APPRENTISSAGE AUTOMATIQUE (MACHINE LEARNING) POUR LE DIAGNOSTIC DU CANCER

Abdi VURAL

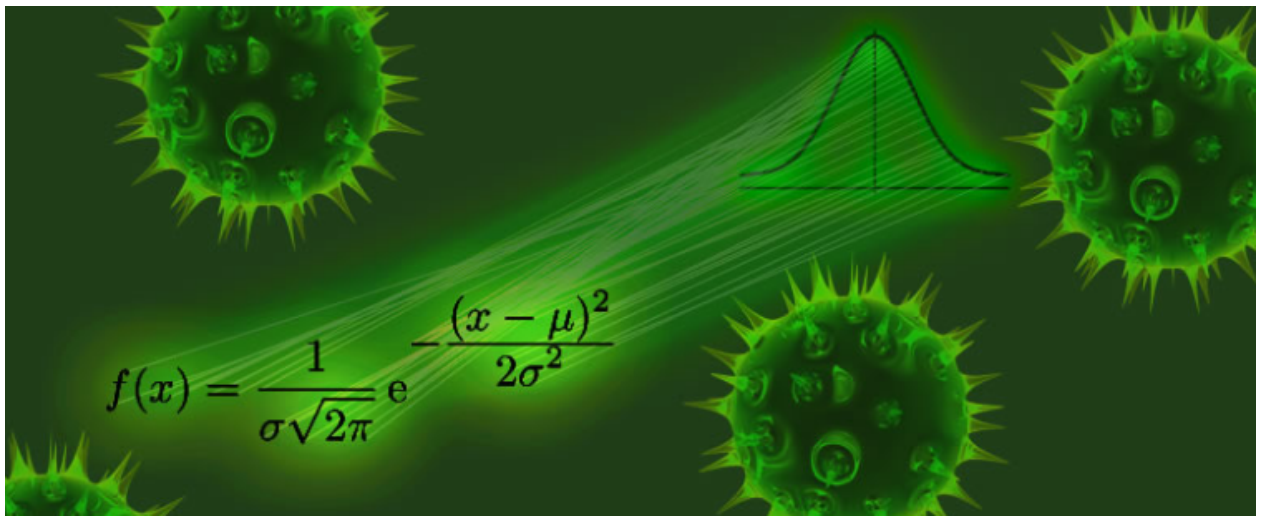
Université de Lausanne

January 22, 2020

Aperçu

1.1 Description

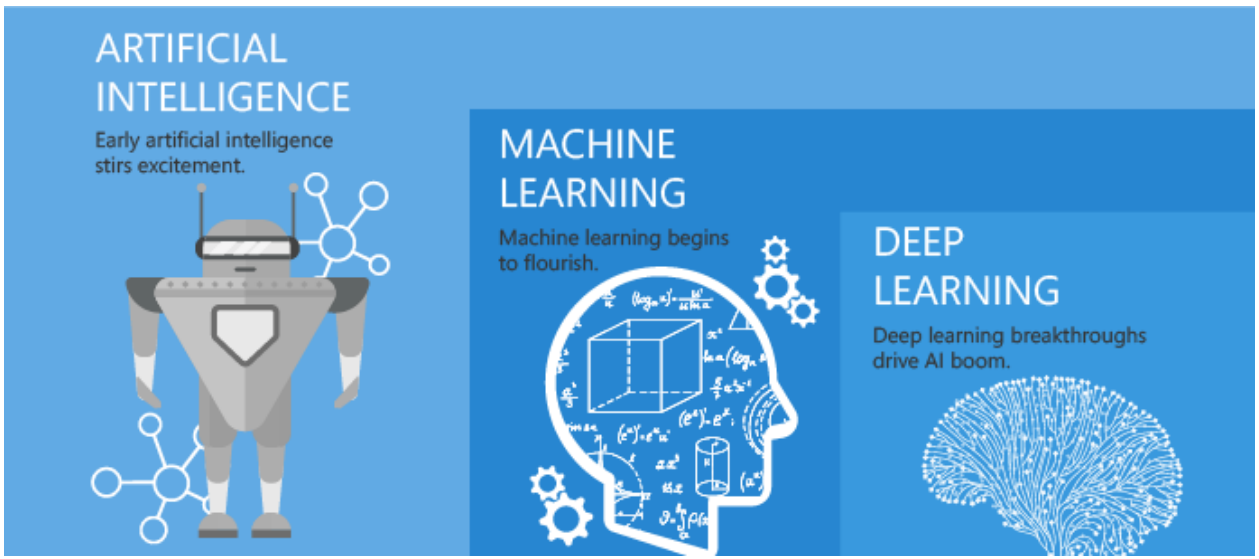
Le cancer est l'une des causes les plus fréquentes de décès dans le monde. Actuellement, le cancer du sein est le plus répandu dans les cancers féminins. Malgré les avancées significatives faites ces dernières décennies en vue d'améliorer la gestion du cancer, des outils plus précis sont toujours nécessaires pour aider les oncologues à choisir le traitement nécessaire à des fins de guérison ou de prévention de récidive tout en réduisant les effets néfastes des ces traitements ainsi que leurs coûts élevés[1].



1.1 Description

Apprentissage Automatique(Machine Learning)

L'apprentissage automatique est une branche de l'intelligence artificielle qui fait référence au développement, l'analyse et l'implémentation d'algorithme permettant à une machine d'apprendre à partir d'un ensemble de données. Ce processus est inductif, il tente de généraliser les relations extraites de l'ensemble d'apprentissage à tout l'espace de données de la source.



1.2 Objectif du projet

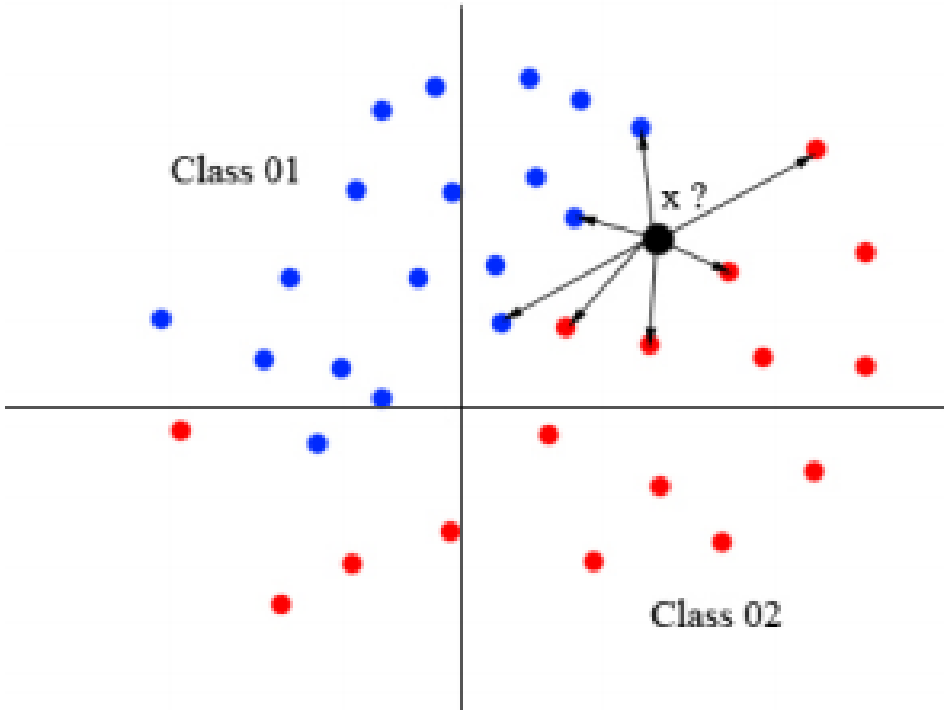
L'Apprentissage Automatique pour diagnostiquer le cancer du sein

Le cancer du sein est l'un des tueurs de cancer dans le monde. Le diagnostic de ce cancer est un gros problème dans les recherches sur le diagnostic du cancer. Dans l'intelligence artificielle, l'apprentissage automatique est une discipline qui permet à la machine d'évoluer à travers un processus. L'apprentissage automatique est largement utilisé en bio l'informatique et en particulier dans le diagnostic du cancer du sein. Un de méthodes les plus populaires est K-voisins les plus proches (K-NN) qui est un méthode d'apprentissage supervisé.

L'objectif de ce projet est de concevoir d'un algorithme de sélection de variable à la methode K-NN.

K plus proches voisins

La méthode des K plus proches voisins (KNN) a pour but de classifier des points cibles (classe méconnue) en fonction de leurs distances par rapport à des points constituant un échantillon d'apprentissage (c'est-à-dire dont la classe est connue a priori).
KNN est une approche de classification supervisée intuitive. Il s'agit d'une généralisation de la méthode du voisin le plus proche (NN). NN est un cas particulier de KNN, où k = 1.



- La distance euclidien entre deux points $\mathbf{p}(x_1, y_1)$ et $\mathbf{q}(x_2, y_2)$:

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

- Si p = (p1, p2, ..., pn) et q = (q1, q2, ..., qn) sont deux points dans l'espace n euclidien, alors la distance (d) de p à q, ou de q à p est donné par la formule de Pythagore

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Validation croisée

La validation croisée (cross-validation) est, en apprentissage automatique, une méthode d'estimation de fiabilité d'un modèle fondé sur une technique d'échantillonnage[2].

$$= \frac{\sum_{i=0}^{k} TE(t_i, DS - t_i)}{k}$$

- TE=(test, éducation), fonction de classification
- DS, jeu de données(dataset)
- k,le nombre de la division de l'échantillon
- $t_i = testset(echantillonselectionné)$

Abstract
In this paper, we classify the breast cancer of medical diagnostic data. Information gain has been adapted for feature selections. Neural fuzzy (NF), k-nearest neighbor (KNN), quadratic classifier (QC), each single model scheme as well as their associated, ensemble ones have been developed for classifications. In addition, a combined ensemble model with these three schemes has been constructed for further validations. The experimental results indicate that the ensemble learning performs better than individual single ones. Moreover, the combined ensemble model illustrates the highest accuracy of classifications for the breast cancer among all models[2].

Abstract
Breast Cancer becomes the life-threatening disease in the female. Breast Cancer can start in breast and spread to different parts of the body. Early detection and diagnosis of Breast Cancer have been pointed at as the most reliable approach to reducing the number of deaths. There are different machine learning techniques available that are widely used in various domains such as classification and prediction process. In the present study, we employed one of the most popular used machine learning technique K-Nearest Neighbor(KNN) for Wisconsin Diagnostic Breast Cancer dataset in R environment[3].

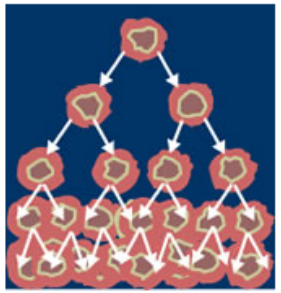
Les Ressources et La Forme Finale

- Dans ce projet, nous allons utiliser l'une des methodes d'apprentissage automatique les plus utilisées K-Nearest Neighbour (KNN) pour le Wisconsin Diagnostic Breast Cancer dataset dans l'environnment R et Phyton.
- Le lien de dataset :
  `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer`

Breast Cancer Wisconsin (Original) Data Set

Download: Data Folder, Data Set Description

Abstract: Original Wisconsin Breast Cancer Database



Data Set Characteristics:	Multivariate	**Number of Instances:**	699	**Area:**	Life
Attribute Characteristics:	Integer	**Number of Attributes:**	10	**Date Donated**	1992-07-15
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	556331

Source:

Calendrier

Table: Calendrier de réalisation du projet

Nom de la tâche	Durée	Semaine
1.Recherche sur sujet	10 heures	Février
2.Recherche bibliographiques	10 heures	Février
3.Construction de l'algorithme	20 heures	Mars
4.Surapprentissage	20 heures	Mars
5.Optimisation de l'algorithme	20 heures	Avril
6.Algorithme finale	20 heures	Avril
7.Rédaction du projet de recherche	20 heures	Mai
8.Présantation du projet de recherche	10 heures	Mai
Total heure de travail	130 heures	

Verrous

- Le nettoyage de donnèes
- La performance de la méthode choisie (KNN).
- L'utilisation d'une autre méthode d'apprentissage automatique au cas où les résultats ne sont pas satisfasiants.

Références bibliographiques

[1] Lyamine Hedjazi,*Outil d'aide au diagnostic du cancer à partir d'extraction d'informations issues de bases de données et d'analyses par biopuces*

[2] Sheau-Ling Hsieh Sung-Huai Hsieh Po-Hsun Cheng Chi-Huang Chen Kai-Ping Hsu I-Shun Lee Zhenyu Wang Feipei Lai. *Design Ensemble Machine Learning Model for Breast Cancer Diagnosis*, 3 August 2011

[3] Arpita Joshi and Ashish Mehta.2018, *Analysis of K- Nearest Neighbor Technique for Breast Cancer Disease Classification*. Int JRecent Sci Res. 9(4), pp. 26126-26130.
DOI:`http://dx.doi.org/10.24327/ijrsr.2018.0904.1997`

[4] Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi,Abdelkader Benyettou.*Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules*,International Journal of Computer Applications (0975 - 8887) Volume 62 - No. 1, January 2013