le cnam

Sujet UE RCP216

Ingénierie de la fouille et de la visualisation des données massives

Année universitaire 2022–2023

Examen 1ère session : 20 juin 2023

Responsable : Michel CRUCIANU

Durée : 2h00

Seul document autorisé : 2 feuilles A4 recto-verso, manuscrites.

Les téléphones mobiles et autres équipements communicants (PC, tablette, etc.) doivent être éteints et rangés dans les sacs pendant toute la durée de l'épreuve. Calculatrice autorisée mais inutile.

Sujet de 5 pages, celle-ci comprise.

Vérifiez que vous disposez bien de la totalité des pages du sujet en début d'épreuve et signalez tout problème de reprographie le cas échéant.

---

1 Fouille de données (13 points)

[4 points] Question 1 : Classification automatique

1. Détailler l'implémentation MapReduce de l'algorithme *K-means*. (**2 points**)
2. Comment est fait le choix des candidats centres dans l'algorithme *K-means*$||$? (**2 points**)

Correction :

1. (voir le support de cours)
2. A chaque itération successive il y a env. l nouveaux candidats centres qui sont choisis par tirage aléatoire suivant une distribution non uniforme qui privilégie les données éloignées des candidats centres déjà choisis aux itérations précédentes. Les candidats centres, en nombre bien supérieur à k, sont ensuite pondérés par le nombre de données dont ils sont plus les proches et ensuite classifiés en k groupes. Voir aussi le support de cours.

[3 points] Question 2 : Filtre de Bloom

1. Indiquer pourquoi le filtre de Bloom permet d'éviter les faux négatifs. (**1 point**)
2. Indiquer pourquoi le filtre de Bloom ne permet pas d'éviter les faux positifs. (**1 point**)
3. Comment réduire le nombre de faux positifs pour le filtre de Bloom ? (**1 point**)

Correction :

1. Parce que la valeur d'interdiction (1) est insérée dès la construction dans toute case mémoire identifiée par le *hash* d'une donnée à filtrer.
2. Les fonctions de hachage présentent des collisions, une données anodine (à ne pas filtrer) peut entrer en collision avec une donnée à filtrer.
3. En utilisant plusieurs fonctions de hachage indépendantes : les collisions ne seront en général pas les mêmes entre les différentes fonctions, une donnée sera filtrée seulement si la valeur d'interdiction est présente dans les cases mémoires identifiées par les *hashs* obtenus suivant toutes les fonctions de hachage.

[3 points] Question 3 : Apprentissage statistique à large échelle

1. Si le nombre de paramètres à optimiser pour un modèle est p, comment pouvons-nous exprimer la complexité algorithmique de la descente de (sous-)gradient employée pour optimiser ces paramètres ? (**1,5 points**)

2. Quelle insuffisance voyez-vous pour la recherche en grille (*grid search*) des valeurs des hyperparamètres pour un modèle (SVM ou autre) ? (**1,5 points**)

Correction :

1. La descente de (sous-)gradient implique de calculer à chaque itération des dérivées partielles par rapport aux p paramètres, la complexité algorithmique (par itération) devrait donc être $O(p)$ (linéaire en p).
2. La recherche en grille se limite aux points de la grille, or l'optimum peut se trouver **entre** des points de la grille. Une recherche en grille **hiérarchique** permet de s'approcher plus de l'optimum (c'est également le cas pour différentes méthodes de recherche stochastiques).

[3 points] Question 4 : Non discrimination

Pour un système d'aide à la décision, on note par A l'attribut protégé, par R le score fourni par le modèle et par Y la variable cible.

1. Quelle est la difficulté majeure dans l'utilisation des critères observationnels ? (**2 points**)
2. Mentionner une difficulté particulière liée à l'utilisation du critère d'indépendance (ou de parité démographique, $R \perp\!\!\!\perp A$). (**1 point**)

Correction :

1. La principale difficulté est l'accès aux variables employées : l'attribut protégé (A) peut ne pas être accessible en interne, le score (R) fourni par le modèle n'est en général pas accessible en externe et la variable cible (Y) est en général disponible seulement pour les cas où $R = 1$ (ou R suffisamment élevé si c'est une variable continue).
2. Si Y n'est pas indépendante de A (malheureusement le cas typique) alors le prédicteur parfait qui donnerait $R = Y$ ne peut pas satisfaire $R \perp\!\!\!\perp A$.

2 Fouille de graphes et visualisation de données (7 points)

[2,5 points] Question 5 : Expliquez ce qu'on appelle une « loi de puissance ». Où est-ce que ça peut intervenir dans les contextes de l'analyse de graphes et réseaux sociaux ? Quelles conséquences cela a-t-il, pour le graphe, et pour le travail de l'analyste ?

[2 points] Question 6 : Expliquez ce que mesure l'indicateur appelé modularité. Pourquoi et comment peut-on chercher à le maximiser ?

[2,5 points] Question 7 : Visualisation de données.

La figure 1 présente une visualisation du résultat d'un sondage sur les baby-boomers (personnes nées entre 1946 et 1964, âgées en 2023 de 59 à 77 ans). En utilisant les connaissances acquises en cours de DataViz, expliquez l'organisation de cette infographie, détaillez ses travers et donnez des pistes d'amélioration.

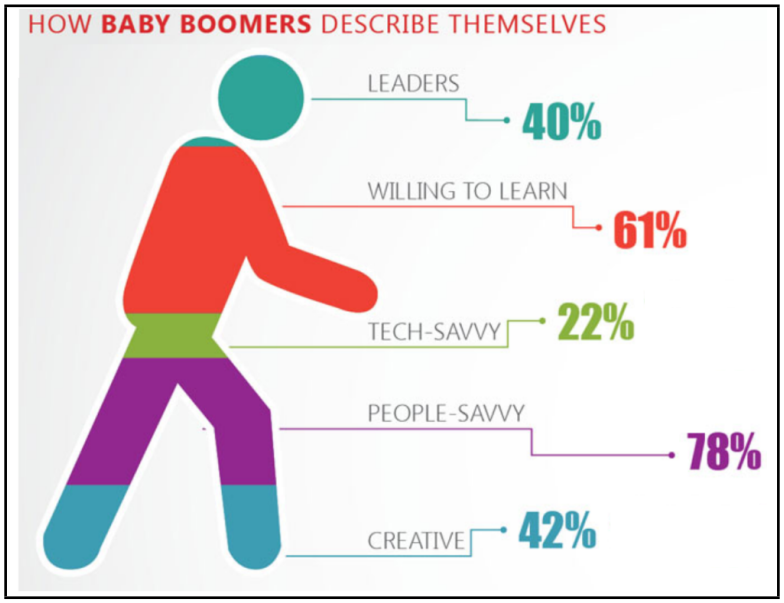

FIGURE 1 – "How Baby Boomers describe themselves ?" (Comment les baby boomers se descrivent-ils ?). source : Elucidat.

La figure présente une silhouette humaine, simplifiée et colorée. De la tête au cou, elle est bleu turquoise, avec un trait surmonté de "leaders" (décideur) et 40%. Du cou aux hanches, elle est orange, et avec un trait "willing to learn" (ayant une soif d'apprendre) et 61%. La zone des hanches est vert pomme, et avec "tech-savvy" (féru de technologie) et 22%. Les jambes sont violettes, avec "people-savvy" (sens des relations humaines) et 78%. Enfin, du milieu des mollets au bas des pieds, la silhouette est bleu clair, avec 42% pour creative ("créatif/ve").

Correction :

1. Une loi de puissance (*power law*) est une répartition statistique que l'on retrouve dans de nombreux contextes, et en particulier pour des graphes de terrains (*complex networks*). Elle se traduit par une proportionnalité entre deux quantités, avec une puissance pour l'une des deux. $f(x) = ax^k$. Les réseaux sociaux en ligne (tels que Twitter) ont une distribution de degré en loi de puissance : il y a quelques dizaines de personnes qui ont "des millions d'abonnés" et des millions de personnes qui n'ont qu'une dizaine d'abonnés (et l'on peut tracer des courbes pour montrer que ça marche pour la plupart des points entre ces deux extrêmes, éventuellement avec des pentes légèrement différentes). Les propriétés d'une loi de puissance rendent la moyenne et la variance définie pour certaines valeurs de l'exposant (et pas pour d'autres) : parler de la moyenne d'une "distribution en loi de puissance" n'a généralement pas de sens.
2. La modularité compare un partitionnement à de l'aléatoire, favorisant les groupes denses et donnant un "score de qualité" indiquant qu'on a "bien trouvé" les communautés du graphe
3. la figure mobilise les variables graphiques de position et de couleur pour présenter les résultats du sondage. L'idée a été d'illustrer, par la silhouette, que les boomers sont décrits par diverses caractéristiques. Cependant :
   - la somme des pourcentages ne fait pas 100%, alors que l'usage d'une silhouette dont diverses parties sont colorées renvoie "aux parties d'un tout". C'est donc trompeur.
   - les couleurs n'ont aucune signification (gradation), ce qui est dommage. Qui plus est, les zones de couleur ne sont pas proportionnelles aux chiffres qu'elles représentent.

   Ainsi, on pourrait proposer une visualisation avec un histogramme *stacked*, par exemple, avec 2 modalités à chaque fois (curieux/pas curieux, créative/pas créative). On pourrait ensuite réfléchir pour croiser les catégories (et ainsi représenter la fraction de ceux qui sont "tech-savvy" ET "créatif"), mais il faudrait plus de données que ce que nous avons sur l'image.