**Université Mohamed Boudiaf M'sila** **M2 SIGL**
**Faculté MI, Dépt Informatique** **M'sila, Le 17/01/2023**

# Examen Final

## Data Mining et Recherche d'Information

**Exercice 1 (6 points) :** Répondre brièvement à ces questions.

**Question 1.1 :** Donner le principe général de l'algorithme K-means. **(1 pt)**
**Question 1.2 :** Quelle est la complexité de K-means ? **(1 pt)**
**Question 1.3 :** Donner deux inconvénients de K-means. **(2 pts)**
**Question 1.4 :** Quelles sont les conditions d'arrêt de l'algorithme K-means ? **(2 pts)**

**Exercice 2 (7 points)**

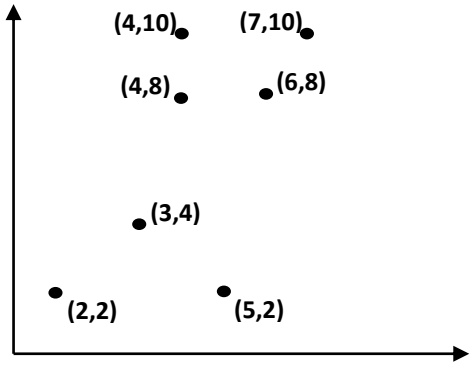
Les données suivantes permettent à un robot de prédire si un objet ayant certaines caractéristiques (Colour, Shape, Size) appartient à la classe (+) ou la classe (-). On désire utiliser les arbres de décision pour réaliser cet apprentissage.

| *Example* | *Colour* | *Shape* | *Size* | *Class* |
|---|---|---|---|---|
| **1** | Red | Square | Big | + |
| **2** | Blue | Square | Big | + |
| **3** | Red | Circle | Big | + |
| **4** | Red | Circle | Small | - |
| **5** | Green | Square | Small | - |
| **6** | Green | Square | Big | - |

**Question 2.1 :** Quel attribut peut-on choisir comme racine de l'arbre si on utilise le Gain en information? **(2 pts)**
**Question 2.2 :** Construire l'arbre de décision complet à partir de ces données. **(5 pts)**

**Exercice 3 (7 points)**

Soit un nuage de 7 points représentés dans un espace bidimensionnel selon la figure ci-dessous.




**Question 3.1 :** Donner un clustering rapide de ces points, selon votre visualisation.**(1 pt)**
**Question 3.2 :** Appliquer l'algorithme k-Means sur cet ensemble de points, avec k=nombre de clusters choisi à la Question 3.1) **(3 Pts).** Comment les centres initiaux ont été choisis ? **(0.5 pt)**
**Question 3.3 :** Comparer votre clustering rapide avec les résultats de k-means **(1 pt)**. Quels sont les avantages et les inconvénients de cette stratégie de clustering? **(1.5 pts)**

**BONNE REUSSITE**

**Dr. Tahar Mehenni**

# Correction de l'Examen Final

## Data Mining et Recherche d'Information

### Exercice 1 (6 points)

Question 1.1 : Donner le principe général de l'algorithme K-means.

Réponse : 1. Choisir k objets formant ainsi k clusters
2. (Ré)affecter chaque objet O au cluster Ci de centre Mi tel que dist(O,Mi) est minimal
3. Recalculer Mi de chaque cluster (le barycentre)
4. Aller à l'étape 2 si on vient de faire une affectation

***(1 Pt)***

Question 1.2 : Quelle est la complexité de K-means ?

Réponse : O(tkn), où n est # objets, k est # clusters, et t est # itérations ***(1 Pt)***

Question 1.3 : Donner deux inconvénients de K-means.

Réponse : - La nécessité de définir le nombre de clusters à priori.
- Les clusters sont construits par rapports à des objets inexistants (les milieux)

***(2 Pts)***

Question 1.4 : Quelles sont les conditions d'arrêt de l'algorithme K-means?

Réponse : - Nombre d'itérations
- Plus de changement sur les centres de gravité (ou changements limités)
- Pas de changement dans la composition des clusters
- Temps d'exécution

***(0.5x4=2 Pts)***

### Exercice 2 (7 points)

| *Colour* | *Shape* | *Size* | *Class* |
|---|---|---|---|
| Red | Square | Big | + |
| Blue | Square | Big | + |
| Red | Circle | Big | + |
| Red | Circle | Small | – |
| Green | Square | Small | – |
| Green | Square | Big | – |

n= 6 E= 1 **(0.5 Pt)**

C1= 3

C2= 3

| **Colour** | + | - | Somme | ni/n | | Entropie | | Gain |
|---|---|---|---|---|---|---|---|---|
| P1 (Red) | 2 | 1 | 3 | 0.50 | 0.91830 | 0.459148 | 0.45915 | 0.54085 |
| P2 (Blue) | 1 | 0 | 1 | 0.17 | 0.00000 | 0 | | |
| P3 (Green) | 0 | 2 | 2 | 0.33 | 0.00000 | 0 | | |

**(0.5 Pt)**

| **Shape** | + | - | Somme | ni/n | | Entropie | | Gain |
|---|---|---|---|---|---|---|---|---|
| P1 Square) | 2 | 2 | 4 | 0.67 | 1.00000 | 0.666667 | 1 | 0 |
| P2 (Circle) | 1 | 1 | 2 | 0.33 | 1.00000 | 0.333333 | | |

**(0.5 Pt)**

| **Size** | + | - | Somme | ni/n | | Entropie | | Gain |
|---|---|---|---|---|---|---|---|---|
| P1 (Big) | 3 | 1 | 4 | 0.67 | 0.81128 | 0.540852 | 0.54085 | 0.45915 |
| P2 (Small) | 0 | 2 | 2 | 0.33 | 0.00000 | 0 | | |

**(0.5 Pt)**

Attribut choisi comme racine: Colour  E(Colour)= 0.540852 **(0.5 Pt)**

Pour Colour = Red

| *Colour* | *Shape* | *Size* | *Class* |
|---|---|---|---|
| Red | Square | Big | + |
| Red | Circle | Big | + |
| Red | Circle | Small | – |

n= 3
C1= 2
C2= 1 **(0.5 Pt)**
E= 0.9183

| **Shape** | + | - | Somme | ni/n | | Entropie | | Gain |
|---|---|---|---|---|---|---|---|---|
| P1 Square) | 1 | 0 | 1 | 0.33 | 0.00000 | 0 | 0.6667 | 0.2516 |
| P2 (Circle) | 1 | 1 | 2 | 0.67 | 1.00000 | 0.66667 | | |

**(0.5 Pt)**

| **Size** | + | - | Somme | ni/n | | Entropie | | Gain |
|---|---|---|---|---|---|---|---|---|
| P1 (Big) | 2 | 0 | 2 | 0.67 | 0.00000 | 0 | 0 | 0.9183 |
| P2 (Small) | 0 | 1 | 1 | 0.33 | 0.00000 | 0 | | |

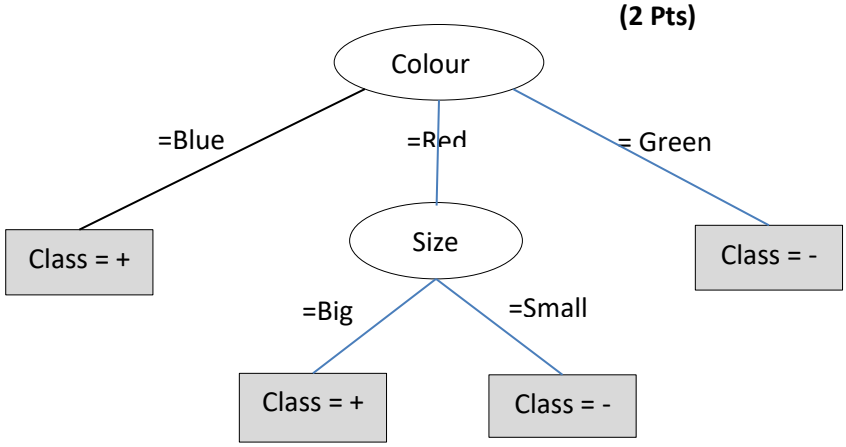
**(0.5 Pt)**

Attribut choisi pour le branchement: **size**  Size= Big → Class= +  Size= Small → Class= -  **(0.5 Pt)**

Colour= Blue → Class= +
Colour= Green → Class= -  **(0.5 Pt)**

**Arbre de decision**

**(2 Pts)**




## Exercice 3 (7 points)

Coordonnées des 7 points du nuage

| A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| (4,10) | (7,10) | (4,8) | (6,8) | (3,4) | (2,2) | (5,2) |

Question 3.1 : Clustering rapide à vue : on peut voir deux clusters : C1=(A,B,C,D), C2=(E,F,G) **(1 Pt)**

Question 3.2 : Application de K-Means. On choisit k=2 selon la réponse à la question 3.1 **(0.25 Pt)**

Initialisation : C1 = {}, C2 = {}, avec c1 = A , c2 = E Les centres sont choisis selon le clustering trouvé à la question 3.1 **(0.5 Pt)**

**Etape 1 :**

Pour B : $d^2(c1,B) = |7-4|^2+|10-10|^2=9$, $d^2(c2,B) = |7-3|^2+|10-4|^2=52$ d'où C1 = C1 + {B}
Pour C : $d^2(c1,C) = |4-4|^2+|8-10|^2=4$, $d^2(c2,C) = |4-3|^2+|8-4|^2=17$ d'où C1 = C1 + {C}
Pour D : $d^2(c1,D) = |6-4|^2+|8-10|^2=8$, $d^2(c2,D) = |6-3|^2+|8-4|^2=25$ d'où C1 = C1 + {D}
Pour F : $d^2(c1,F) = |2-4|^2+|2-10|^2=68$, $d^2(c2,F) = |2-3|^2+|2-4|^2=5$ d'où C2 = C2 + {F}
Pour G : $d^2(c1,G) = |5-4|^2+|2-10|^2=65$, $d^2(c2,G) = |5-3|^2+|2-4|^2=8$ d'où C2 = C2 + {G}

**(1 Pt)**

- ⇨ C1 = { A,B,C,D }, c1 = ((4+7+4+6)/4, (10+10+8+8)/4)= (5.25,9)
- ⇨ C2 = { E,F,G,}, c2 = ((3+2+5)/3, (4+2+2)/3)= (3.33,2.66)

**(0.25 Pt)**

**Etape 2 :**

C1 = {}, C2 = {} avec c1 = (5.25,9) , c2 = (3.33,2.66) **(0.25 Pt)**

Pour A : $d^2(c1,A) = |4-5.25|^2+|10-9|^2=2.56$, $d^2(c2,A) = |4-3.33|^2+|10-2.66|^2=54.32$ d'où C1 = C1 + {A}
Pour B : $d^2(c1,B) = |7-5.25|^2+|10-9|^2=4.06$, $d^2(c2,B) = |7-3.33|^2+|10-2.66|^2=67.34$ d'où C1 = C1 + {B}
Pour C : $d^2(c1,C) = |4-5.25|^2+|8-9|^2=2.56$, $d^2(c2,C) = |4-3.33|^2+|8-2.66|^2=28.96$ d'où C1 = C1 + {C}
Pour D : $d^2(c1,D) = |6-5.25|^2+|8-9|^2=2.56$, $d^2(c2,D) = |6-3.33|^2+|8-2.66|^2=36.64$ d'où C1 = C1 + {D}
Pour E : $d^2(c1,E) = |3-5.25|^2+|4-9|^2=30.06$, $d^2(c2,E) = |3-3.33|^2+|4-2.66|^2=1.90$ d'où C2 = C2 + {E}
Pour F : $d^2(c1,F) = |2-5.25|^2+|2-9|^2=59.56$, $d^2(c2,F) = |2-3.33|^2+|2-2.66|^2=0.54$ d'où C2 = C2 + {F}
Pour G : $d^2(c1,G) = |5-5.25|^2+|2-9|^2=49.06$, $d^2(c2,G) = |5-3.33|^2+|2-2.66|^2=3.22$ d'où C2 = C2 + {G}

**(1 Pt)**

- ⇨ Il y a stabilité des clusters. Les clusters sont C1 = { A,B,C,D }, C2= { E,F,G } **(0.25 Pt)**

(a) K-means a donné les mêmes résultats que le clustering rapide. **(1 Pt)**

Avantages : **(1 Pt)**

- k est choisi selon le clustering rapide
- Convergence rapide (stabilité rapide, nombre d'étapes réduit, temps d'exécution court) .
- Confirmation du clustering à vue.

Inconvenients : **(0.5 Pt)**

- Résultat biaisé, dû au clustering rapide (choix de k et des centres initiaux)

**Rédigé par T. Mehenni**