le cnam

Sujet UE RCP216

Ingénierie de la fouille et de la visualisation des données massives

Année universitaire 2024–2025

Examen 1ère session : janvier 2025

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 6 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)

[La réponse à chaque question doit comporter au moins 2 lignes de texte manuscrit.]

[4 points] Question 1 : Réduction du volume de données, réduction de la complexité

1. Nous souhaitons réduire fortement à la fois le nombre d'observations, par échantillonnage, et le nombre de variables, par réduction de dimension avec l'analyse en composantes principales (ACP). Laquelle de ces réductions appliquer en premier et pourquoi ? (**2 points**)
2. Lorsque les observations disponibles font partie de classes très déséquilibrées (c'est à dire d'effectifs très différents), quelle méthode d'échantillonnage doit être employée et pourquoi ? (**1 point**)
3. Qu'est-ce qui permet de réduire la complexité sans réduire le volume de données ? (**1 point**)

Correction :

1. Si l'ACP est appliquée avant échantillonnage, la détermination des axes principaux est fiable (calcul à partir de toutes les données) mais le coût de l'ACP est plus élevé (application sur toutes les données). Si l'échantillonnage est appliqué avant l'ACP, le coût de l'ACP diminue mais la détermination des axes principaux est moins fiable car les calculs sont faits sur un petit échantillon.
2. Lorsque les classes sont très déséquilibrées l'échantillonnage stratifié est préférable car il permet de s'approcher plus du taux de sélection visé, surtout pour les classes minoritaires.
3. La complexité peut être réduite sans réduire le volume de données car en général les calculs (pour l'estimation de modèles comme pour la décision avec un modèle) privilégient les données proches et des méthodes d'indexation multidimensionnelle permettent de sélectionner à faible coût les données proches.

[3 points] Question 2 : Fouille de données textuelles

1. Les n-grammes de mots sont des séquences de n mots consécutifs. Un modèle vectoriel de texte de type « sac de mots » (avec pondérations TF-IDF, par exemple) peut être employé sur les n-grammes plutôt que sur les mots individuels : dans le vecteur qui représente un texte, chaque composante indique la présence ou non d'un n-gramme particulier dans ce texte. Quel pourrait être l'intérêt de l'emploi de n-grammes par rapport à l'emploi de mots individuels ? (**2 points**)

2. Pourquoi les mots les plus rares (par exemple les mots qui apparaissent moins de 5 fois dans la base de données textuelles) sont en général éliminés des représentations par « sac de mots » ? (**1 point**)

Correction :

1. Dans les représentations par « sac de mots » les relations de proximité entre mots dans le texte (par ex. mots successifs) sont ignorées, alors qu'elles sont importantes pour la signification de ces mots (par ex. pour traiter la négation, les locutions nominales ou verbales, etc.). Les n-grammes de mots représentent, dans une certaine mesure, ces relations de proximité entre mots dans le texte.
2. Les mots très rares ont des valeurs excessives pour leur pondération IDF et donc un impact fort sur les comparaisons de vecteurs TF-IDF, alors que leur très faible présence indique un intérêt très limité dans l'analyse.

[3 points] Question 3 : Apprentissage à large échelle

1. Pouvons-nous réduire en même temps le taux de faux positifs et le taux de faux négatifs ? Justifier brièvement. (**1 point**)
2. Indiquez les étapes de la procédure employée pour choisir les valeurs des hyperparamètres d'un modèle. (**2 points**)

Correction :

1. Comme on peut le voir Fig.77 du cours, lorsque la frontière définie par le modèle est loin de la région de séparation entre classes il est possible de réduire un de ces taux sans augmenter l'autre, en revanche à proximité de la région de séparation entre classes la diminution d'un des taux fait augmenter l'autre (sauf si les classes sont linéairement séparables, avec une marge). Pour continuer à réduire en même temps les deux taux il faut changer de classe de modèles, par ex. passer de modèles linéaires à des modèles non linéaires (voir par ex. la Fig.80).
2. Voir la section « Grid search pour le choix des hyperparamètres » du cours.

[3 points] Question 4 : Aspects éthiques dans la fouille de données

1. En quoi consiste le critère de séparation (égalisation des chances ou *equalized odds*) ? (**1 point**)
2. Quel est le principe de la méthode de post-traitement **vue en TP** permettant d'obtenir un modèle respectant le critère de séparation à partir d'un modèle initial qui ne respecte pas ce critère ? (**2 points**)

Correction :

1. Le score et les attributs protégés sont indépendants *conditionnellement* à la variable cible.
2. Pour le modèle initial les taux de faux positifs et de vrais positifs sont différents pour les différentes valeurs de l'attribut protégé (c'est en cela que le critère de séparation n'est pas respecté). Pour corriger le modèle afin de satisfaire à ce critère il faut augmenter le taux de prédictions positives (et ainsi à la fois le taux de vrais positifs et le taux de faux positifs) lorsque l'attribut protégé prend la valeur « défavorisée » et/ou réduire le taux de prédictions positives lorsque l'attribut protégé prend la valeur « favorisée ».

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

[2 points] Question 5 : Centralités

Pour le graphe suivant indiquez, sans calcul mais en justifiant votre réponse :

1. Quel nœud a la centralité d'intermédiarité la plus élevée ?
2. Quel nœud a la centralité PageRank la plus élevée ?

Correction :

1. Le nœud 4 a probablement la centralité d'intermédiarité la plus élevée car il est intermédiaire entre 3 sous-graphes plus fortement connectés.
2. Le nœud 5 a probablement la centralité PageRank la plus élevée car c'est vers lui que convergent le plus grand nombre de chemins.

[2 points] Question 6 : Visualisation d'information

Quels défauts présente la visualisation de la Fig. 2 et comment l'améliorer ?

Correction : D'abord, les différentes couleurs n'ont ici aucune signification, elles servent seulement à séparer des pays voisins. Aussi, pour certaines valeurs numériques les caractères sont petits et peu lisibles. Enfin, dans certains cas les nombres sont penchés sans raison. Il est préférable de supprimer la couleur de l'océan et des mers, représenter les frontières par des traits et employer des niveaux d'une même couleur pour représenter les ordres de grandeur des valeurs numériques.

[3 points] Question 7 : Visualisation de graphes

1. Quels sont les principaux avantages et défauts de l'algorithme de Tutte ? (**1 point**)

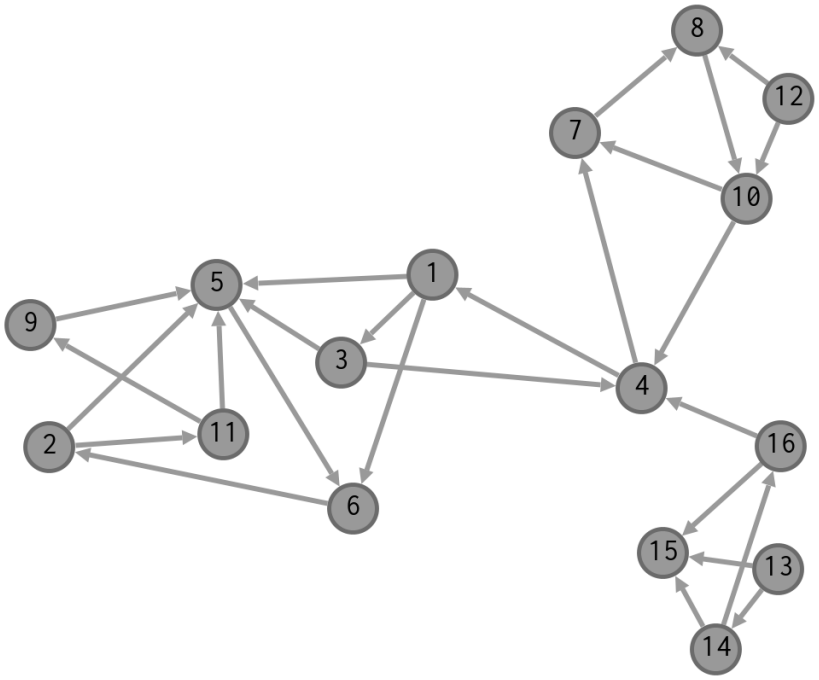

FIGURE 1 – Un graphe

2. Quel est le principe des méthodes de visualisation de graphes dites à « forces/ressorts » ? (**2 points**)

Correction :

1. Avantages : complexité relativement faible, représentations planaires pour les graphes planaires. Défauts : résolution sommitale limitée, parfois résolution angulaire limitée.
2. Le principe consiste à représenter le graphe par un système dynamique où chaque arête est un ressort et à ajouter différentes forces qui s'exercent sur les nœud pour contraindre la visualisation, ainsi que pour y inclure des contraintes particulières du domaine.

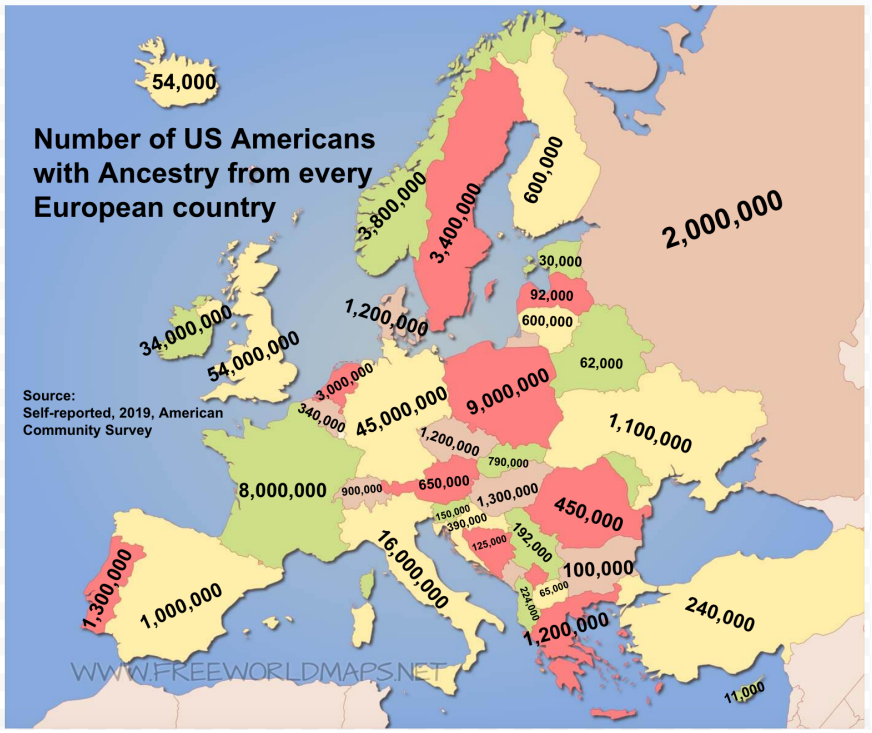

FIGURE 2 – Une visualisation (image en couleurs)