Apprentissage automatique

Projet 1

Manon FLEURY-BENOIT

Erwan MARTIN

M1 Sciences Cognitives

2023 / 2024

Projet 1 : Prédiction de la langue d'un texte

On s'intéresse au comportement d'un perceptron multicouche permettant de prédire la langue d'un texte, à partir de statistiques sur la fréquence des lettres du texte.
On entraîne notre modèle avec des données d'apprentissage (du texte brut dans chacune des langues) puis on évalue ses performances sur des données de test, sur lesquelles il doit faire des prédictions. Mais les performances obtenues par le classifieur dépendent du nombre d'exemples fournis lors de l'apprentissage. Ainsi, on peut s'intéresser en premier lieu à la taille des données d'apprentissage et analyser l'évolution des prédictions. En second lieu, on peut se focaliser sur la couche d'entrée du modèle, qui contient les statistiques extraites du texte, soit la fréquence de lettres. On peut donner en entrée les fréquences des bigrammes (séquence de deux lettres), mais également des unigrammes (une lettre) ou des trigrammes (trois lettres), et analyser la performance.

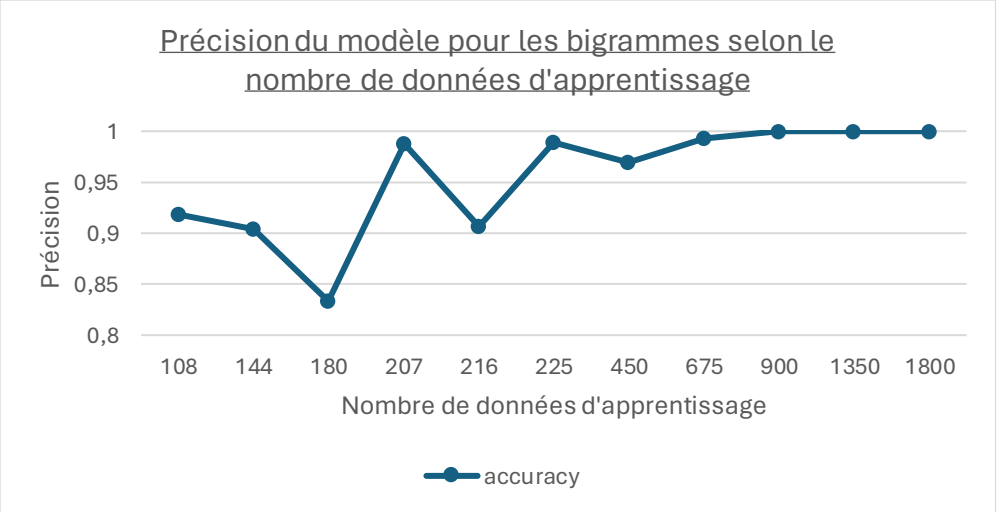
I. Taille des données d'apprentissage

Nous allons faire varier le nombre d'exemples d'apprentissage et regarder comment cela influe sur les performances de prédiction de notre modèle. On travaille sur des bigrammes.

Hypothèses :

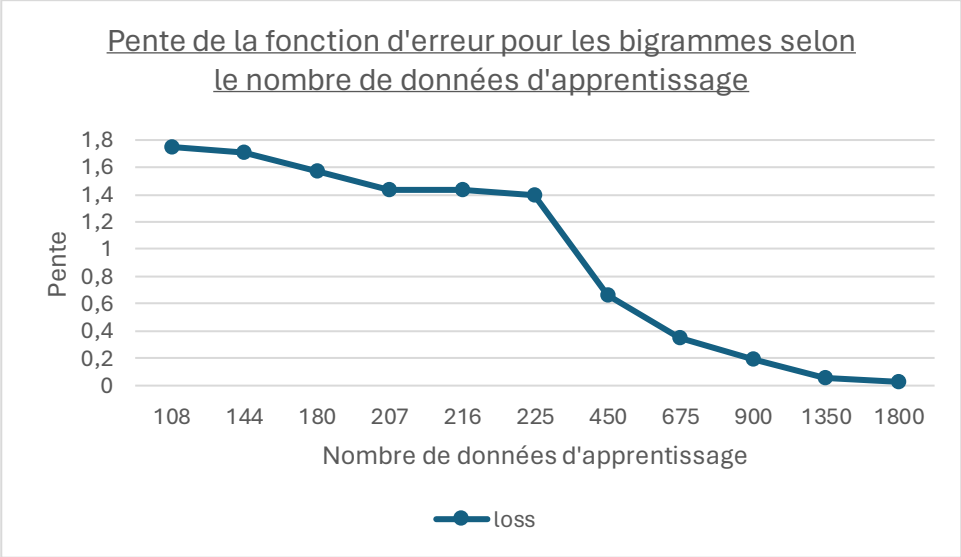
- On peut s'attendre à obtenir une amélioration du modèle avec une précision de plus en plus proche de 1 et une pente de la fonction d'erreur de plus en plus proche de 0.
- L'augmentation de la taille des données d'apprentissage améliorerait les performances du modèle jusqu'à un certain seuil où les prédictions seront trop collées aux données d'apprentissage et la solution n'est pas générale (surapprentissage).
- Les langues ayant des caractéristiques linguistiques similaires auront besoin de plus de données d'apprentissage pour être complètement identifiées, que les langues avec moins de lien avec les autres.

On trace des courbes d'apprentissage donnant les performances du réseau en fonction de la taille des données d'apprentissage, pour N = {12, 16, 20, 23, 24, 25, 50, 75, 100, 150, 200}, le nombre de bigrammes et nombre de tirages, et on obtient comme nombres d'exemples lus les valeurs ci-dessous en ordonnée (nombre de données d'apprentissage). On a par exemple 900 exemples lus pour N = 100. On trace la précision du modèle en fonction de nombre de données d'apprentissage :



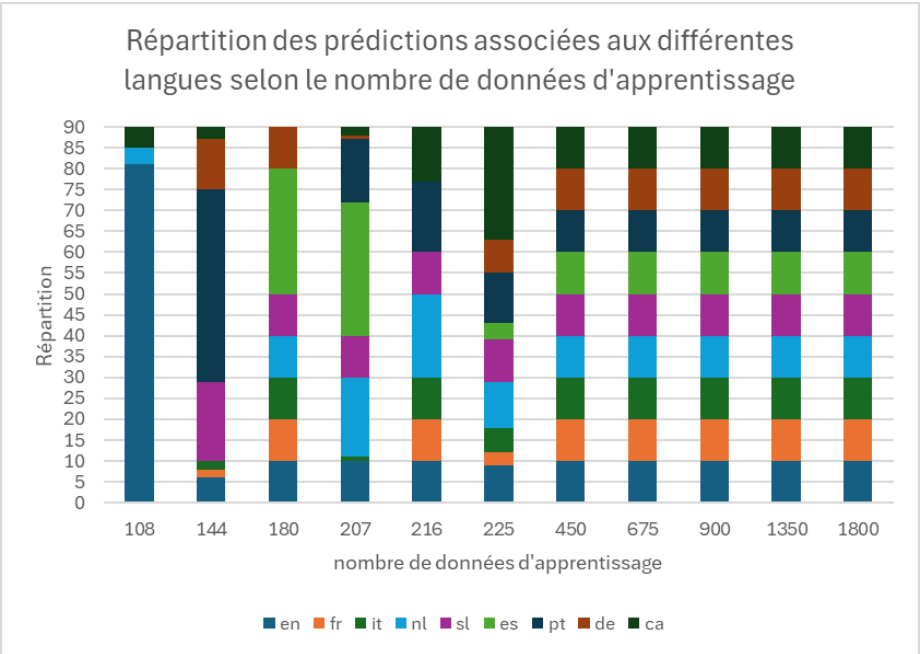
La courbe nous indique que lorsque le nombre de données d'apprentissage est faible, la précision du modèle n'est pas fiable. Elle varie entre 0.85 et 0.95 pour un nombre de données d'apprentissage variant entre 12 et 75 pour le nombre de bigrammes et de tirages. On remarque qu'on atteint une précision de 1 pour N = 100 (900 essais) : la fonction prédit parfaitement la langue du texte, il n'y a aucune erreur entre la prédiction de la langue faite par le modèle et la langue du texte.

On trace également l'évolution de la pente de la fonction d'erreur du modèle selon le nombre de données d'apprentissage :



On observe une baisse successive entre les nombres d'essais variant de 12 à 100. La fonction d'erreur diminue donc bien selon le nombre de données d'apprentissage. On a d'ailleurs une faible valeur de la fonction d'erreur (0,0601) pour 150 données d'apprentissage, ce qui nous montre que la fonction d'erreur tend vers 0 plus le nombre de données d'apprentissage est important.

Toutes les langues ont-elles besoin d'autant de données ?



Pour que le modèle soit fiable, il faut qu'il identifie de manière distincte 10 fois chaque langues (c'est pour cela que l'on a 90 données en ordonnées), le nombre de données d'apprentissage est sur l'axe des abscisses N = {108, 144, 180, 207, 216, 225, 450, 675, 1350, 1800}. Ce diagramme met en avant le nombre de fois où une langue a été prédite par rapport au nombre de données d'apprentissage.

Pour 108 données d'apprentissage, seulement 3 langues sont identifiées parmi les 9 langues et sur 90 exemples. Sur ces 3 langues, l'anglais est identifié 81 fois sur les 90, tandis que le catalan et le néerlandais sont identifiés 5 et 4 fois sur les 90 exemples.

Pour 144 données d'apprentissage, on remarque qu'un nombre plus important de langues sont identifiées, mais toujours pas de manière équivalente. Le portugais est identifié sur 50 % des exemples, on peut supposer que cette langue a été identifiée à la place de l'espagnole, du catalan et de l'italien ou encore du français, qui sont des langues identifiées faiblement.

Pour 180 données d'apprentissage, l'identification des langues commencent à être optimale pour certaines (comme l'anglais, le français, l'italien, l'allemand, le norvégien, le slovène) alors que pour d'autres elle est toujours difficile, comme pour le catalan et le portugais, dont l'espagnol a dû être identifié à la place.

Pour 207, 216 et 225 données d'apprentissage, le modèle évolue encore, l'identification varie toujours autant selon les langues, mais de manière moindre. On s'aperçoit que selon l'identification de certaines langues, d'autres le sont moins. C'est le cas pour l'espagnol et le portugais lorsque le catalan est majoritairement identifié, également entre l'allemand, le norvégien et le slovène.

Pour 450, 675, 900,1350 et 1800, la précision du modèle étant de 1, l'identification des différentes langues est fiable et sans erreur.

Pour conclure, l'augmentation de la taille des données d'apprentissage améliore les performances du modèle, mais l'on ne rencontre pas de seuil où les prédictions sont trop collées aux données d'apprentissage et que l'on soit victime de surapprentissage. Peut-être qu'il faudrait un nombre de données d'apprentissage plus important afin d'atteindre ce stage de surapprentissage. On obtient une amélioration du modèle avec une précision de plus en plus proche de 1, jusqu'à l'atteindre et une pente de plus en plus proche de 0. Les langues ont besoin d'un nombre important de données d'apprentissage afin d'être identifiées précisément. On s'aperçoit, mais on ne peut pas confirmer, que les langues présentant des similarités linguistiques nécessitent davantage de données d'apprentissage pour être correctement identifiées.

II. Trigrammes et Unigrammes

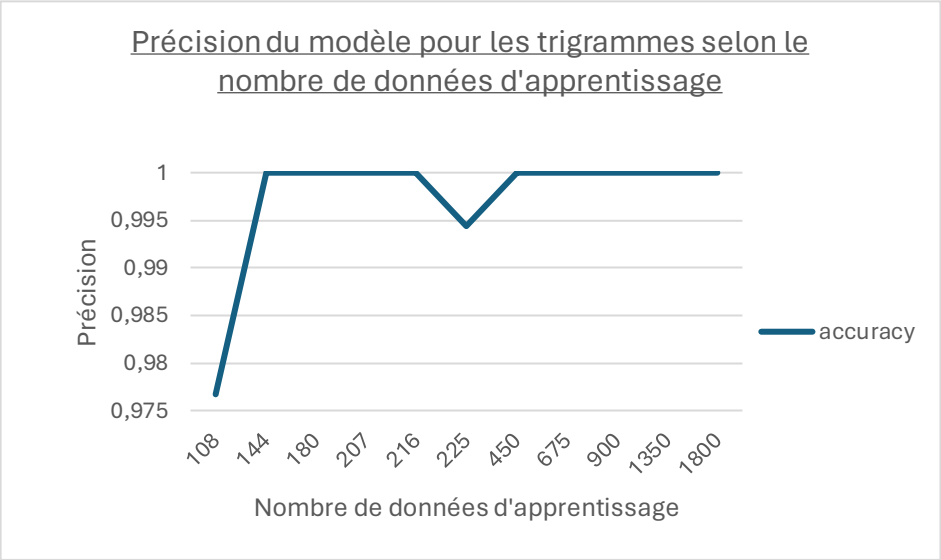
On se demande si modèle unigramme suffit par rapport au modèle bigramme, et si un modèle trigramme obtiendra de meilleures performances. Nous allons donc comparer ces trois modèles entre eux.

Hypothèses :

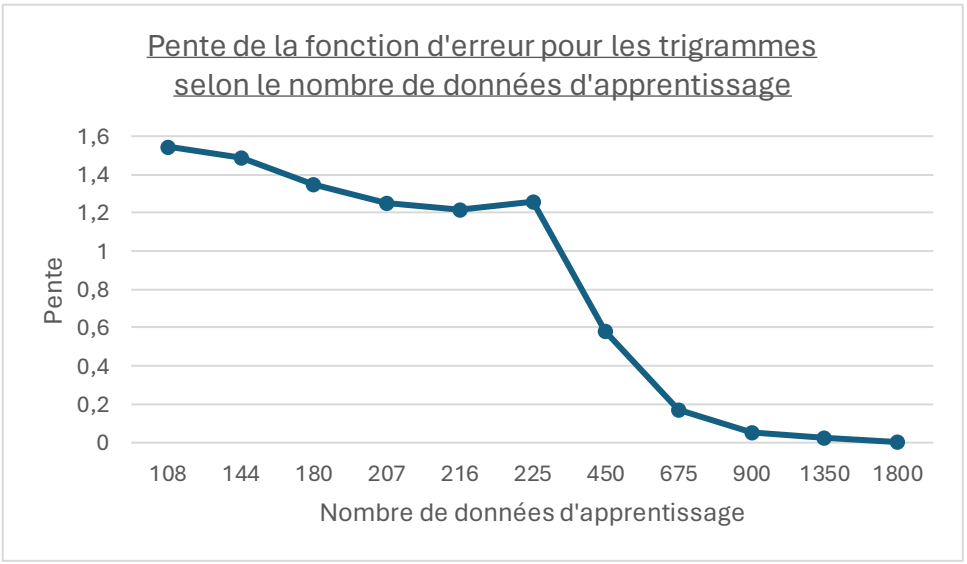
- Le modèle unigramme suffit par rapport au modèle bigramme.
- Le modèle trigramme obtiendra de meilleures performances que les deux autres modèles.
- Le nombre de données d'apprentissage est moindre pour le modèle trigramme que pour le modèle bigramme ou unigramme afin d'atteindre une précision de 1.

Afin de répondre à nos hypothèses, nous traçons des courbes d'apprentissage de la précision et de la pente, avec les mêmes valeurs de N (taille des données d'apprentissage) données précédemment.

1. Trigrammes

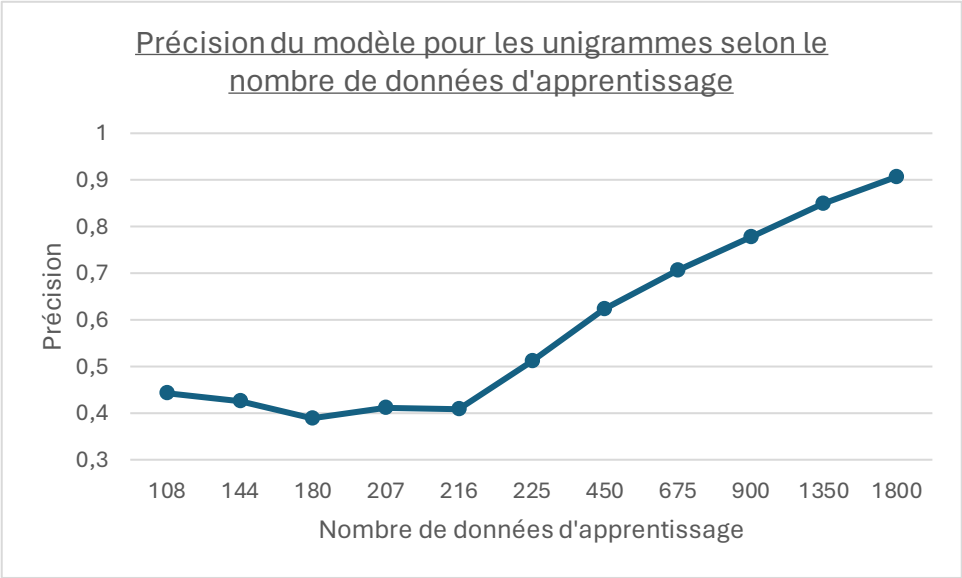


La courbe de précision nous indique qu'il faut très peu de données d'apprentissage avant que le modèle n'arrive à identifier de manière optimale les différentes langues. En effet, dès 144 données d'apprentissage, on atteint une précision de 1.

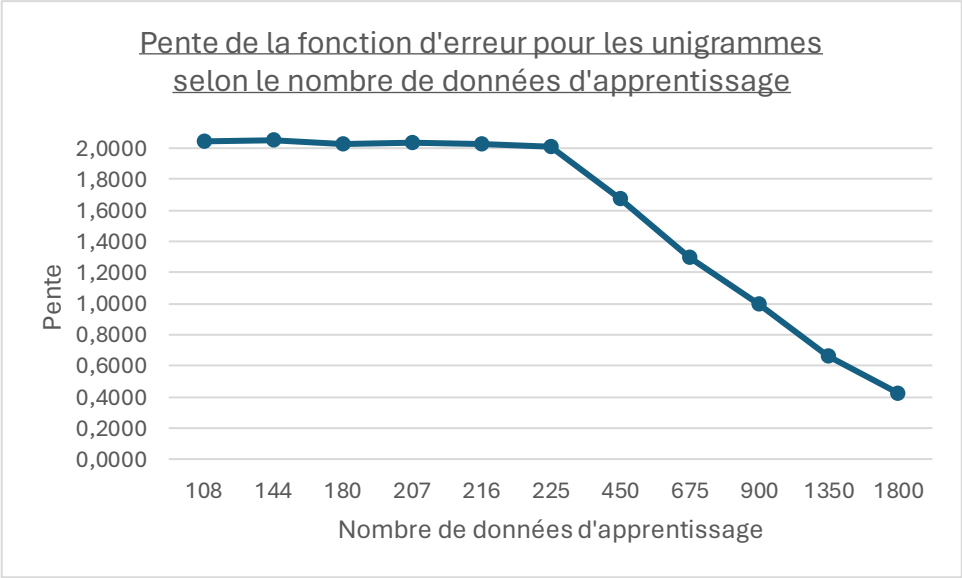


La courbe de la pente du modèle des trigrammes montre que plus le nombre de données d'apprentissage est important, plus la pente de la fonction d'erreur est faible et tend vers 0. Elle est quasiment à 0 à N = 900.

2. Unigrammes



La courbe de précision indique que la précision du modèle est très faible avec un peu de données d'apprentissage. A partir de 225 données d'apprentissage, la précision est de 0.5 soit 50 %. Et il faut plus de 1800 données d'apprentissage pour espérer avoir une identification correcte de chaque langue.



La courbe indique que la pente pour les unigrammes est très élevée (loss = 2.0) jusqu'à 225 données d'apprentissage où la fonction d'erreur diminue en lien avec une amélioration de la précision du modèle. Mais même pour N = 1800, la valeur de la pente n'atteint pas 0.

Nos résultats montrent que le modèle trigramme atteint une précision de 1 avec seulement 144 données d'apprentissage, contre 900 données pour le modèle bigramme, et contre plus de 1800 pour le modèle unigramme pour prétendre à une performance similaire. De plus, la

fonction d'erreur du modèle trigramme tend vers 0 plus rapidement avec l'augmentation des données d'apprentissage par rapport aux autres modèles.

On peut alors conclure que le modèle unigramme ne suffit pas par rapport au modèle bigramme, puisqu'on voit qu'il faudrait un nombre bien plus important de données d'apprentissage afin que les différentes langues soient identifiées sans erreur. Le modèle trigramme obtient de meilleures performances par rapport aux deux autres modèles, il a besoin de bien moins de données avant d'atteindre une précision de 1 contrairement aux deux autres modèles.

En conclusion, ce projet a exploré l'utilisation d'un perceptron multicouche pour la prédiction de la langue d'un texte, en variant la taille des données d'apprentissage et en comparant les performances des modèles utilisant des unigrammes, des bigrammes et des trigrammes. Notre analyse révèle que l'augmentation de la taille des données d'apprentissage améliore les performances du modèle sans atteindre un point de surapprentissage. Les résultats suggèrent également que la représentation trigramme offre de meilleures performances prédictives par rapport aux unigrammes et bigrammes, nécessitant moins de données d'apprentissage pour atteindre une parfaite précision.