le cnam

Sujet UE RCP216

Ingénierie de la fouille et de la visualisation des données massives

Année universitaire 2021–2022

Examen de rattrapage : 30 août 2022

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)

[3 pts] Question 1 : Passage à l'échelle

1. Quel taux maximal d'accélération des calculs peut être espéré en utilisant un *cluster* composé de n nœuds de calcul par rapport à un seul nœud de calcul ? (**1 point**)
2. Pouvons-nous espérer une augmentation de la rapidité des calculs lorsqu'on remplace des RDD par des *DataFrames* capables de conserver les mêmes données sous une forme beaucoup plus compacte ? Justifier brièvement. (**2 points**)

Correction :

1. Nous pouvons espérer une accélération maximale de n fois ; qui ne sera vraisemblablement pas atteinte pour plusieurs raisons : les échanges de certaines données entre nœuds de calcul introduisent des ralentissements, l'équilibrage de la charge entre nœuds est imparfait, etc.
2. Oui, réduire la place mémoire nécessaire pour stocker les données permet d'accélérer les calculs pour plusieurs raisons : réduction du temps nécessaire pour transférer les données entre stockage de masse et mémoire vive (voir la figure sur la hiérarchie de stockage), réduction du ralentissement introduit par les échanges de données entre nœuds de calcul, possibilité de conserver plus de structures de données en mémoire vive, etc.

[4 pts] Question 2 : Classification automatique

1. Expliquer en quoi diffère k-means$\|$ de k-means++. (**2 points**)
2. Expliquer la signification de la borne $O(\log \psi)$ sur le nombre d'itérations à faire pour la génération de centres dans l'algorithme d'initialisation k-means$\|$. Évolue-t-elle à chaque itération ? (**2 points**)

Correction :

1. K-means++ génère un seul nouveau centre (d'initialisation de k-means) à chaque itération suivant une loi non uniforme (qui privilégie les points éloignés des centres déjà choisis antérieurement), alors que k-means$\|$ génère de l'ordre de l nouveaux centres à chaque itération suivant cette même loi non uniforme. Cela permet de dépasser en peu d'itérations le nombre de centres nécessaires.
2. La borne $O(\log \psi)$ indique que plus les centres déjà générés sont bien distribués dans les données (valeur faible de ψ), plus le nombre d'itérations qui restent à faire sera faible car il n'est pas nécessaire de générer des centres en excès pour qu'ils soient suffisamment bien répartis. Cette borne évolue à chaque itération, avec l'évolution de ψ.

[3 pts] Question 3 : Systèmes de recommandation

1. Dans le filtrage collaboratif basé sur la factorisation régularisée, la matrice utilisateurs-articles est approchée par le produit $\mathbf{X} \simeq \mathbf{U} \cdot \mathbf{A}^t$. Chaque colonne de $\mathbf{U}$ (et chaque ligne correspondante de $\mathbf{A}^t$) est associée à un facteur latent. Comment peut-on interpréter **chacun des éléments** d'une ligne de $\mathbf{U}$? Et **les éléments** d'une colonne de $\mathbf{A}^t$? (**2 points**)
2. Comment procéder pour régler la constante de régularisation λ pour la factorisation régularisée ? (**1 point**)

Correction :

1. Chaque ligne de $\mathbf{U}$ caractérise un utilisateur et chaque colonne de $\mathbf{U}$ correspond à un facteur latent. L'élément ligne i colonne j indique dans quelle mesure le facteur latent j est important pour l'utilisateur i. Chaque colonne de $\mathbf{A}^t$ caractérise un article et chaque ligne de $\mathbf{A}^t$ correspond à un facteur latent. L'élément ligne j colonne k indique dans quelle mesure l'article k possède le facteur j. En conséquence, en faisant le produit de la ligne i de $\mathbf{U}$ avec la colonne k de $\mathbf{A}^t$ on cumule, sur tous les facteurs latents, les produits entre l'importance du facteur latent pour cet utilisateur et la mesure dans laquelle l'article possède ce facteur latent.
2. La technique classique de l'échantillon-test peut être employée : on met de côté un échantillon-test des données disponibles, sur les données restantes on estime différents modèles (matrices $\mathbf{U}$ et $\mathbf{A}^t$) pour différentes valeurs de λ, on compare leurs performances sur l'échantillon-test et on choisit la valeur de λ qui permet d'obtenir les meilleures performances.

[3 pts] Question 4 : Absence de discrimination

Pour un système d'aide à la décision, on note par A l'attribut protégé, par R le score sur la base duquel une décision est prise et par Y la variable cible. On souhaite appliquer un des trois critères observationnels pour détecter la discrimination.

1. Quelle difficulté se manifeste dans l'utilisation du critère d'indépendance ? (**1 point**)
2. Quelle difficulté se manifeste dans l'utilisation du critère de séparation ou du critère de suffisance ? (**2 points**)

Correction :

1. Dans l'utilisation du critère d'indépendance ($R \perp\!\!\!\perp A$) la difficulté majeure est le fait qu'en général la variable cible n'est pas indépendante de l'attribut protégé, il est donc difficile d'envisager une indépendance du score de l'attribut protégé.
2. Dans l'utilisation du critère de séparation ($R \perp\!\!\!\perp A|Y$) ou du critère de suffisance ($Y \perp\!\!\!\perp A|R$) la difficulté majeure est que souvent nous avons accès à la valeur de Y seulement pour les individus pour lesquels R a une certaine valeur (« acceptation ») ou sa valeur est supérieure à un seuil.

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

[1.5 pts] Question 5 : Qu'est-ce qu'une composante connexe dans un graphe ? Complétez votre explication rédigée par un schéma simple.

Correction : Une composante connexe dans un graphe est un ensemble de nœuds tels qu'il existe, pour chaque paire de nœuds de l'ensemble, un chemin entre eux.

[1 pt] Question 6 : Qu'est-ce que la centralité d'intermédiarité ? Que permet-elle de quantifier ?

Correction : La centralité d'intermédiarité est calculée sur les plus courts chemins dans un graphe. Pour un nœud/un lien, on compte le nombre de plus courts chemins du graphe qui passent par ce sommet/lien.

Selon les réseaux auxquels on l'applique, elle permet de caractériser l'influence/le pouvoir d'un nœud dans un réseau. En télécommunications par exemple, un nœud avec une forte centralité voit passer davantage d'information (et peut éventuellement la contrôler).

[1 pts] Question 7 : Que signifie, pour un graphe "avoir une distribution de degrés en loi de puissance" ? Quelles précautions statistiques doit-on observer, dans ce cas ?

Correction : 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. $f(x) = ax^k$.

La distribution des degrés d'un graphe indique le nombre de sommets qui ont pour degré une certaine valeur. 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 pts] Question 8 : Visualisation de graphes
Expliquez le principe algorithmique des méthodes de visualisation de graphes dites à "masses-forces-ressorts". Quels sont leurs avantages ? Et leurs défauts ?

Correction :

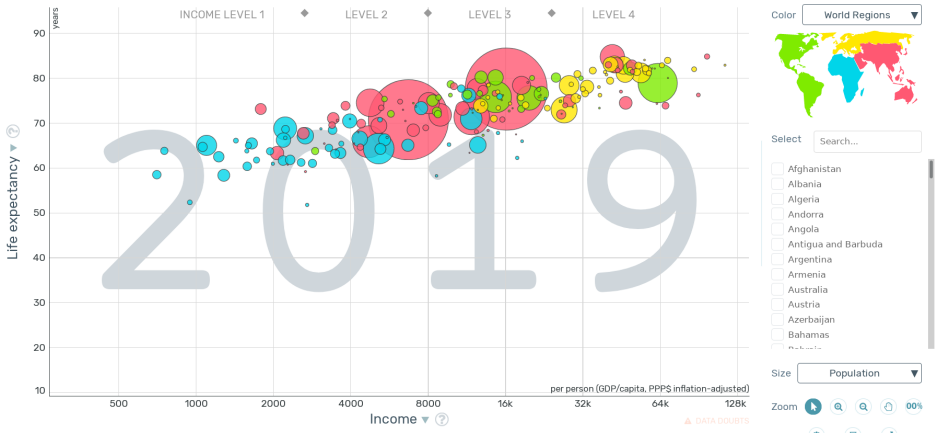
Ces méthodes sont issues de l'algorithme barycentrique de Tutte, raffiné par une vision physicienne du problème, reposant sur l'énergie d'un diagramme. Des masses, figurant les nœuds du graphes, sont attachées à des ressorts (les liens du graphes). On calcule les forces s'exerçant sur chaque masse : celles des ressorts et d'éventuelles forces complé-

mentaires (magnétiques, pour aligner ou espacer les nœuds). On garde la spatialisation obtenue du graphe quand le système atteint un équilibre.

Défauts : esthétiquement, cela peut générer des chevauchements. Et les graphes planaires ne sont pas forcément représentés sous forme "planaire".

Avantage : vitesse d'exécution, taille des graphes traités, facilité à implémenter et expliquer.

[1.5 pts] Question 9 : Quelles sont les variables graphiques qui ont été mobilisées pour créer cette visualisation ? Expliquez l'emploi de chacune.



Correction : Variables : Taille, position, couleur, forme, transparence.