

## EXAMEN NORMAL DE MAI 2022

### TECHNIQUE AVANCÉES DE DATA MINING

Examinateur : Dr OLLE OLLE Niveau d'étude : MASTER 2 Durée : 2 heures Documents : oui, sauf internet

#### Exercice 1

##### Question de cours

1. Quels sont les grands défis actuels de l'analyse des grands volumes de données ?
2. Définissez, représentez et expliquez chaque étape du CRISP-DM.
3. Consider the problem of finding the  $K$  nearest neighbors of a data object. A programmer designs Algorithm this task.

##### Algorithm

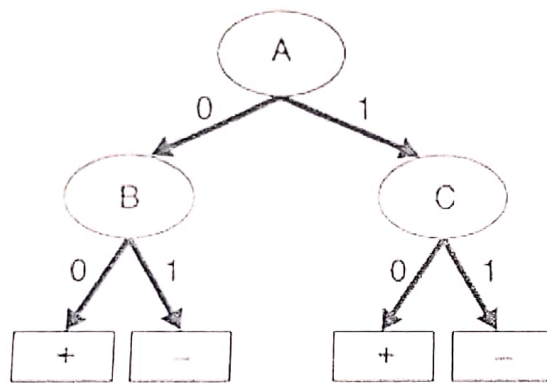
1. for  $i = 1$  to number of data object do
  2. Find the distance of the  $i^{th}$  object to all other object.
  3. Sort these distances in decreasing order.  
(Keep track of which object is associated with each distance.)
  4. return the object associated with the first  $K$  distances of the sorted list
  5. end for
- a) Describe the potential problems with this algorithm if there are duplicate objects in the data set. Assume the distance function will only return a distance of 0 for objects that are the same.
- b) How would fix this problem?

#### Exercice 2

Consider the decision tree show in Figure 1

- a) Compute the generalization error rate of the tree using the optimistic approach.

- b) Compute the generalization error rate of the tree using pessimistic approach. (For simplicity, use the strategy of adding a factor of 0.5 to each leaf node.)
- c) Compute the generalization error rate of the tree using the validation set shown above. This approach is known as *reduced error pruning*.



Training:

Instance	A	B	C	Class
1	0	0	0	+
2	0	0	1	+
3	0	1	0	+
4	0	1	1	-
5	1	0	0	+
6	1	0	0	+
7	1	1	0	-
8	1	0	1	+
9	1	1	0	-
10	1	1	0	-

Validation:

Instance	A	B	C	Class
11	0	0	0	+
12	0	1	1	+
13	1	1	0	+
14	1	0	1	-
15	1	0	0	+

Figure 1: Decision tree and data sets for Exercice 2

## Problème

En raison des relations étroites entre plusieurs départements ou entreprises, il est nécessaire en fouille des données d'effectuer une exploration de données sur plusieurs bases de données interconnectées. En comparaison avec l'exploration de données multi relationnelle, une difficulté majeure avec l'exploration de plusieurs bases de données est l'hétérogénéité sémantique entre les bases de données. Par exemple : "William Nelson" dans une base de données pourrait être "Bill Nelson" ou "B Nelson" dans une autre.

1. Proposez une méthode d'exploration de données qui peut consolider ces objets en explorant les liens d'objet entre plusieurs bases de données.
2. Décrivez maintenant une méthode efficace qui peut effectuer une classification dans plusieurs bases de données hétérogènes.

### Astuce pour la question 1 :

Dans la réconciliation d'objets, la tâche consiste à prédire si deux objets sont, en fait, identiques, en fonction de leur attributs et liens. Cette tâche est courante dans l'extraction d'informations, l'élimination

des duplications(déduplication), consolidation d'objets et l'appariement de citations, également appelé couplage d'enregistrement ou incertitude d'identité.

**Astuce pour la question 2 :**

Proposez juste des objectifs, méthodes ou algorithmes et décrire comment on s'y prendra par des exemples.

**Bonne chance !**



EXAMEN NORMAL DE AVRIL 2021

TECHNIQUES AVANCEES DE DATA MINING

Examineur : Dr OLLE OLLE Niveau d'étude : MASTER 2 Durée : 3 heures Documents : oui,  
sauf internet

Exercice 1. Question de Cours

1. Quels sont les grands défis actuels de l'analyse des grands volumes de données ? (1 pts)
2. Citez et expliquez-en quoi consistent les 2 grandes catégories d'activités de Fouille des données. (1pt)
3. Définissez, représentez et expliquez-chaque étape du CRISP-DM (2 pts)
4. Présentez et expliquez l'architecture globale d'un système décisionnel dans le contexte d'implémentation d'un système de fouille de données en OLAP? (1pt)

Problème 1

Suppose your task as a software engineer at Big-University is to design a data mining system to examine their university course database, which contains the following information: the name, address, and status (e.g., undergraduate or graduate) of each student, the courses taken, and their cumulative grade point average (GPA). You are given the architecture you would choose as to be:

- ✓ A database, data warehouse, or other information repository
- ✓ A database or data warehouse server
- ✓ A knowledge base
- ✓ A data mining engine
- ✓ A pattern evaluation module
- ✓ A graphical user Interface

Describe and explain what will be the purpose of each component of this architecture?

Page 109

## **Problème 2**

En raison des relations étroites entre plusieurs départements ou entreprises, il est nécessaire de fouiller des données d'effectuer une exploration de données sur plusieurs bases de données interconnectées. En

Comparaison avec l'exploration de données multi relationnelle, une difficulté majeure avec l'exploration de plusieurs bases de données est l'hétérogénéité sémantique entre les bases de données. Par exemple : « William Nelson » dans une base de données pourrait être « Bill Nelson » ou « B Nelson » dans une autre.

1. Proposez une méthode d'exploration de données qui peut consolider ces objets en explorant les liens d'objet entre plusieurs bases de données.
2. Décrivez maintenant une méthode efficace qui peut effectuer une classification dans plusieurs bases de données hétérogènes

### **Astuce pour la question 1 :**

Dans la réconciliation d'objets, la tâche consiste à prédire si deux objets sont, en fait, identiques, en fonction de leur attributs et liens. Cette tâche est courante dans l'extraction d'informations, l'élimination des duplications (déduplication), consolidation d'objets et l'appariement de citations, également appelé couplage d'enregistrements ou incertitude d'identité.

### **Astuce pour la question 2**

Proposez juste des objectifs, méthodes ou algorithmes et décrivez comment on s'y prendra par des exemples.

**Bonne chance !**

Composition de fin de semestre 1 2021/2022

Epreuve de INF5029 : Science de Données

Durée : 3 heures

**NB :** documents, téléphones, ordinateurs, calculatrices fermés. Ne rien écrire, ni dessiner au crayon.

**Partie 1 : SID**

**10 points**

**Questions de cours**

1. Faites une comparaison entre un SIO et un SID sur trois critères que vous choisissez
2. Faites une comparaison entre l'architecture des entrepôts suivant B. Inmon et l'architecture suivant R. Kimball sur trois critères que vous choisissez.
3. Définir : indicateur, dimension, sujet, datamart

**Application**

Une entreprise de fabrication de pâtes alimentaires souhaite mettre en place un système d'information décisionnel sous la forme d'un datamart pour observer son activité de ventes aux niveaux des différents lieux de distributions de ses articles et cela dans plusieurs villes. Ces lieux de distributions sont renseignés par leur enseigne, leur type (en fonction de leur surface), leur adresse (arrondissement et localité), leur département, leur région. Les ventes sont renseignées selon une période qui se décline en mois, en trimestre et année. Les ventes sont observées par le nombre d'articles selon le type, et le chiffre d'affaire.

1. Quel est le fait à observer ? quels sont les indicateurs ?
2. Quels sont les axes d'analyse ?
3. Construire le modèle en étoile de ce datamart.
4. Construire le modèle en flocon de datamart.

**Partie 2 : Datamining**

**10 points**

**Questions de cours**

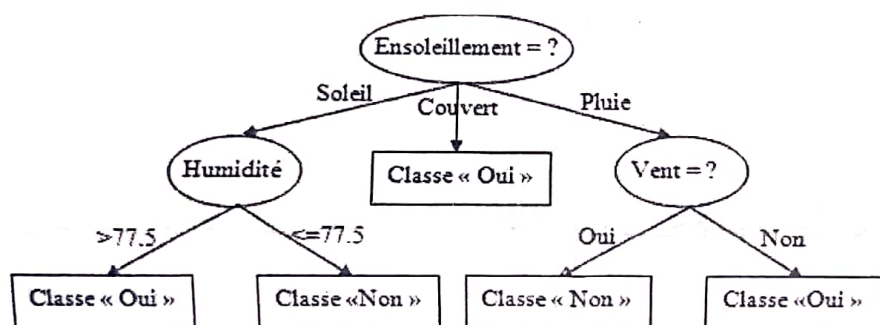
1. Dans votre spécialité, décrire un problème qui peut être résolu par une approche de fouille de données
2. En quoi consiste le problème ?
3. Comment sont décrites les données ?
4. Comment sont collectées les données
5. Votre problème peut être transformé comment en quel problème de fouille de données ? Comment évaluer la solution ?

**Problème**

Soit un ensemble de données décrites avec quatre attributs : Ensoleillement, Température, Humidité, Vent et l'attribut à prédire Jouer. L'arbre de décision de la page suivante a été construit pour classer les données du tableau (de la même page)

1. Décrire le principe validation croisé d'ordre k et celui du holdout
2. Donner le principe général de construction d'un arbre de décision.
3. Evaluer la précision, le rappel et la F-mesure de cet arbre. Donner une interprétation de chacune de ces mesures.

N°	Ensoleillement	Température	Humidité	Vent	Jouer
1	Soleil	75	70	Oui	Oui
2	Soleil	80	90	Oui	Non
3	Soleil	85	85	Non	Non
4	Soleil	72	95	Non	Non
5	Soleil	69	70	Non	Oui
6	Couvert	72	90	Oui	Oui
7	Couvert	83	78	Non	Oui
8	Couvert	64	65	Oui	Oui
9	Couvert	81	75	Non	Non
10	Pluie	71	80	Oui	Non
11	Pluie	65	70	Non	Non
12	Pluie	75	80	Non	Oui
13	Pluie	68	80	Non	Oui
14	Pluie	70	96	Oui	Oui



## Master 2

### DATASCIENCE – Durée : 2 heures

NB : Documents, téléphones, calculatrices et ordinateurs fermés

#### SID

5 points

1. Pourquoi dit-on que les systèmes d'information opérationnels ne sont pas appropriés au pilotage de l'entreprise ?
2. Quelles sont les opérations du processus ETL, dire en quoi consiste chacune d'elles.
3. Que signifie base de données normalisées? Données dénormalisées ? Présenter les avantages et les inconvénients dans chaque cas

#### Problème (SID)

7 points

Une entreprise de location de véhicules vous demande de lui concevoir un entrepôt de données, dans le but d'optimiser son fonctionnement et/ou son processus de prise de décision. Ses souhaits sont relativement vagues, on vous demande donc de l'aider à définir le contenu, les buts et l'architecture de cet entrepôt de données. La compagnie est à l'échelle internationale : elle se compose de plusieurs sociétés nationales.

Chaque société nationale comprend un grand nombre de succursales couvrant chacun des pays.

- Les clients peuvent réserver un véhicule au guichet, par téléphone, par internet.
  - Les clients peuvent rendre le véhicule dans n'importe laquelle des agences de la marque.
  - Lorsqu'un client rend un véhicule, on connaît le nombre de kilomètres qu'il a parcouru, et les éventuels incidents qui ont eu lieu pendant la location (panne, accident. . .).
  - Il existe plusieurs catégories de véhicules.
  - Il existe plusieurs tarifs de location (à la journée, au kilomètre, à la semaine . . .)
  - La compagnie a souvent recours à des promotions sur ses tarifs.
1. Identifiez trois sujets d'entreprise (décisionnels) que vous pensez pouvoir modéliser à l'aide de cette description
  2. Choisissez un de ces trois sujets pour approfondir votre analyse et proposition :
    - a) A quels types de questions votre entrepôt permettra-t-il de répondre ?
    - b) Quel sera le fait de base, et sa granularité ?
    - c) Définir chacune des tables de dimensions.
    - d) Faire un modèle en étoile correspondant
    - e) Identifier les hiérarchies (si elles existent), et proposer le modèle en flocon correspondant

#### DATAMINING

8 points

Les données ont été recueillies dans le but de prédire si un client allait souscrire à un service bancaire ou pas. Les données ont donc un attribut *souscription* dont les valeurs possibles sont soit YES (95%), soit NO (5%). Mais ces données peuvent aussi être exploitées à d'autres tâches de fouille.

1. Décrire le processus de CRISPDM
2. En quoi consiste le prétraitement ?
3. Il existe des corrélations et des implications entre les attributs décrivant les clients. Donner le principe de l'algorithme APRIORI pour l'extraction des règles. Proposer aussi des solutions pour rendre cet algorithme plus efficace (en temps et en mémoire). Donner un exemple d'exploitation des résultats de cet algorithme.
4. La tâche suivante consiste à construire à partir des données recueillies un modèle permettant de prédire si le client va souscrire au service ou pas. Quelles sont les mesures à utiliser pour évaluer ce modèle ? donner les formules et les interprétations de ces formules.

## Partie II : Veille scientifique et Technologique

15. En quoi consiste la veille scientifique et technologique ? (0,5 pt)
16. Quel pourrait être pour le Cameroun, l'un des objectifs de cette veille dans le cadre de la riposte au Covid 19 ? (1 pt)
17. Quel rôle pourraient jouer les TIC dans la mise en oeuvre de cette veille ? (1 pt)
18. Quel projet de recherche directement lié à votre sujet de mémoire (que vous prendrez soin de rappeler) pourrait être lancé dans le cadre de cette riposte ? (2 pts)
19. Quelle est la publication récente que vous avez lue et qui pourrait être utile au Cameroun dans le cadre de cette riposte ? (1 pt)
20. Donner un secteur d'activité lié aux TIC qui a été boosté au Cameroun à cause de la pandémie du Covid 19 (1 pt)
21. Donner un secteur d'activité lié aux TIC qui s'est effondré au Cameroun à cause de la pandémie du Covid 19 (1 pt)
22. Donner quatre objectifs possibles (2 sur le plan individuel, 2 sur le plan collectif) que peut viser une campagne de vaccination (2 pts)
23. Quel est l'objectif prioritaire visé par les campagnes actuelles de vaccination anti COVID-19 en Europe ? (1 pt)
24. Quelle est l'efficacité du vaccin AstraZeneca ? (donner une source crédible) (1 pt)
25. Sachant qu'au Cameroun  $R_0$  est estimé à 3,5, la vaccination de TOUTE LA POPULATION CAMEROUNAISE par AstraZeneca permet-elle d'avoir l'immunité collective ? (1 pt)
26. Citez un bénéfice à court terme, de la vaccination immédiate par AstraZeneca ? (1 pt)
27. Donner un conseil pour la stratégie vaccinale anti COVID-19 de notre pays à moyen terme ? (1 pt)

## Partie III : Ethique

28. Rappeler la définition de l'éthique proposée dans le séminaire du Prof Kirchner (0,5 pt)
29. Citez un exemple très couramment évoqué dans les media, de problème éthique (0,5 pt)
30. Citez une question éthique liée au numérique (non précédemment mentionné) et qui est spécifique aux pays africains (1 pt)
31. Citez un exemple de problème éthique (non précédemment mentionné) qui est né avec le Covid 19 (1 pt)
32. Citez un exemple de problème éthique (non précédemment mentionné) ayant de très fortes implications économiques (1 pt)

## Partir IV : Le Droit

33. Comment peut-on définir le droit ? (1 pt)
34. Donnez deux concepts proches du droit en précisant ce qui les distingue du droit (1 pt)
35. Citez un exemple de problème de droit très ancien qui a été complètement modifié avec la révolution numérique (1 pt)
36. Citez un problème nouveau de droit (non précédemment mentionné), qui est né suite à la révolution numérique (1 pt)
37. Citez un problème nouveau de droit (non précédemment mentionné), qui est né suite à la pandémie de Covid 19 (1 pt)
38. Citez un problème de droit (non précédemment mentionné), qui est né de la combinaison Révolution numérique + pandémie de Covid 19 (1 pt)



Computer Science Department

Examen INF5099 Apprentissage Distribué, 11 Mai 2021

Docteur Thomas MESSI

Exercice 1 : Parallélisation de K-means avec Map Reduce (10 points).

1. Algorithme séquentiel
  - (a) Rappeler le principe de K-Means
  - (b) Rappeler le fonctionnement de ce algorithme à l'aide du pseudo-code
2. Version parallèle
  - (a) Donner une stratégie permettant de le paralléliser.
  - (b) Donner la version parallèle de l'algorithme.
  - (c) Rappeler la loi d'Amhdhal et dire quelle est la fraction parallélisable de votre code.
  - (d) Utiliser cette fraction pour donner le speedup maximal de votre implémentation.
  - (e) Réaliser une implémentation de l'algorithme en utilisant Map Reduce et posix thread.  
(On donnera juste l'essentiel : code des mappers, code des reducers).

Exercice 2 : Parallélisation d'un algorithme d'apprentissage artificiel. (10 points)

Considérez les algorithmes suivant : SVM, RNN et CNN

1. Faire un choix, puis donner une description de l'algorithme séquentiel (pseudo code). On suppose que vous utilisez la descente du gradient (batch, minibatch ou stochastique).
2. Dire comment paralléliser l'algorithme en parallélisant uniquement la descente du gradient.
3. Donner les grandes lignes d'une implémentation avec posix threads.
4. Après avoir donné une estimation de la fraction parallélisable, donnez une estimation du speedup maximal.

..... Bon Courage! .....

INF5099 : Apprentissage Distribué

Enseignants : Dr. Paulin Melatagia, Dr. Thomas Messi

Examen de Fin de Semestre, Durée : 3 heures

**Exercice 1: (PSVM : 5 points)**

1. Rappeler la formulation du problème d'apprentissage avec SVM.
2. Pourquoi est-il important de disposer d'une implémentation de SVM qui passe à l'échelle?
3. Pourquoi SMO est difficilement parallélisable comparé à IPM?
4. Rappeler le principe de la méthode IPM with Incomplete Cholesky Factorization.
5. Quels sont les gains de PSVM par rapport à SVM; expliquer comment ces gains sont obtenus.

**Exercice 2: (LambdaMart : 5 points)**

1. Rappeler la définition du problème de Ranking vu en cours.
2. Quelle est l'idée maîtresse de l'algorithme LambdaRank?
3. Décrire le fonctionnement des algorithmes MART.
4. Quelles sont les deux approches proposées par Svore et al. pour distribuer LambdaMart?
5. LambdaRank est mis en oeuvre pour un cluster de machines. Expliquer en quoi ce type de distribution du calcul est différent d'une implémentation sur une architecture massivement parallèle de type FPGA.

**Exercice 3: (MapReduce : 6 points)**

1. Rappeler le principe de Map Reduce tel que vu en cours
2. Dire dans quel cas Map Reduce peut-être vu comme un modèle et dans quel cas il peut-être vu comme une technologie.
3. Répondre aux questions suivantes sur l'algorithme K-means :
  - (a) Rappel la version séquentielle de l'algorithme.
  - (b) Donner une stratégie de parallélisation et justifier en quoi elle permettra de réduire le temps d'exécution.
  - (c) Donner les codes des mappers et des reducers pour une implémentation de l'algorithme en utilisant posix thread.
  - (d) Rappeler et expliquer la formule du calcul du temps d'exécution d'un programme parallèle d'après la loi d'Amdahl. Qu'en est-il de votre implémentation?

**Exercice 4: (DryadLink : 4 points)**

1. DryadLink est un modèle ou une technologie? Justifier votre réponse.
2. Prendre un algorithme d'apprentissage de votre choix et dites comment réaliser sa version parallèle avec DryadLink (sous forme de grandes lignes).
3. Le problème principal de DryadLink est son lien fort avec les technologies de Microsoft. On voudrait résoudre ce problème en réalisant une implémentation plus libre. Après avoir rappelé le fonctionnement de DryadLink, dites-nous comment réaliser un outil équivalent en se servant de Map Reduce.



## Faculté des Sciences Département d'Informatique

## Examen INF0505 Session Normale

Avril 2021

Professeur Maurice TCHUENTE, Docteur Thomas MESSI NGUELE

**Partie I – Compréhension d'un Modèle exposé lors d'un Séminaire**

Thème : Epidémiologie mathématique – Séminaire du Professeur Samuel Bowong

1. Que signifie l'acronyme COVID-19 (0,5 pt)
2. Quelle est la classe des maladies concernées par les modèles d'épidémiologie mathématique exposés par Professeur Samuel Bowong ? (0,5 pt)
3. Donnez une approche de modélisation épidémiologique autre que celle présentée par Prof Bowong (1 pt).
4. Expliquer comment est construit le modèle SIR : Hypothèses, variables S, I, R, paramètres  $\beta$  et  $\gamma$ , équations (2 pts)
5. Expliquer et donner la représentation graphique de l'une des variantes suivantes du modèle SIR : SIS, SIRS (1 pt)
6. Donner les paramètres et le schéma correspondant à l'un des modèles suivants : SIR avec confinement, SIR avec incubation (2 pts)
7. Dans chacun des trois cas ci-dessous, justifiez l'affirmation qui est proposée, si vous êtes consulté par le Gouvernement en tant qu'expert de modélisation épidémiologique, pour donner des conseils sur l'évolution de la pandémie de COVID 19 au Cameroun. A chaque fois, vous justifierez brièvement et précisément votre réponse.

**Cas 1 : A court terme (moins de 45 jours) (1 pt)**

A court terme un modèle épidémiologique peut être utilisé dans un but prédictif

**Cas 2 A moyen terme (de 45 jours à 3 mois) (1 pt)**

A moyen terme un modèle épidémiologique est à utiliser plus dans un but explicatif que dans un but prédictif

**Cas 3 A long terme (plus de 3 mois) (1 pt)**

A long terme un modèle épidémiologique est à utiliser essentiellement dans un but explicatif.

Dans toute la suite on considère le modèle SIR.

8. Qu'appelle-t-on taux de reproduction de base  $R_0$  (0,5 pt)
9. Qu'appelle-t-on taux de reproduction  $R$  (0,5 pt)
10. Qu'appelle-t-on immunité individuelle ? (0,5 pt)
11. Qu'appelle-t-on immunité collective ? (0,5 pt)

On considère un premier accès d'épidémie liée à une maladie infectieuse, dans une population totalement saine et dont le taux de reproduction de base est  $R_0$ . On suppose que cette épidémie se termine en laissant la population survivante dans un état où on a une proportion  $p$  de personnes guéries et immunisées, les autres survivants étant tous sains mais susceptibles d'attraper la maladie lors d'une seconde vague. On suppose que les deux paramètres  $\beta$  et  $\gamma$  restent inchangés.

12. Donner en fonction de  $p$  et  $R_0$ , le taux de reproduction de base  $R'_0$  de la seconde vague (1 pt)
13. Déduire en fonction de  $R_0$ , le seuil  $p_0$  que doit avoir  $p$  pour qu'on ait l'immunité collective (1 pt)
14. Remplir le tableau ci-dessous en fonction de votre réponse à la question précédente (1 pt)

$R_0$	2	2,5	2,75	3	3,25
$p_0$					69%

N.B. Comme documentation de base, vous pouvez vous servir aussi de l'exposé de Prof Gauthier Sallet.

REPUBLIQUE DU CAMEROUN

*Paix – Travail – Patrie*

UNIVERSITÉ DE YAOUNDÉ I

Faculté des Sciences

Département d'Informatique

B.P. 812 Yaoundé



REPUBLIC OF CAMEROON

*Peace – Work – Fatherland*

UNIVERSITY OF YAOUNDÉ I

Faculté des Sciences

Department of Computer Science

P.O.Box 812 Yaoundé

EXAMEN : S1

NIVEAU : M2

INTITULE : Vision par ordinateurs

SALLE : S111

DATE : 12 mai 2022

SEMESTRE : 01

CODE UE : INF5109

HORAIRE : 14H<sup>45</sup> → 16H<sup>45</sup>

### Exercice 1 (7pts)

1. Définir les expressions suivantes : (1pt)
  - a. Texture
  - b. False Accept Rate
2. Donnez la différence entre l'authentification et l'identification (1pt)
3. Donnez la différence entre un système biométrique permissif et un système biométrique de haute sécurité. L'argumentaire sera axé sur le FAR et le FRR. (2pt)
4. Énumérez dans l'ordre les phases d'un système biométrique. (1pt)
5. Définir segmentation et citer 2 techniques de segmentation. (2pt)

### Exercice 2 (6pts)

6. .... est utilisé pour la détection des coins dans une image (1pt)
  - a. Le détecteur de Harris-Stephens
  - b. L'algorithme SIFT
  - c. Le filtre de Canny
  - d. Le détecteur de Moravec
7. .... est utilisé comme approche de l'analyse de texture. (1pt)
  - e. Matrices de co-occurrences
  - f. Méthode globale
  - g. Statistiques du premier ordre
  - h. Méthode spatiale
8. On peut calculer plusieurs paramètres d'une texture à partir de l'image ou de son histogramme : (1pt)
  - i. Moyenne
  - j. Distance
  - k. Uniformité
  - l. Entropie
  - m. Ecart type
9. Le..... permet de détecter les contours dans une image. (1pt)
  - n. Filtre Gaussien
  - o. Filtre de Canny
  - p. Filtre médian
  - q. Filtre de Sobel
10. Le .... est un filtre numérique non linéaire, souvent utilisé pour la réduction de bruit. (1pt)
  - r. Filtre moyenneur
  - s. Filtre de Robert
  - t. Filtre médian
  - u. Laplacien
11. Le..... permet de flouter une image. (1pt)
  - v. Filtre Gaussien
  - w. Filtre de Canny
  - x. Filtre médian
  - y. Filtre de Prewitt

### Exercice 3 : (7pts)

1	4	4	3
4	2	3	2
1	2	1	4
1	2	2	3

Considérant l'image ci-dessus :

12. Donner la matrice de co-occurrence de cette image pour la distance 2 et direction  $45^\circ$  (2pts)
13. A partir de cette matrice calculer les paramètres suivants : (5pts)
- a. Moyenne
  - b. Energie