Découverte de l'Apprentissage Profond

Application au Traitement du Langage Naturel

Présenté par

Radu CRACIUN

Encadré par

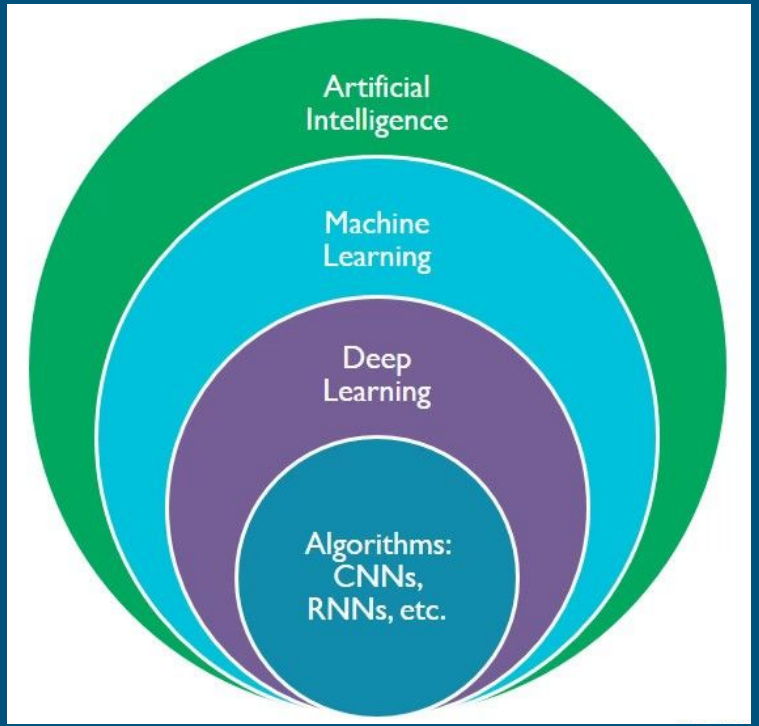
Devan SOHIER

Sommaire

Sommaire

Introduction

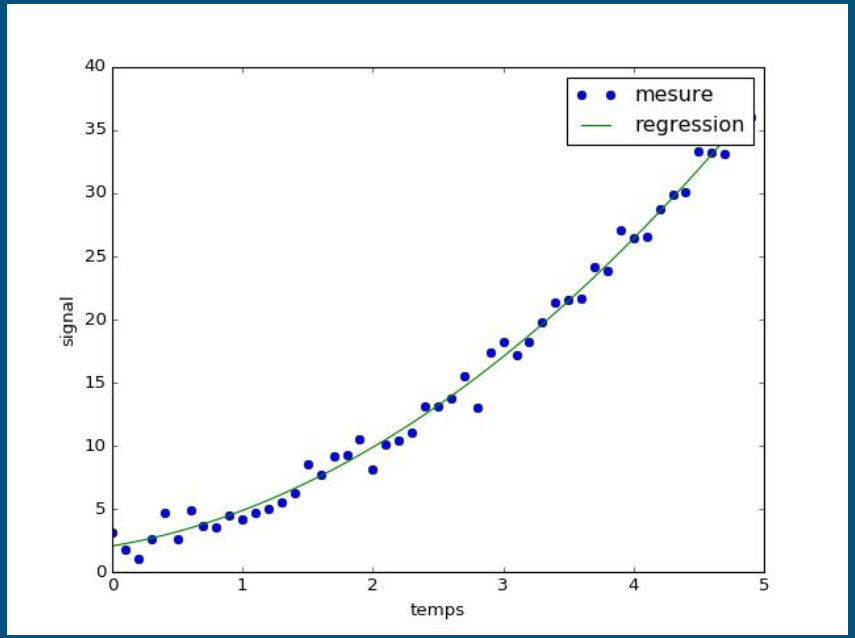
- Cadre de travail :
  - Projet de fin d'études
  - Apprentissage profond ? Intelligence artificielle ? *Machine learning* ?
  - Données textuelles
- Problématique : Peut-on reconnaître des auteurs à leur style d'écriture ?



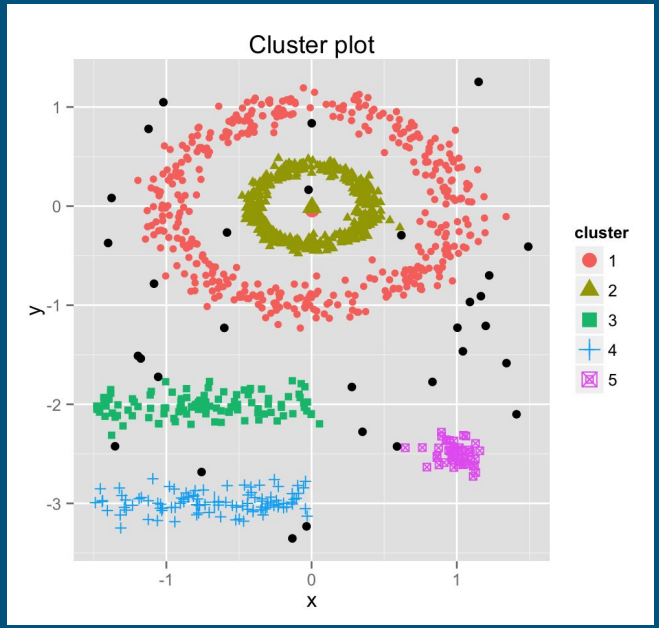
Sommaire

Régression *vs* classification

Entrées continues → sortie continue



Entrées quelconques → sortie discrète



Supervisé *vs* non supervisé

Apprentissage par l'exemple

- On questionne le modèle en lui donnant la réponse
- Définition d'un "coût" (⇔ performance) explicite à optimiser
- Entraînement sur certaines données, tests sur d'autres

Exploration

- Recherche de structures dans les données
- Réduction de dimensionnalité
- Détection d'anomalies

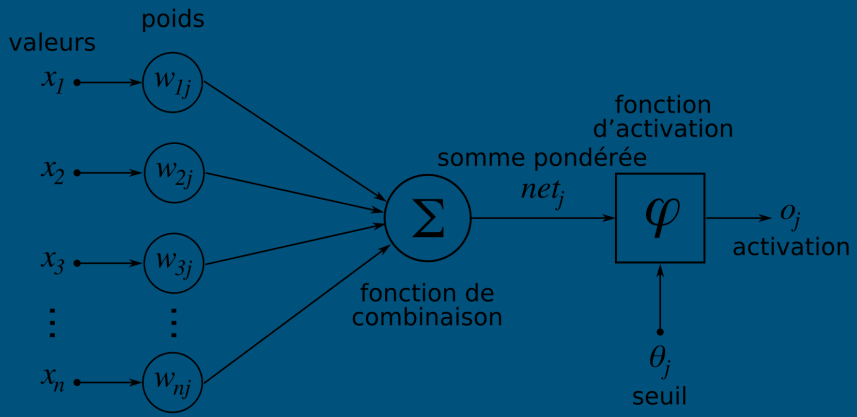
Techniques mixtes

- On peut déduire une classification après une régression (cf. régression ridge)
- Apprentissage semi-supervisé possible
- Une 3e manière d'apprendre : le renforcement

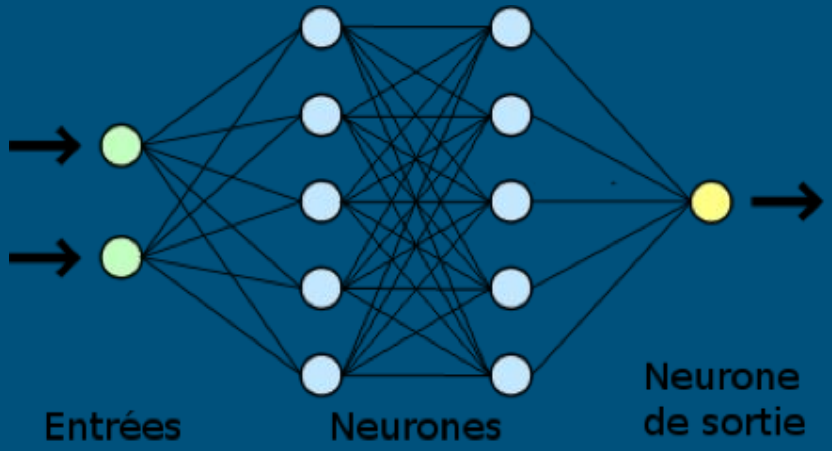
⇒ L'apprentissage brouille les frontières artificielles qu'on s'impose

Réseaux de neurones

Neurone seul



Réseau profond (un exemple)



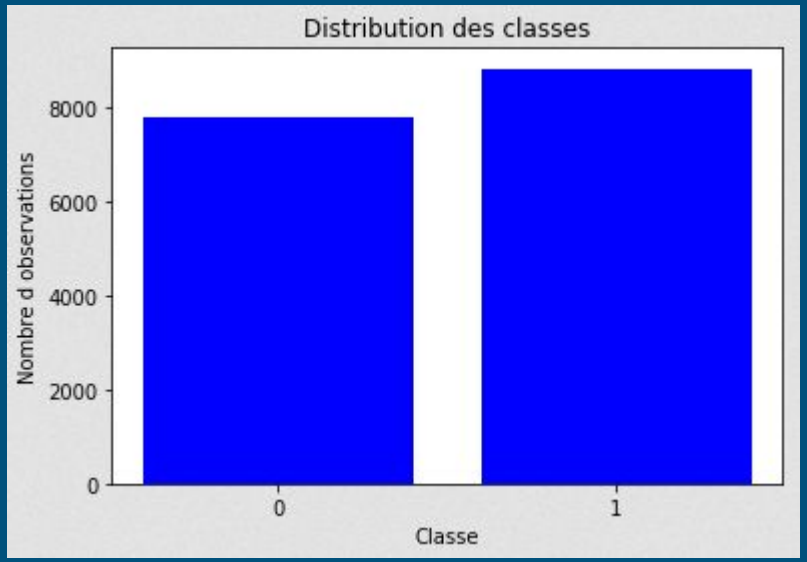
Méthodologie

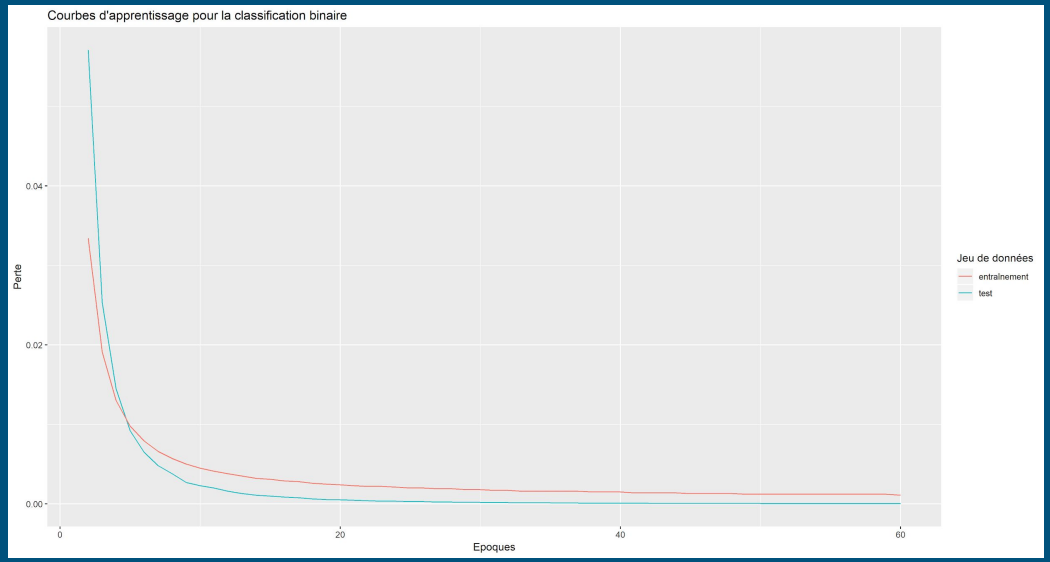
Etapes vers la résolution d’un problème d’apprentissage

Toutes les étapes sont importantes et ont des difficultés

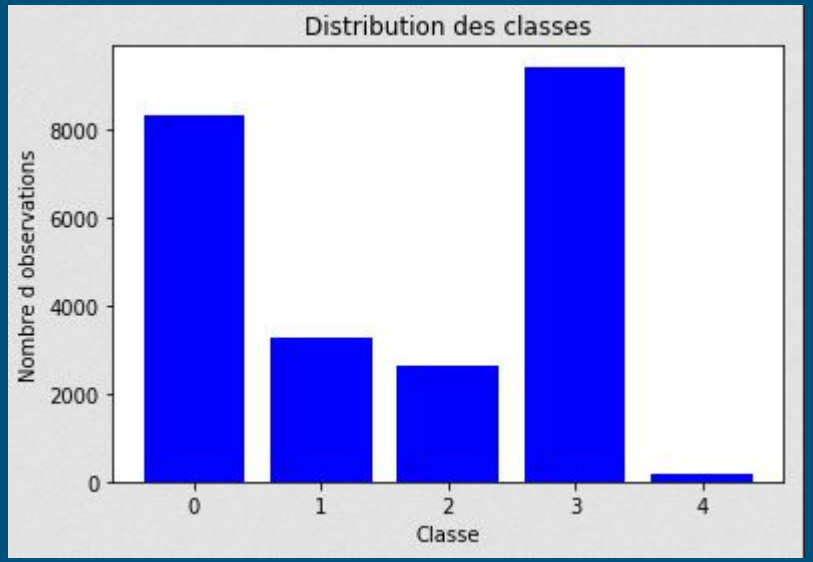
Sommaire

Classification binaire : Balzac ou Zola ?

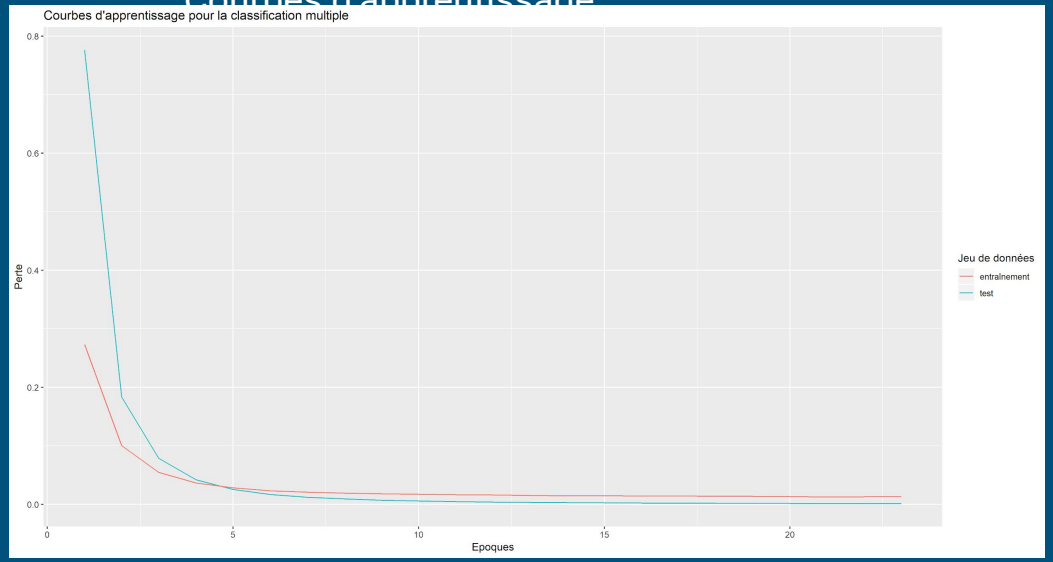




Classification multiple : quel auteur ?



Courbes d'apprentissage



Classification - Remarques

Le classifieur reconnaît-il le style ?

- Style : notion abstraite et subjective
- Modèle basé sur les bigrammes, peu informatifs
- Pas assez d'extraits pour lancer une construction sémantique
- Pour s'en assurer il faudrait générer du texte

Code (adapté de Google) disponible et testable :

https://colab.research.google.com/drive/1NZ1NrFBn2ArlpgkgI0PPFFQ0Mu0hZk0n

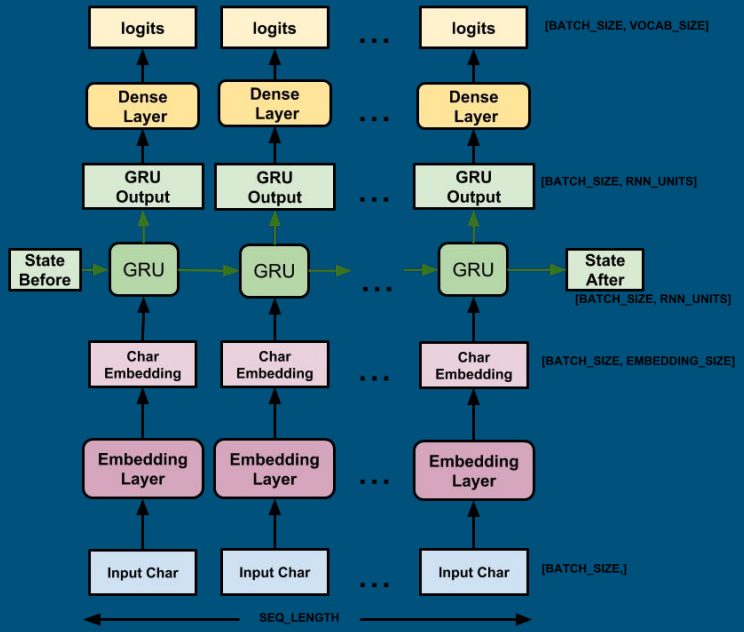
Sommaire

“ Monsieur, le derrière, en hâte. Il semblait que tout le coron se tuait de sa taille, pressé d’envoyer la crosse farouche.

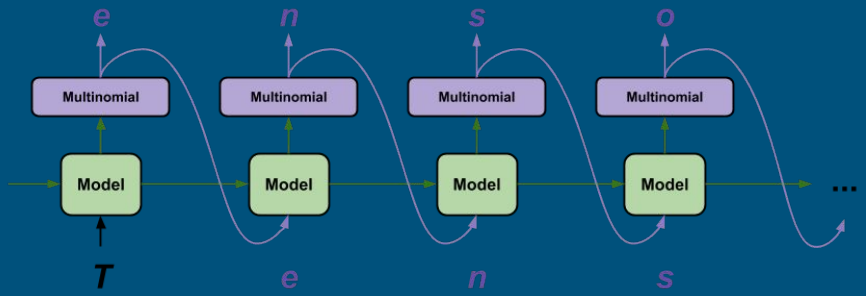
— Veux-tu bien de d’où en les autre, il faut tirer En bas de la cheminée, ils n’en routes pas avec les petits, et quand il eut découdé que deux patients, qu’ils voyaient une décoration, le flaire était achevé de défendre. Il baiserait tout de même potiller. On avait là, fondu entrer, ler maintenant, éternelle moins de toilette, M. Grégoire accrochant de tuerelle, alors, pour l’empêcher de ricaner encore, avec ses deux mères les petits jupes dont les mineurs se méfiant. S’il ne s’y donnait pas sa main, il bégaya : [...] ”

Texte “inspiré” de Germinal, généré automatiquement

Principes de la prédiction



Réseau de neurones profond récurrent



Génération - Remarques

Le générateur comprend-il le style ?

- Code basé sur les lettres et les mots, peu informatifs
- Non comparable avec le classifieur : réseaux très différents
- Pour s'en assurer il faudrait l'entraîner longtemps

Code (adapté de Google) disponible et testable :

https://colab.research.google.com/drive/1W6QNRw-n-vK-iULLNJWsOL8dSCixixlU

Sommaire

Conclusion

Multiples points positifs

- Découverte d'une science jeune et prometteuse
- Ressources très nombreuses (livres, sites internet)
- Bagage utile pour la suite (stages, etc.)

Des difficultés

- Beaucoup de notions théoriques
- Peu de temps pour tout maîtriser
- Outils informatiques très pointus