

Examen INF5119 - Techniques avancées de fouille de données et applications (Durée : 02h00)

Enseignant : NZEKON NZEKO'O Armel Jacques, PhD

Exercice 1 : Techniques de fouille de données (2 + 2 + 2 + 2 = 8 points)

- Décrire le principe de la descente du gradient.

Indices : Dérivées, optimisation.

- En quoi est-ce que l'application d'un filtre de convolution vue comme opération matricielle correspond à une opération effectuée au niveau d'un neurone ?

Indices : Ne pas hésiter à faire des schémas pour illustrer cette correspondance.

- Décrire le principe de l'algorithme LIME pour expliquer la conclusion d'un modèle pour un exemple donné ; LIME propose pour chaque attribut descriptif de l'exemple d'entrée, son taux de contribution à la conclusion du modèle.

Indices : Régression linéaire d'un nuage de points.

- Décrire le principe du Forward Selection pour la sélection de quelques attributs les plus pertinent et suffisant pour résoudre efficacement un problème de classification.

Indices : Gain d'information, arbre de décision, Accuracy.

Exercice 2 : Évaluation des modèles de fouille de données (2 + 3 + 1 = 6 points)

- Pour l'évaluation du résultat d'un algorithme de clustering, on peut calculer la silhouette

$$s(x) = \frac{b(x) - a(x)}{\max(b(x), a(x))}$$

— Dans quel intervalle $s(x)$ prend ses valeurs

— Donner les significations thématiques de $b(x)$ et $a(x)$

— Quand est-ce que $s(x)$ est maximale, minimale ou nulle et dire à quoi ça correspond ?

- Dessiner un exemple de matrice de confusion dans le cas d'une classification où on dispose de 03 classes, et dire comment est-ce qu'on procède pour le calcul des métriques ci-dessous.

— Dessiner une matrice de confusion générique pour une classification où on dispose de 03 classes (A, B, C) et dessiner également un exemple avec des valeurs dans les cellules de la matrice.

— Pour l'évaluation du modèle pour lequel la matrice de confusion est construite, donner les formules de calcul de l'Accuracy, la Précision et le Rappel, et calculer à chaque fois ces métriques pour l'exemple précédent.

- Pourquoi la métrique Précision est déconseillé lorsque le jeu de données est déséquilibré ?

Prendre un exemple théorique d'une classification avec 03 classes pour votre illustration.

Exercice 3 : Applications des techniques de fouilles de données (0.5 x 6 x 4 = 12 points)

Considérez 04 problèmes parmi les 09 problèmes ci-dessous :

- Système de recommandation 2 b
- Explicabilité des réseaux de neurones 1
- Sélection d'attributs 1
- Prédiction de la résistance de la compression du béton 1
- Prédiction de l'évolution du stock d'un produit 2
- Analyse de l'opinion sur l'aspect d'un objet 1
- Détection d'une maladie 7
- Extraction des patterns dans une image 1 i
- Clustering des spectrographes des molécules 1

- Décrire en quoi consiste chacun des problèmes en un paragraphe.
- Décrire en un paragraphe, un contexte applicatif réel dans lequel la solution à ce problème est importante.
- Donner un exemple de modèle de fouille de données vu en cours pour résoudre ce problème.
- Décrire le format d'entrée des données sur lesquels ce modèle peut être appliqué.
- Décrire une métrique d'évaluation pour évaluer la qualité du modèle proposer.
- Dire ce que cette métrique permet de mesurer en vous référant au domaine d'application.



Computer Science Department

Examen INF5099 Apprentissage Distribué, 08 Mars 2023
Dr MESSI NGUELÉ Thomas. Senior Lecturer

Exercice 1 : Analyse des performances (4.5 points). Soit T_p le temps d'exécution avec p thread.

1. Définir speedup puis donner la formule.
2. Soit s la fraction séquentielle d'un code donnée. On considère la formule suivante :
$$T_p = s * T_1 + (1 - s) * \frac{T_1}{p}$$
 - (a) Dire de façon littérale ce que traduit la loi d'Amdhal.
 - (b) Déterminer l'expression du speedup maximal à partir de cette formule.
 - (c) Déterminer le speedup max pour : $s = \frac{1}{2}$, $s = \frac{1}{4}$

Exercice 2 : Produit de matrices d'ordre n avec T Threads (4.5 points).

1. Donner une stratégie de parallélisation du produit matriciel.
2. Rappeler le rôle des fonctions `pthread_create`, `pthread_join`.
3. Écrire la fonction principale exécutée par un thread.
4. Analyser votre code puis déterminer la fraction séquentielle s .
5. Déterminer le speedup maximal.

Exercice 3 : Parallélisation de K-means avec Map Reduce (4.5 points).

1. Rappeler le principe de K-Means
2. Rappeler le fonctionnement de cet algorithme à l'aide du pseudo-code
3. Donner une stratégie permettant de le paralléliser.
4. Donner la version parallèle de l'algorithme.
5. 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 4 : Parallélisation du MLP. (5 points) Considérez l'algorithme MLP.

1. 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 parallisant uniquement la descente du gradient.
3. Donner les grandes lignes d'une implémentation avec posix threads.

Exercice Bonus : Parallélisation d'un algorithme d'apprentissage artificiel. (4 points)
Paralleliser l'un des algorithmes suivant qui vous a été donné pendant le projet CNN, RNN, GNN. Donner tous les éléments permettant de montrer que vous avez réalisé le projet par vous-même.

..... Bon Courage!



EXAMEN : S1

NIVEAU: M2

INTITULE : Vision par Ordinateur

SALLE : S107

DATE : 09/03/2023

SEMESTRE : 01

CODE UE : INF5109

HORAIRE: 14H⁴⁵-16H⁴⁵

EXERCISE I (5Pts)

1. Les filtres a et b permettent d'améliorer les bords (0,5Pts)
 - a. Prewitt
 - b. Sobel
 - c. Laplacien
 - d. Gaussien
2. La détection des bords dans une image est généralement accomplie en effectuant un(e) ----- spatial(e) du champ de l'image. (0,5Pts)
 - c. Filtre de lissage
 - f. Intégration
 - g. Différenciation *
 - h. Filtre Min
3. Quels sont les filtres et détecteurs qui permettent la détection des coins dans une image: (0,5Pts)
 - i. Le détecteur de Harris-Stephens
 - j. le filtre de Robert
 - k. le detecteur de Moravec
 - l. Filtre de Canny
4. L'un des filtres suivants est non linéaire (0,5Pts)
 - m. Filtre Gaussien
 - n. Filtre de moyenneur
 - o. Filtre Laplacien
 - p. Filtre Médian
5. On peut calculer plusieurs paramètres d'une texture à partir de l'image ou de son histogramme : (0,5Pts)
 - q. Moyenne
 - r. Distance
 - s. Uniformité
 - t. Entropie
 - u. Ecart type
- v. Le..... permet de détecter les contours dans une image. (0,5Pts)
 - w. Filtre Gaussien
 - x. Filtre de Canny
 - y. Filtre médian
 - z. Filtre de Sobel
6. Le G. est un filtre numérique non linéaire, souvent utilisé pour la réduction de bruit. (0,5Pts)
 - aa. Filtre moyenneur
 - bb. Filtre de Robert
 - cc. Filtre médian
 - dd. Laplacien
7. Le L.... permet de flouter une image. (0,5Pts)
 - ee. Filtre Gaussien
 - ff. Filtre de Canny
 - gg. Filtre médian
 - hh. Filtre de Prewitt

EXERCISE II (6Pts)

8. Quelle sont les différents types de codage et description des couleurs ? (1,5Pts)
9. Quels est la différence entre apprentissage supervisé et non supervisé et leur champ d'application ? (1,5Pts) ✓
10. Citer 2 méthodes d'extraction de caractéristiques ? (1,5Pts) ✓
11. Quelle différence entre la programmation traditionnelle et le machine learning? (1,5Pts) ✓
12. Définir segmentation et citer 2 techniques de segmentation. (1,5Pts) ✓
13. Que signifie l'apprentissage par transfert (1,5Pts) ✓
14. Quels sont les différents types d'apprentissage par transfert ? (1,5Pts)
15. Citer 8 algorithms de detection d'objet (1,5Pts) ✓

EXERCISE III (10Pts)

16. Citer deux algorithmes de machine learning utilisés pour faire de la classification d'images, en expliquant brièvement leur fonctionnement. (3pts)
17. Décrire l'architecture de réseau de neurones convolutifs (2Pts)
18. Expliquer le principe de l'algorithme AdaBoost (2Pts)
19. Comparer SIFT and SURF (3pts)

slow
lent faster comp
rapide tem de calcul

Examen sur la méthodologie de la recherche, Mars 2023

Répondez aux questions principales et aux questions sous-jacentes suivantes :

- 1. Qu'est-ce que la recherche en sciences sociales ?** 7
 - a. Qu'est-ce que la connaissance scientifique ?
 - b. Quels sont les fonctions et objectifs de la recherche scientifique ? ✓
 - c. Quelles sont les caractéristiques des connaissances scientifiques ?
 - d. Quelles sont les étapes de la recherche scientifique ? ✓
 - e. Quelle est l'objectivité et la subjectivité dans la recherche scientifique ? ✓
 - f. Quelle est l'éthique de la recherche scientifique ?
- 2. Quelle est la méthodologie de la recherche scientifique ?** 4
 - a. Comment déterminer la méthodologie de recherche ?
 - b. Quelle est la différence entre la recherche quantitative et la recherche qualitative ?
 - c. Qu'est ce que la triangulation ?
 - d. Qu'est ce que la démarche comparative ?
 - e. Quels sont les types d'échantillons ?
 - f. Quelles sont les étapes d'examen de l'échantillon ? ✓
 - g. Comment déterminer le nombre approprié d'échantillons ?
- 3. Qu'est-ce que la recherche qualitative ?** 5
 - a. Quelles sont les caractéristiques de la recherche qualitative ?
 - b. Quelles sont les stratégies de recherche qualitative ?
 - c. Quelles sont les méthodes de collecte d'informations dans la recherche qualitative ?
 - d. Comment rédige-t-on les questionnaires d'entretien ou d'enquête ?
 - e. Comment s'effectue l'analyse des données ?
 - f. Quelles sont les erreurs courantes dans la formulation des résultats de recherche qualitative ?
- 4. Qu'est-ce que la recherche quantitative ?** 6
 - a. Quelles sont les caractéristiques de la recherche quantitative ?
 - b. Quelles sont les hypothèses de recherche ?
 - c. Quels sont les avantages des hypothèses ?
 - d. L'hypothèse nulle et l'hypothèse alternative
 - e. Quelles sont les caractéristiques des hypothèses ?
 - f. Qu'est-ce qu'une variable quantitative et variable qualitative ?
- 5. Comment analyser les données quantitatives ?**
 - a. Quelles sont les conditions de la causalité entre les variables dans une recherche quantitative ?
- 6. Comment la recherche scientifique est-elle rédigée et publiée ?** 3
 - a. Quel est le produit final de la recherche ?
 - b. Méthode scientifique et méthodologique de formulation ? ✓
 - c. Quelles sont les règles à respecter de citations, d'attribution et de documentation ?
 - d. Quelle est la structure du rapport / de la recherche ?
- 7. Comment évalue-t-on la qualité d'un article ?** 1
 - a. Quelle est l'importance d'un résumé dans un article scientifique ?
 - b. Comment lire de manière efficiente et efficace un article scientifique ?
 - c. Comment apprécie-t-on un sujet d'un article ou d'une thèse ?
 - d. Quelles différences faites-vous d'un résumé long et d'un résumé court ?
 - e. Quelles sont les erreurs à ne pas commettre lors de l'écriture d'un résumé
- 8. Quelles sont les qualités inhérentes au métier de chercheur ?** 2
 - a. Quelles sont les qualités d'un chercheur ?
 - b. Quel est l'attitude d'un bon chercheur ?
 - c. Quelles sont les qualités d'une femme ou d'un homme scientifique ?
 - d. Qu'est-ce qu'un bon scientifique ?
 - e. Qu'est-ce qui caractérise le travail d'un chercheur ?
 - f. Comment le scientifique élabore-t-il un sujet de recherche ?

INF5029 : Science de données
Examen de fin de semestre 1 (2022-2023)
Durée : 2 heures

NB : Documents, téléphones, ordinateurs, calculatrices fermés

Questions

- 1) Définir : Fouille de données, indicateur, flocon, classification
- 2) Définir la règle des 5 V
- 3) Définir les propriétés ACID
- 4) Définir les opérations CRUD du paradigme clé/valeur

Exercice 1

Une agence de voyage aimerait pouvoir analyser ses données afin de planifier de meilleures campagnes de promotion auprès de ses clients. Plus particulièrement, elle aimerait analyser le nombre et le montant des ventes en fonction:

- ✓ De la destination: hôtel, ville, pays, région, catégorie de région (ex: bord de mer, alpine, etc.), catégorie de destination (ex: familial ou non), catégorie hôtel (ex: 14 étoiles);
 - ✓ De la date d'achat: jour de l'année, jour de la semaine, mois, année, saison touristique (ex: basse ou haute saison);
 - ✓ De la date de départ: jour de l'année, jour de la semaine, mois, année, saison touristique (ex: basse ou haute saison);
 - ✓ Du forfait: nombre de personnes, nombre de nuits, type de forfait (ex: tout inclus, repas inclus, etc.), type de chambre (ex: standard, suite, penthouse, etc.) ;
 - ✓ Du client: groupe d'âge, sexe, adresse, type d'acheteur (ex: nouveau, récurrent, etc.) ; Du canal de vente: catégorie (ex: magasin, internet, etc.) ;
 - ✓ De la promotion: catégorie (ex: 2 pour 1, rabais 10%, rabais 25%, etc.), début et fin de validité ;
 - ✓ Du mode de paiement: catégorie (ex: crédit, comptant, etc.) ;
1. Définir clairement les indicateurs à analyser pour ce problème
 2. Proposer un schéma en étoile permettant de faire ces analyses en faisant ressortir clairement les clés primaires et étrangères des tables de faits et de dimension;
 3. Identifier, pour chaque table de dimension, une hiérarchie de niveaux de granularité et proposer un schéma en flocon
 4. Proposer une stratégie d'agrégation pour mettre ensemble plusieurs étoiles ou flocons.

Exercice 2

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%).

1. Décrire chaque étape du processus de CRISPDM
2. En quoi consiste le prétraitement ? Décrire en quoi consiste les opérations de normalisation et de discréttisation ?
3. Les proportions de YES et NO dans le jeu de données posent elles un problème pour faire cette classification ? Si oui lequel ? Proposer une solution si oui.
4. Ces données vont servir pour construire un modèle permettant de classer ces clients ? Proposer trois mesures pour évaluer ce modèle. Donner une interprétation de chacune d'elle.
5. Décrire un problème de votre spécialité qui peut être résolue par une approche de fouille de données. Comment collecter les données ? Comment construire et évaluer la solution ?