le cnam

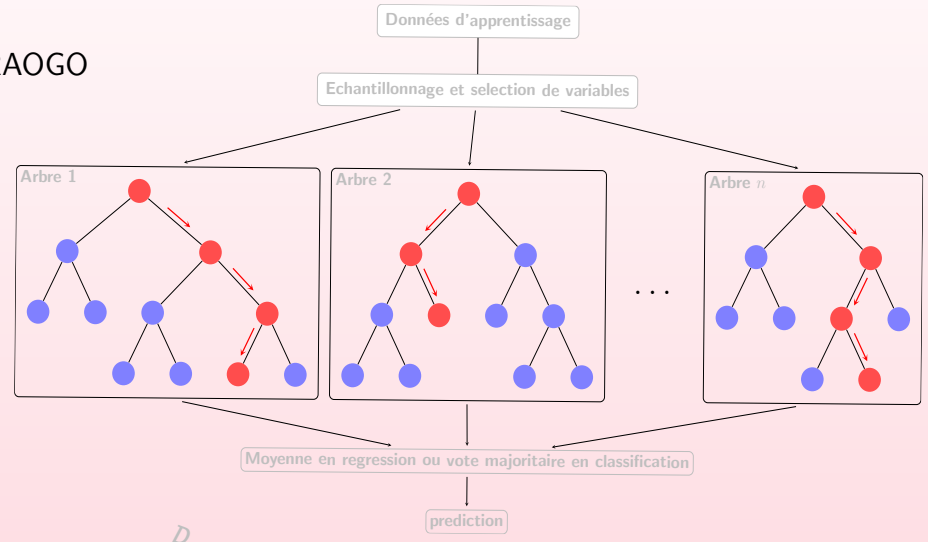
Certificat de spécialisation analyste de données massives

Unité d'enseignement : Entreposage et fouilles de données (STA211)

Synthèse du cours

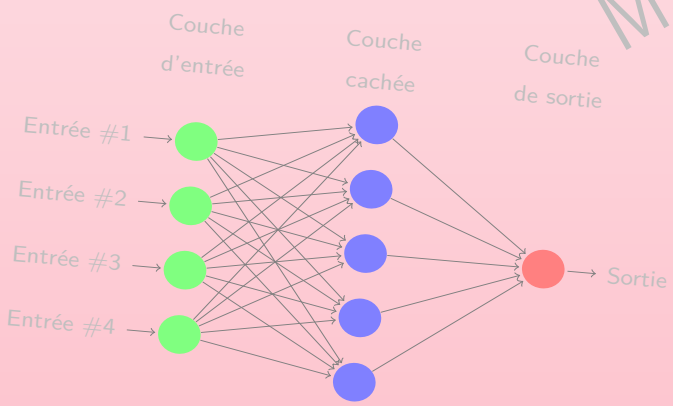
Réalisé par :

Boukary OUEDRAOGO



$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$



Professeur :
Mme Niang

Année académique :
2021-2022

Résumé : Avec l'explosion des données de ces dernières années, le besoin de tirer de la valeur des grands gissements de données est de plus en plus d'actualité. Les recherches théoriques qui restaient jadis dans les centre de recherche ont eu un regain d'intérêt grâce notamment au évolutions technologiques dans les domaines de l'informatique. Ces évolutions technologiques ont rendu possible le stockage et le traitement des données de grandes dimensions. Ces données opérationnelles qui sont optimisées pour le stockage et pas pour le traitement nécessité une certaine ingénierie pour permettent aux entreprises de tirer de la valeur de celles-ci.

Mots-clés : Analyse statistique des données, Statistique décisionnelle, Data mining, Base de données, fouille de donnees

1. Définition et contexte d'émergence du data mining

1.1 Définition

Il n'est pas aisé de trouver une définition exacte et claire associé au *data mining*. Ce que l'on peut retenir c'est le *data mining* désigne un ensemble de méthodes à l'intersection entre plusieurs disciplines (statistiques, intelligence artificielle, informatique,...) permettant d'explorer et d'analyser de grandes bases de données en vue d'extraire des informations utiles. Dans le langage courant, on confond souvent les termes *big data* et *data mining*. Ces deux termes désignent pourtant des choses différentes. Le*big data* renvoie à une quantité importante de données. Quand au *data mining*, il désigne l'utilisation des techniques diverses pour extraire de ces énormes quantités de données (big data), de la connaissance utile permettant de tirer un avantage comparative.

A partir de quand parle t'on de big data ? On parlera de big data quand les techniques traditionnelles de traitement des données ne sont plus opérantes sur les données. Le big data ne se résume pas seulement au seul aspect volumétrie. En effet, le volume est un aspect relatif. En effet avec l'amélioration de capacité de stockage, des données jadis considérées comme volumineuses deviennent moins problèmatiques pour les entreprises. Le big data est souvent caractérisé par ce qu'on appelle les **3V**. Les 3V font référence aux trois aspects suivants : volume, vitesse, variété.

- **Le volume** désigne la quantité de données astronomiques générées.
- **La vitesse** est la rapidité à laquelle les données sont générées, stockées, transformées et exploitées.
- **La variété** fait référence à la diversité des sources de données et leurs types (données structurées, données non-structurées). Les données prennent le plus souvent des formes très variées et très hétérogènes (son, images, vidéos,textes, etc.)

Aux 3V ci-dessus, certains y adjoignent deux autres V pour en faire 5. Ces 2V additionnels sont :

- La valeur : La valeur désigne la valeur ajoutée
- la véracité fait référence à la fiabilité des données

1.2 Dans quel contexte émergent le data mining ?

Les raisons du succès du data mining s'explique principalement par deux facteurs :

- L'environnement concurrentiel : La mondialisation a intensifié la concurrence entre les entreprises et saturé les marché. Dans ce environnement de plus en plus concurrentiel le client devient l'acteur principal de l'entreprise. Cela fait naître chez les entreprises le besoin de mieux connaître leurs clients.
- L'état de l'entreprise : Le développement des systèmes d'informations a entraîné une réduction des coûts de stockages et augmentation des puissances de calcul permettant aux entreprises de stocker et d'exploiter des quantités gigantesques de données. Avec la concurrence et la disponibilité des données, le besoin de transformer les données d'entreprises en connaissance s'est fait sentir. Avant toute démarche de data mining il faut être capable de poser de manière clair le problème business auquel on souhaite répondre. Est ce que les données disponibles permettent de répondre à la question qu'on se pose.

  Il existe deux types de data mining. Data mining de vérification et data mining de découverte.

2. Analyse de données

2.1 Analyse en composantes principales

2.2 Analyse factorielle des correspondances

2.3 Analyse des correspondances multiples

2.4 Analyse factorielles mixtes

```
library(dplyr)
library(tidyverse)
library(FactoMineR)
library(missMDA)
library(caret)
```

3. SVM

Les supports vecteurs machines

4. Business intelligence avec BO

Références

[1] Donald E. Knuth. Computer programming as an art. *Commun. ACM*, pages 667–673, 1974.