INF8111 — Recueil d'exercices

Quentin Fournier, Daniel Aloise

25 mai 2023

Préparation des données

1. Quelles affirmations sont vraies ?
   a) Les données académiques sont inutilisables, car elles ont déjà été traitées.
   b) La discrétisation perd de l'information.
   c) Les métadonnées sont une source importante de données.
   d) J'ai toujours le droit de gratter le web (web scrapping), car j'ai accès aux données.
   e) La standardisation ou z-score est la meilleure façon de normaliser les données.
   f) Il est possible de convertir n'importe quel type de données en graphe.
   g) Les séries temporelles ont généralement des dépendances explicites.
   h) Les artefacts sont des erreurs.
   i) Les données dépendantes sont généralement plus complexes que les données indépendantes.
   j) Mettre les valeurs manquantes à zéro est généralement une mauvaise idée.
2. Soit le jeu de données suivant sur lequel nous souhaitons faire une régression linéaire afin de prédire le montant des prêts futurs. Selon vous, quelles sont les 5 à 7 étapes du prétraitement des données nécessaires pour obtenir un meilleur modèle ? Justifier chaque étape en une ligne.

Âge	Diplôme	Salaire ($)	Date du prêt	Montant du prêt ($)
21	Bac	34000	08/2020	10000
64	Licence	67000	01/98	20000
35		43000	12/2003	37000
34	Bac	37000	7/2013	5000
42	Phd	98000	11/2009	30000
19	Bac	63000	01/2020	5000
26	PHD	113000	04/2016	14000

3. a) Découvrez ce qui est étrange à propos du mois de septembre 1752. Quelles mesures prendriez-vous pour normaliser des statistiques concernant ce mois-ci ?
   b) Dans le cas où aucune mesure spéciale n'est prise, seriez-vous devant une erreur ou un artefact ? Justifiez votre réponse.
4. Considérez la pluviométrie et la température journalière des villes de Montréal et de Québec. Laquelle de ces deux quantités, pluviométrie ou température, devrait être la plus corrélée dans le temps entre ces deux villes ? Pourquoi ?

Réduction et transformation de données

5. Considérez le jeu de données ci-dessous avec 8 individus, 2 attributs (X^1 et X^2), et une classe (y). Quel attribut a le plus grand pouvoir discriminant selon le score de Fisher ? Donnez vos résultats avec 3 chiffres après la virgule.

X^1	X^2	y
2	1	1
4	3	1
4	2	1
6	5	1
3	5	1
1	3	2
3	6	2
2	5	2

6. Imaginez une entreprise qui souhaite lancer un nouveau produit ou repositionner un produit existant sur le marché. Elle peut utiliser ?? pour analyser la façon dont les consommateurs perçoivent les différentes marques ou produits en termes de similarités ou de dissimilarités. En collectant des données à travers des enquêtes ou des questionnaires où les consommateurs évaluent ou comparent divers caractéristiques des produits, ?? peut être utilisée pour créer une carte perceptuelle. Cette carte représente les dimensions ou les facteurs sous-jacents que les consommateurs utilisent pour évaluer et différencier les produits. Chaque produit est ensuite positionné sur la carte en fonction de sa similarité ou de sa dissimilarité perçue par rapport aux autres.

   Quelle méthode vue en classe mieux corresponde à la méthode ??

7. Vous disposez d'un ensemble de données contenant 10 attributs.

   a) Expliquez le concept de réduction de dimension dans le contexte de l'Analyse de Composantes Principales (ACP).

   b) Quel est l'objectif de la standardisation des attributs avant d'appliquer l'ACP ?

   c) Après avoir effectué l'ACP, vous obtenez les valeurs propres des composantes principales. Comment pouvez-vous interpréter ces valeurs propres ?

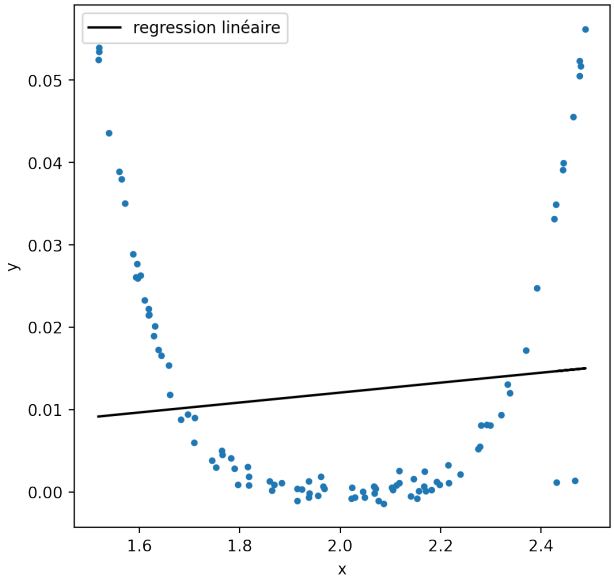
   d) Quelle est la signification du taux de variance expliquée en ACP ? Comment est-il calculé ?

   e) Supposons que les trois premières composantes principales expliquent 80% de la variance totale dans l'ensemble de données. Comment interpréteriez-vous ce résultat ?

Regression linéaire

8. Les coefficients d'une régression linéaire reflètent-ils à l'importance des attributs qu'ils multiplient ?

9. Soit un jeu de données contenant 100 individus (points) ayant un seul attribut (axe des abscisses x) et une seule valeur de sortie numérique (axe des ordonnées y). Une régression linéaire sans biais a été appliquée sur les données, cependant les résultats ne sont pas bons (voir figure). Comment améliorer ce modèle ?



10. Supposons que nous voulons trouver la meilleure fonction d'ajustement $y = f(x)$ où $y = w^2x + wx$.Comment utiliseriez-vous la régression linéaire pour trouver la meilleure valeur de w ?
11. Écrivez la formulation d'optimisation pour la régression linéaire de la forme $y_i = w^T X_i + b$ avec un terme de biais b. Fournit une solution pour les valeurs optimales de w et b en termes de matrice de donnéees X et du vecteur y. Montrer que la valeur optimale du terme de biais b est toujours égale à 0 lorsque la matrice de données X et le vecteur y sont tous les deux centrées sur la moyenne.

Classification

12. Pourquoi la régression logistique est un classificateur linéaire ?
13. En quoi consiste le *kernel trick* ? Quel est son principal avantage ?
14. On doit construire un arbre de décision afin de déterminer si on entre dans un restaurant, selon le type de cuisine et le prix. Voici les données :

Cuisine	Prix	Entrer
chinoise	$	oui
française	$$$	non
brésilienne	$	non
italienne	$$$$	oui
portugaise	$$	non
chinoise	$$	oui
française	$$$	non
brésilienne	$$	oui
italienne	$	non
portugaise	$$$$	non

Dans un arbre de décision par entropie, quel sera le premier attribut testé ?

15. Nous souhaitons déterminer à l'aide d'un arbre de décision si l'on va à la plage en fonction de la météo (soleil et vent). Considérer les données suivantes :

Soleil	Vent	Plage
Oui	Léger	Oui
Non	Moyen	Oui
Non	Léger	Oui
Oui	Léger	Oui
Non	Léger	Oui
Oui	Fort	Non
Non	Moyen	Non
Oui	Moyen	Non
Oui	Fort	Non
Non	Fort	Non

a) Quelle est l'entropie initiale ?

b) Quel est le gain de l'attribut vent ?

c) Quel attribut choisir pour le premier test ?

16. Le tableau suivant résume un jeu de données où chaque enregistrement possède trois attributs binaires (A, B et C) et appartient à la classe positive ou négative ($+$ ou $-$).

			# d'enregistrements	
A	B	C	$+$	$-$
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	2

Par exemple, la première ligne du tableau signifie qu'il y a 5 enregistrements du jeu de données pour lesquels $(A, B, C) = (T, T, T)$, et qui appartiennent à la classe $+$. De même, la quatrième ligne signifie qu'il existe 5 enregistrements de la classe $-$ pour lesquels $(A, B, C) = (F, F, T)$.

a) Construisez un arbre de décision de deux niveaux avec la méthode de maximisation du gain d'entropie vue en cours. Appuyez votre réponse avec vos calculs.

b) Combien d'enregistrements sont mal classés par l'arbre de décision obtenu en (a) ?

c) Êtes-vous capable de présenter un meilleur arbre que celui obtenu en (a) en termes du nombre d'enregistrements mal classés ?

17. Soit un jeu de données **linéairement séparables**. Donnez deux avantages et un inconvénient des machines à support de vecteurs (SVM) linéaires simples par rapport à la régression logistique. Justifiez vos choix en une ligne.

18. Quels sont les principaux avantages de l'apprentissage profond par rapport aux autres méthodes de l'apprentissage automatique ?

Évaluation de modèles

19. Expliquer avec vos propres mots :

a) Qu'est-ce que le surapprentissage (overfitting) ?

b) Comment le détecter ?

c) Comment le limiter ?

20. Expliquez pourquoi il y a un compromis entre la précision (*precision*) et le rappel (*recall*). Comment peut-on augmenter la précision ?

21. Soit un échantillon de N objets, dont p appartiennent à la classe positive. Soit un classifieur "gourmand" qui classifie tous les objets comme appartenant à la classe majoritaire. Quelle est la précision (accuracy), la précision de la classe positive (precision), et le rappel (recall) de ce modèle ? Donnez explicitement les calculs ou justifiez vos réponses.

22. Vous êtes invité à évaluer les performances de deux modèles de classification, M_1 et M_2. L'ensemble de test que vous avez choisis contient 10 individus avec 26 attributs binaires nommés de A à Z . Le tableau ci-dessous montre les probabilités postérieures obtenues en appliquant les modèles à l'ensemble de test. Seules les probabilités postérieures pour la classe positive sont reportées, parce que pour un problème à deux classes, $P(-) = 1 - P(+)$ et $P(-|A, ..., Z) = 1 - P(+|A, ..., Z)$.

Instance	True Class	$P(+\|A, .., Z, M_1)$	$P(+\|A, .., Z, M_2)$
1	+	0.73	0.61
2	+	0.69	0.03
3	–	0.44	0.68
4	–	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09
7	–	0.08	0.38
8	–	0.15	0.05
9	+	0.45	0.01
10	–	0.35	0.04

a) Tracez sur le même graphique les courbes ROC pour les deux modèles M_1 et M_2. À votre avis, quel est le meilleur modèle ? Justifiez.

b) Pour le modèle M_1, supposons que le seuil choisi soit $t = 0.5$. En d'autres termes, toute instance de test dont la probabilité postérieure est supérieure à t sera classée comme un exemple positif. Calculez la précision, le rappel et le F-score pour le modèle M_1 à cette valeur de seuil.

c) Pour le modèle M_2, supposons que le seuil choisi soit aussi $t = 0.5$. Calculez la précision, le rappel et le F-score pour le modèle M_2 à cette valeur de seuil. Comparez le F-score obtenu pour M_2 avec celui obtenu en (b) pour M_1. Lequel des deux modèles est supposé d'être le meilleur en fonction des F-scores calculés ? Les résultats sont-ils cohérents avec les courbes ROC tracée en (a) ?

23. Considérez un modèle de régression logistique qui classifie des courriels en deux classes : spam (classe positive) et non-spam (classe négative).Un courriel est classé comme spam lorsque la valeur retournée par la fonction sigmoïde est supérieure à une valeur de seuil (τ) fixée à 0.5. Dénotez z le rappel (recall) de ce modèle. Supposez maintenant que vous utilisez une autre valeur de seuil ($\tau' > \tau$) pour le même modèle. Est-ce que le rappel z augmente, diminue ou demeure constant ? Justifiez votre réponse.

Clustering

24. Utilisez k-means et la distance euclidienne pour regrouper en 2 clusters les 7 points : $x_1 = (2, 5)$, $x_2 = (5, 5)$, $x_3 = (6, 1)$, $x_4 = (7, 2)$, $x_5 = (8, 1)$, $x_6 = (4, 4)$, et $x_7 = (3, 6)$. Vous utiliserez les centroïdes initiaux $c_1 = (2, 4)$ et $c_2 = (5, 2)$. Détaillez les calculs intermédiaires.

25. Utilisez dbscan et la distance euclidienne pour regrouper les 7 points : $x_1 = (2, 4)$, $x_2 = (5, 4)$, $x_3 = (6, 1)$, $x_4 = (7, 2)$, $x_5 = (4, 4)$, $x_6 = (2, 6)$, et $x_7 = (1, 1)$. Vous utiliserez $m = 2$ (le nombre minimum de points pour créer un cluster) et $\epsilon = 2$ (la distance minimum pour considérer deux points voisins). Détaillez les calculs intermédiaires.

26. Appliquez l'approche agglomérative *single-linkage* avec la distance de Manhattan aussi appelé L_1. Les points de cet exercice sont les mêmes que pour le précédent : $a = (2, 4)$, $b = (0, 1)$, $c = (2, 0)$, $d = (8, 2)$, $f = (9, 1)$, $= (2, 6)$, et $x_7 = (1, 1)$. **Tracez le dendrogramme** sans vous soucier de l'échelle. Détaillez les calculs intermédiaires.

27. Étant donné un ensemble d'enregistrements $X = \{x_1, x_2, \ldots, x_n\}$, où chaque enregistrement $x_i \in X$ est un vecteur réel de dimension d, l'algorithme k-moyennes (en anglais k-means) vise à partitionner les n enregistrements en $k(\leq n)$ clusters $C = \{C_1, C_2, \ldots, C_k\}$ afin de minimiser la somme des distances au carré entre chaque enregistrement x_i et le centroïde de son cluster. Formellement, l'objectif est de minimiser :

$$\sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \eta_j\|^2 \tag{1}$$

où η_j représente le centroïde du cluster j, pour $j = 1, \ldots, k$.

a) Supposez que l'algorithme k-moyennes a été exécuté sur X. Est-ce que la partition obtenue à la fin correspond au **minimum global** pour (1) ? Justifiez votre réponse.

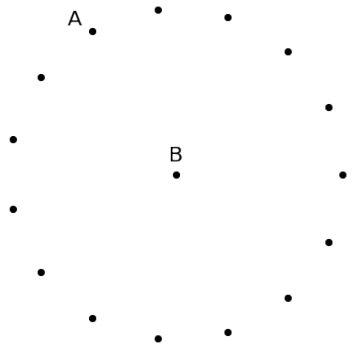
b) Supposez que l'algorithme k-moyennes a été exécuté sur X pour k clusters, et qu'après convergence, une partition avec $k - 1$ clusters est retournée par l'algorithme. Comment pourriez-vous améliorer de façon triviale l'inertie de cette partition ?

Détection de données aberrantes

28. Comment utiliser les k-plus proches voisins pour détecter les données aberrantes ? Quels sont les avantages et les inconvénients de cette approche ?

29. La distance euclidienne est-elle généralement bien adaptée pour détecter des données aberrantes ? Justifiez et donnez un exemple.

30. Considérez les données ci-dessous.

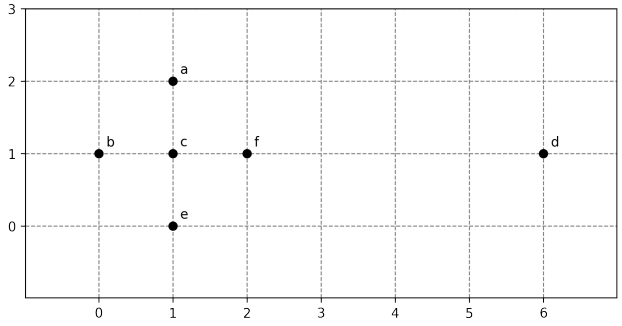


Quel point de A ou B a le plus grand *outlier score* :

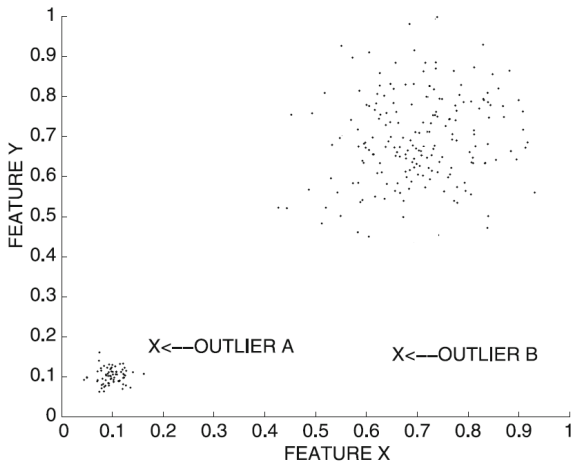
a) Selon un modèle de *clustering*. Considérez que tous les points appartiennent au même cluster et que la méthode utilise la distance euclidienne.

b) Selon un modèle basé sur les distances. Considérez k=2 et que la méthode utilise la distance euclidienne.

31. Considérez le jeu de données suivant par ordre alphabétique et utilisez la distance euclidienne. Nous souhaitons détecter le point avec le plus grand *outlier score* tel que mesuré par la distance à son plus proche voisin. Autrement dit, $k = 1$ et $r = 1$. Appliquez la méthode du Sampling avec l'échantillon $S = \{a, b\}$. Quelles distances n'ont pas été calculées ?



32. Considérez les données présentées dans la figure ci-dessous où on note clairement deux clusters distincts. Dans la figure, c'est aussi possible d'observer deux *outliers* A et B.



a) Lequel de deux *outliers* est plus difficile à détecter automatiquement ? Pourquoi ?

b) Est-ce que l'*outlier* indiqué en (a) peut être identifié par son *outlier score* obtenu en utilisant un modèle de k plus proches voisins basé sur des distances euclidiennes ? Justifiez.

Fouille de graphe

33. Quelle est la distance d'édition entre G_1 et G_2 ? Donnez les opérations.

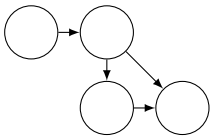


FIGURE 1 – Graphe $G_1 = (N_1, A_1)$

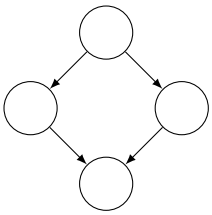


FIGURE 2 – Graphe $G_2 = (N_2, A_2)$

34. Quel est le graphe produit de G_1 et G_2 ? À quoi sert-il ?

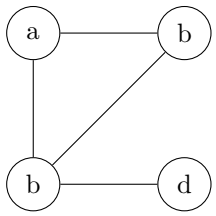


FIGURE 3 – Graphe $G_1 = (N_1, A_1)$

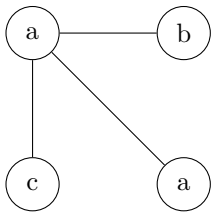


FIGURE 4 – Graphe $G_2 = (N_2, A_2)$

35. Cochez la ou les caractéristiques du graphe correspondant à chaque scénario.

a) Un réseau social comme Facebook, où deux utilisateurs peuvent être amis :

- ◯ orienté / *directed*
- ◯ complet / *dense*
- ◯ pondéré / *weighted*
- ◯ libre / *free*
- ◯ simple / *simple*
- ◯ topologique / *topological*

b) Un réseau social comme Twitter, où un utilisateur peut en suivre un autre :

- ◯ orienté / *directed*
- ◯ complet / *dense*
- ◯ pondéré / *weighted*
- ◯ libre / *free*
- ◯ simple / *simple*
- ◯ topologique / *topological*

c) Le réseau routier canadien :

- ◯ orienté / *directed*
- ◯ complet / *dense*
- ◯ pondéré / *weighted*
- ◯ libre / *free*
- ◯ simple / *simple*
- ◯ topologique / *topological*

36. Quel est le maximum commun sous-graphe (MCS) à G_1 et G_2 ? Quelle est la distance normalisée selon le MCS ?

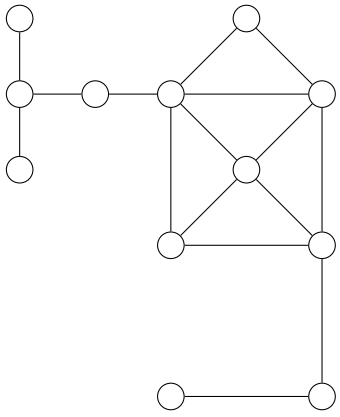


FIGURE 5 – Graphe $G_1 = (N_1, A_1)$

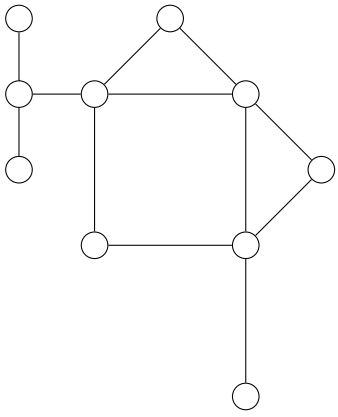


FIGURE 6 – Graphe $G_2 = (N_2, A_2)$

37. Quel est le maximum commun sous-graphe (MCS) à G_1 et G_2 ? Quelle est la distance normalisée selon le MCS ?

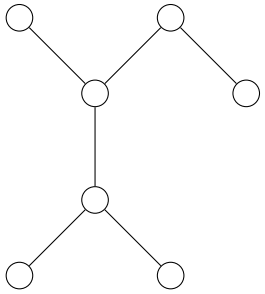


FIGURE 7 – Graphe $G_1 = (N_1, A_1)$

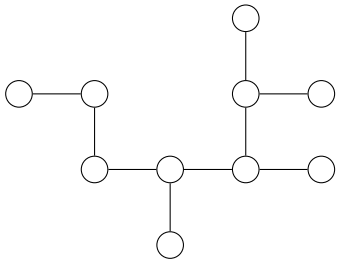


FIGURE 8 – Graphe $G_2 = (N_2, A_2)$

Big Data

38. Quels sont les trois Vs qui définissent le *big data* ? Donnez un exemple de données pour chaque caractéristique.
39. Expliquez deux biais ou limitations du *big data* ?
40. Qu'est-ce que MapReduce ? Donnez un schéma de son fonctionnement.
41. Soient deux matrices $\boldsymbol{M} \in \mathbb{R}^{p \times q}$ et $\boldsymbol{N} \in \mathbb{R}^{q \times r}$, et $\boldsymbol{P} = \boldsymbol{MN}$ leur produit matriciel. Comment utiliser `Map` et `Reduce` pour calculer P. Rappel : $p_{i,k} = \sum_j m_{i,j} \times n_{j,k}$.
42. Soit un ensemble de documents $\mathcal{D}$ où le i-ième document est identifié par un nom unique $name_i$ et contient une liste de mots $d_i = [w_1^{(i)}, w_2^{(i)}, w_2^{(i)}, \dots]$. Implémentez en pseudo-code une seule fonction Map et une seule fonction Reduce afin de calculer la *term frequency* (nombre d'occurrences du mot) et la *document frequency* (nombre de documents qui contiennent le mot) où la *document frequency* df_i du i-ème mot t_i est $df_i = |\{d_j : t_i \in d_j\}|$. Utilisez des noms de variables explicites et/ou commentez votre code.
43. Concevez un algorithme MapReduce qui prend en entrée de gros fichiers contenant des nombres entiers et qui produit comme sortie les entiers plus petits que 50 sans répétition. Utilisez des noms de variables explicites ou commentez votre code.
    Exemple d'entrée : `(10, 102, 100, 12, 200, 2, 10)`
    Exemple de sortie : `(10, 12, 2)`

Fouille de flot de données

44. Donnez trois contraintes de la fouille de flots de données. À quelle caractéristique du *big data* les flots de données correspondent-ils ?
45. Soit le flot de données suivant. Considérez une sous-fenêtre de taille 8, une de taille 4, une de taille 2 et deux de taille 1. Estimez le nombre de 1 pour $k = 11$ avec DGIM, et calculez l'erreur relative.

FIGURE 9 – Flot de données binaires.

46. Soit $\alpha = \frac{|B|}{|S|}$ le ratio entre le nombre de *bucket* et le nombre de clés à filtrer. Entre quelles valeurs de α est-il optimal d'avoir un filtre de Bloom avec 5 fonctions de hash différentes ?
47. Supposons que l'on souhaite filtrer $|S| = 10^9$ adresses mail non-spam avec une seule fonction de hash. Cependant, la quantité de mémoire est limitée à $|B| = 100\text{Mo} = 8.10^8$.
    a) Donnez une première estimation du taux de faux positifs en n'utilisant que la propriété suivante : *la fonction hash chaque clé uniformément.*

b) Donnez une meilleure estimation du taux de faux positifs. Rappel $(1-\epsilon)^{\frac{1}{\epsilon}} = \frac{1}{e}$ pour ϵ petit.

c) Supposons que l'on utilise maintenant un filtre de Bloom. Quel est le nombre de fonctions de hash optimal ? Commentez.

Fouille du web

48. Donnez l'équation algébrique de PageRank. Expliquez chacun des termes c.-à-d. ce qu'ils contiennent et à quoi ils servent.

49. Considérez le graphe ci-dessous. Calculez 4 itérations de PageRank avec une probabilité de téléportation de 0.1.

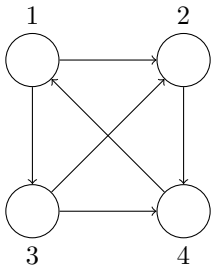


FIGURE 10 – Graphe $G = (N, A)$

50. Considérez le graphe ci-dessous. Calculez 4 itérations de PageRank avec une probabilité de téléportation de $(1-\beta) = 0.2$.

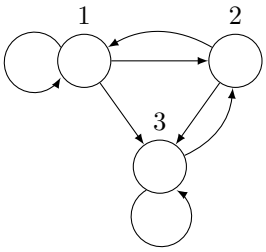
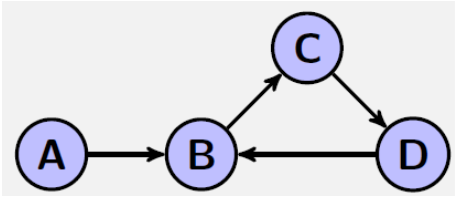


FIGURE 11 – Graphe $G = (N, A)$

51. Obtenez la plus grande valeur de $rang(v)$ pour le graphe ci-dessous après trois itérations de *PageRank*. Utilisez $d = 0.8$. Initialisez $rang(v) = 0.25$ pour tous les sommets.



52. Plutôt que d'utiliser la téléportation, le problème des dead-ends est résolu en les supprimant. Pour supprimer une dead-end, le sommet et toutes ses arrêtes entrantes sont retirés du graphe. En se faisant, d'autres dead-ends sont créées et doivent être aussi supprimées récursivement.

Quel va être l'impact sur la structure du web considérée dans le calcul de PageRank ? Autrement dit, quelle(s) partie(s) du web va (vont) être retirée(s) avant le calcul de PageRank ?

Fouille des réseaux sociaux

53. Expliquez ce qu'est une mesure de centralité et de prestige d'un sommet ? Donnez un exemple de chaque (équation).
54. Qu'est-ce que le modèle nul G' d'un graphe $G = (N, A)$? Montrer que le nombre attendu d'arêtes pour le multigraphe G' est égal à $|A|$.
55. Considérez le graphe ci-dessous. Pour chacune des questions, donnez la formule/équation utilisée.

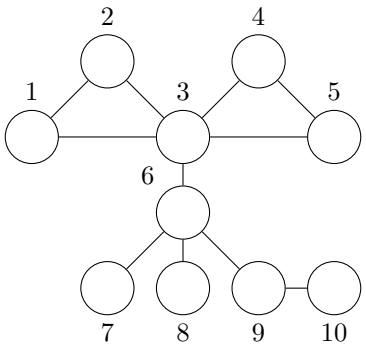


FIGURE 12 – Graphe $G = (N, A)$

a) Quel est le coefficient de regroupement de G ?
b) Quelle est la centralité de degré du sommet 3 ?
c) Quelle est la centralité de proximité du sommet 6 ?
d) Quelle est la centralité d'intermédiarité du sommet 6 ?

Considérez le graphe ci-dessous. Pour chacune des questions, donnez la formule/équation utilisée.

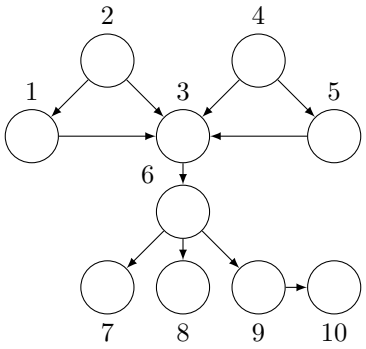


FIGURE 13 – Graphe $G = (N, A)$

a) Quel est le prestige de degré du sommet 3 ?
b) Quel est le prestige de proximité du sommet 6 ?

56. Considérez le graphe ci-dessous. Classifiez le sommet 4 en utilisant 3 itérations de l'algorithme des marches aléatoires.

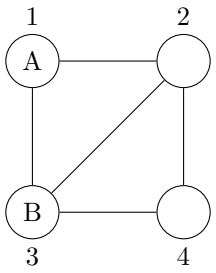
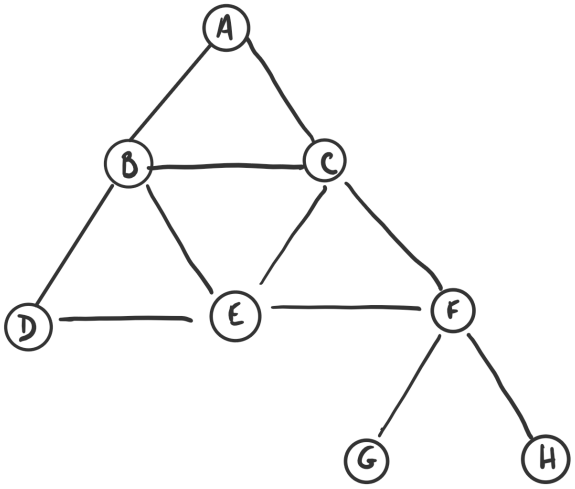


FIGURE 14 – Graphe $G = (N, A)$

57. Calculez pour le graphe ci-dessous :



a) La valeur de son coefficient de regroupement.

b) La plus grande valeur de la *centralité de proximité* trouvée parmi ses sommets.