améliorés par rapport au cas précédent.

Finalement, losque l'on utilise l'ensemble des textes pour ré-entraîner le CamemBERT, on obtient des accuracy de 0.9167 pour le set de validation et de 0.9565 pour le test set. On note que l'accuracy pour le set de validation a diminué faiblement par rapport au cas précédent.

On peut donc conclure que l'utilisation des prénoms comme unique source d'information pour prédire le sexe d'un individu a ses limites, et qu'en rajoutant plus d'information, comme la profession et la relation on améliore la prédiction. Ceci semble logique pour les mêmes raisons évoquées en partie 1 (analogie à la regression linéaire). Cependant, il reste une question à répondre. L'utilisation de plus de mots, même si ceux-ci, a priori, ne semblent pas avoir beaucoup de sens, permet-elle de mieux prédire le sexe lorsque l'on utilise un modèle pré-entrainé comme CamemBERT? Nos résultats nous mènent à penser que non, puisque les résulats sont meilleurs (un petit peu) lorsqu'on n'utilise que le prénoms, la relation et la profession plutôt que lorsqu'on utilise tout le texte. Un nettoyage des données semble donc important pour maximiser la performance de CamemBERT. Néanmoins, la différence d'accuracy entre ces deux derniers cas est très faible. Le CamemBERT est donc capable de sélectionner les parties du texte qui semblent avoir un sens pour la prédiction du sexe de l'individu.

Voici une matrice de confusion pour le set de validation:

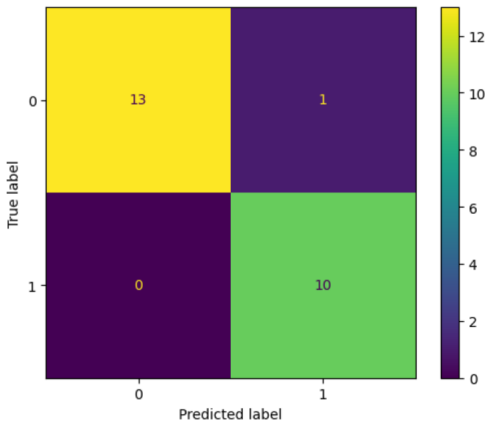

Figure 2: Matrice de confusion prédiction sir validation set avec informations: prénom, relation, profession