Interrogation IMC : 30 minutes

Exercice 1

Une école propose les 3 modules suivants : « machine learning (ML) », « big data (BD) » et « information retrievel (IR) ».

L'étudiant 1 a choisi de faire les 3 modules ;
L'étudiant 2 a choisi « machine learning » et « big data » ;
L'étudiant 3 a choisi « machine learning » et « information retrievel » ;
L'étudiant 4 a choisi « big data » et « information retrievel » ;
L'étudiant 5 a choisi les 3 modules.

A. Que permet cette équation d'estimer ?

$$\begin{pmatrix}1\\1\\1\\0\\1\end{pmatrix} = \alpha \begin{pmatrix}1\\1\\0\\1\\1\end{pmatrix} + \beta \begin{pmatrix}1\\0\\0\\1\\1\end{pmatrix} + \varepsilon$$

1- L'influence du choix des modules « BD » et « IR » sur le choix du module « ML »
2- L'influence du choix des modules « ML » et « BD » sur le choix du module « IR »
3- L'influence du choix des modules « ML » et « IR » sur le choix du module « BD »

B. Donnez les valeurs des coefficients α et β en utilisant la fonction lm de R.
C. Qu'en concluez-vous ?

Exercice 2 :

Soit $X = \{1, 2, 9, 12, 20\}$

A. Appliquer l'algorithme des K-Means pour identifier 2 clusters.

B. A l'inertie intra-classes, on rajoute le terme suivant : $2kNlogN$. Ce terme permet de jouer sur la complexité du modèle (principe du rasoir d'Occam). Il s'agit de trouver le modèle qui regroupe le mieux les points, tout en évitant de "sur-apprendre". On obtient le résultat suivant :

k	2	3	4	5
J_w	70	5	0.5	0
$2kNlogN$	32.1888	48.2831	64.3775	80.4719
J_w^{new}	102.1888	53.2831	64.8775	80.4719

Quel est le meilleur regroupement possible ?
1- k=2
2- k=3
3- k=4