

TD Extraction

1. Soit $T(A, B, C, D, E, \text{Classe})$ une table relationnelle à partir de laquelle on construit un arbre de décision permettant d'expliquer ou de prédire l'attribut Classe en fonction des autres attributs. Est-il possible que dans l'arbre de décision obtenu on retrouve les deux règles
r1 : $A=a1, B=b1, C=c1 \rightarrow \text{Classe}=1$ et
r2 : $A=a1, B=b1, C=c1, E=e1 \rightarrow \text{Classe}=1$?
Expliquer.
2. Expliquer comment une mesure de distance entre données de type booléen peut être utilisée pour comparer des données définies par des variables de type catégoriel.
3. Soit une base de transactions T qui est partitionnée en $T_1, T_2, \dots, T_m$. Montrer que si un itemset X est fréquent dans T , c'est-à-dire sa fréquence est supérieure ou égale à un seuil s dans T , alors il existe au moins une part T_i telle que X est fréquent dans T_i
4. Reprendre le fichier DeuxCarrés vu au TD précédent. On a vu qu'avec $\text{Eps}=1.1$ et $\text{MinPts}=2$, l'algorithme DBSCAN permet de retrouver les deux carrés en les considérant comme deux groupes distincts. On a vu que K-means était incapable d'identifier ces deux carrés. Est-il possible de retrouver ces deux mêmes carrés en utilisant un algorithme hiérarchique agglomératif ? Expliquer
5. Soit $R(A, B, C, D, E, M)$ une table relationnelle à partir de laquelle on veut évaluer toutes les requêtes de la forme

```
SELECT X, SUM(M)
FROM R
GROUP BY X
HAVING COUNT(*) > S
```

Où S est un nombre entier et X est un sous-ensemble de $\{A, B, C, D, E\}$

En vous inspirant d'A priori, proposer une méthode efficace pour évaluer cet ensemble de requêtes.