Projet : Analyse de la Polarisation

Le sujet est écrit au féminin générique. On étudie dans ce sujet la question de la polarisation des préférences exprimées lors de processus électoraux. On dit qu'une élection est polarisée lorsque l'électorat peut être divisé en deux clusters de votantes aux positions idéologiques très distinctes. On cherche donc un homomorphisme de $(\mathcal{E}, R)$ vers $(\mathbb{R}, \geq)$, où $\mathcal{E}$ est l'ensemble des élections possibles et R est une relation binaire sur $\mathcal{E}$ telle que xRy ssi l'élection x est jugée plus polarisée que l'élection y. Le sujet explore différentes manières de définir une telle mesure de la polarisation. On conseille de lire l'intégralité du sujet avant de commencer le projet.

Quelques consignes. Ce travail est à réaliser en trinôme et en utilisant le langage de programmation Python. Votre travail devra être rendu sur Moodle pour le 29/03/2026.

Rendu. Le rendu de votre projet doit comprendre un rapport au format PDF et une archive de votre code. On attend bien sûr un code bien organisé et commenté, muni d'un fichier ReadMe. Votre rapport contiendra les réponses ou des explications pour les questions 1, 2, 4, 6, 7, 9, 10, 11, et 15. Vous mentionnerez toute source ou site Web qui a été utilisé pour réaliser le projet. Votre rapport contiendra également un retour critique et honnête sur votre utilisation des IAs génératives. Vous communiquerez 1) les différents types et le nombre de prompts utilisés, 2) les retours critiques que vous avez eus sur le code et le texte généré, et 3) les avantages et les inconvénients que vous avez identifiés sur l'utilisation de ces outils.

1 Votantes, Candidates, et Bulletins de Vote

Lors d'une élection, un ensemble de n votantes $V = \{v_1, \ldots, v_n\}$ expriment leurs préférences sur un ensemble de m candidates $C = \{c_1, \ldots, c_m\}$, donnant n bulletins de vote. On supposera que n est m sont pairs. On considère deux types de bulletins de vote.

- *Votes par approbations* : chaque votante indique les candidates qu'elle approuve. On obtient alors pour la votante v_i un vecteur $a_{v_i} \in \{0,1\}^m$, où $a_{v_i}[j] = 1$ ssi v_i approuve c_j. On note $\mathcal{A} = \{0,1\}^m$ l'ensemble des votes par approbations possibles. Soit $a \in \mathcal{A}$, on note $\overline{a}$ le bulletin opposé où $\overline{a}[j] = 1 - a[j], \forall j \in \{1, \ldots, m\}$. Enfin, soit $a \in \mathcal{A}$ et π une permutation sur $\{1, \ldots, m\}$, on note a^π le bulletin où $a^\pi[j] = a[\pi(j)], \forall j \in \{1, \ldots, m\}$, i.e., le vecteur où les indices sont permutés par π.
- *Votes par ordres totaux* : chaque votante indique un rangement sur les candidates, de sa candidate la plus préférée à celle la moins préférée. On obtient alors pour la votante v_i un ordre total $\succeq_{v_i}$ sur C. À tout ordre total $\succeq_{v_i}$, on associe une fonction de rang $r_{\succeq_{v_i}} : C \to \{1, \ldots, m\}$ où $r_{\succeq_{v_i}}(c_j)$ désigne la position de la candidate c_j dans le classement de la votante v_i (e.g., $r_{\succeq_{v_i}}(c_j) = 1$ si c_j est la candidate préférée de v_i). On note $\mathcal{L}$ l'ensemble des votes par ordres totaux possibles, ce qui peut être vu comme l'ensemble des bijections de C dans $\{1, \ldots, m\}$. Soit $\succeq \in \mathcal{L}$, on note $\overline{\succeq}$ le bulletin opposé où $r_{\overline{\succeq}}(c) = m - r_{\succeq}(c) + 1, \forall c \in C$. Enfin, soit $\succeq \in \mathcal{L}$ et π une permutation sur C, on note $\succeq^\pi$ le bulletin où $r_{\succeq^\pi}(c) = r_{\succeq}(\pi(c)), \forall c \in C$, i.e., l'ordre total où les candidates sont permutées par π.

On nomme profil de l'élection (ou profil pour faire court) l'ensemble des n bulletins de vote exprimés par les votantes. Un profil est donc un élément de $\mathcal{A}^n$ ou $\mathcal{L}^n$. Soit π une permutation sur $\{1, \ldots, m\}$ (resp. C) et p un profil dans $\mathcal{A}^n$ (resp. $\mathcal{L}^n$), on note p^π le profil où chaque bulletin $a \in p$ (resp. $\succeq \in p$) est remplacé par a^π (resp. $\succeq^\pi$). Soit σ une permutation sur $\{1, \ldots, n\}$ on note ${}^\sigma p$ le profil où $p[i]$ est remplacé par $p[\sigma(i)]$ pour tout $i \in \{1, \ldots, n\}$. Dit autrement, alors que p^π permute les labels des candidates, ${}^\sigma p$

permute les labels des votantes. Enfin, soit $k \in \mathbb{N}^*$ on note kp le profil de taille $n \times k$ où $p[i]$ est remplacé par k copies de $p[i]$.

On présente désormais deux profils aux niveaux de polarisation opposés.

— Votes par approbations. Soit $a \in \mathcal{A}$: on note p_a le profil consistant en la répétition n fois du bulletin a. On note $p_{a,\overline{a}}$ le profil consistant en la répétition $n/2$ fois du bulletin a et la répétition $n/2$ fois du bulletin $\overline{a}$. Alors que p_a est un profil où la polarisation semble minimale, $p_{a,\overline{a}}$ est un profil où la polarisation semble maximale.

— Votes par ordres totaux. Soit $\succeq \in \mathcal{L}$: on note $p_{\succeq}$ le profil consistant en la répétition n fois du bulletin $\succeq$. On note $p_{\succeq,\overline{\succeq}}$ le profil consistant en la répétition $n/2$ fois du bulletin $\succeq$ et la répétition $n/2$ fois du bulletin $\overline{\succeq}$. Alors que $p_{\succeq}$ est un profil où la polarisation semble minimale, $p_{\succeq,\overline{\succeq}}$ est un profil où la polarisation semble maximale

Question 1. Implémenter une méthode permettant de générer aléatoirement un profil dans $\mathcal{A}^n$. Votre méthode prendra en paramètre un ou des arguments pour contrôler le niveau de polarisation de l'élection. Une élection peu polarisée devra ressembler à un profil p_a tandis qu'une élection fortement polarisée devra ressembler à un profil $p_{a,\overline{a}}$. On vous laisse la liberté de choisir comment générer ces élections et comment contrôler d'une certaine façon le niveau de polarisation.

Question 2. Implémenter une méthode permettant de générer aléatoirement un profil dans $\mathcal{L}^n$. Votre méthode prendra en paramètre un ou des arguments pour contrôler le niveau de polarisation de l'élection. Une élection peu polarisée devra ressembler à un profil $p_{\succeq}$ tandis qu'une élection fortement polarisée devra ressembler à un profil $p_{\succeq,\overline{\succeq}}$. On vous laisse la liberté de choisir comment générer ces élections et comment contrôler d'une certaine façon le niveau de polarisation.

2 Axiomes sur la Polarisation

On présente désormais des axiomes qu'une mesure de polarisation φ devrait satisfaire.

Axiome 1 (Régularité). *$\varphi(p) \in [0,1]$ pour chaque $p \in \mathcal{A}^n$ (resp. $p \in \mathcal{L}^n$). De plus, on demande les conditions suivantes :*

— *$\varphi(p) = 0$ ssi $p = p_a$ pour un bulletin $a \in \mathcal{A}$ (resp. $p = p_{\succeq}$ pour un bulletin $\succeq \in \mathcal{L}$).*

— *$\varphi(p) = 1$ ssi $p = p_{a,\overline{a}}$ pour un bulletin $a \in \mathcal{A}$ (resp. $p = p_{\succeq,\overline{\succeq}}$ pour un bulletin $\succeq \in \mathcal{L}$).*

Axiome 2 (Neutralité). *$\varphi(p) = \varphi(p^\pi)$ pour toute permutation π sur $\{1,\dots,m\}$ (resp. C) et tout profil $p \in \mathcal{A}^n$ (resp. $p \in \mathcal{L}^n$).*

Axiome 3 (Anonymité). *$\varphi(p) = \varphi({}^\sigma p)$ pour toute permutation σ sur $\{1,\dots,n\}$ et tout profil $p \in \mathcal{A}^n$ (resp. $p \in \mathcal{L}^n$).*

Axiome 4 (Invariance par réplication). *$\varphi(kp) = \varphi(p)$ pour tout profil $p \in \mathcal{A}^n$ (resp. $p \in \mathcal{L}^n$) et entier $k \in \mathbb{N}^*$.*

Soit C^2 l'ensemble des sous-ensembles de C de taille 2. Étant donné un profil p et $\{c_k, c_l\} \in C^2$, on note $n_{c_k c_l}(p)$ le nombre de votantes qui préfèrent c_k à c_l, i.e., $n_{c_k c_l}(p) = |\{v | a_v[k] = 1 \wedge a_v[l] = 0\}|$ pour des votes par approbations et $n_{c_k c_l}(p) = |\{v | c_k \succeq_v c_l\}|$ pour des votes par ordres totaux. Soit $d_{c_k c_l}(p) = |n_{c_k,c_l}(p) - n_{c_l,c_k}(p)|$ la différence absolue entre le nombre de votantes préférant c_k à c_l et celles préférant c_l à c_k pour le profil p.

Question 3. Implémenter une méthode permettant de calculer l'ensemble des valeurs $\{d_{c_k c_l}(p), \{c_k, c_l\} \in C^2\}$ pour un profil $p \in \mathcal{A}^n$ et une méthode permettant de calculer l'ensemble des valeurs $\{d_{c_k c_l}(p), \{c_k, c_l\} \in C^2\}$ pour un profil $p \in \mathcal{L}^n$.

En inspectant les profils portant uniquement sur deux candidates, on peut avoir l'intuition que la polarisation d'un profil sera corrélée aux valeurs $d_{c_k c_l}(p)$ pour $\{c_k, c_l\} \in C^2$. Cela nous mène à la mesure de polarisation φ^2 suivante :

$$\varphi^2(p) = \sum_{\{c_k,c_l\}\in C^2} \frac{n - d_{c_k c_l}(p)}{n\binom{m}{2}}.$$

Pour $\{c_k, c_l\} \in C^2$, $n - d_{c_k c_l}(p)$ est égale à 0 si toutes les votantes ont la même préférence entre les candidates c_k et c_l (polarisation faible), et est égale à n si la moitié des votantes préférent c_k et c_l et l'autre moitié préférent c_l et c_k (polarisation forte).

Question 4. Quels axiomes sont vérifiés par φ^2 ? Si un axiome est vérifié, vous fournirez une preuve. Si un axiome n'est pas vérifié, vous fournirez un contre-exemple.

Question 5. Implémenter une méthode permettant de calculer $\varphi^2(p)$ pour un profil $p \in \mathcal{A}^n$ et une méthode permettant de calculer $\varphi^2(p)$ pour un profil $p \in \mathcal{L}^n$.

Question 6. En utilisant les méthodes des questions 1,2, et 5 tracer l'évolution de la mésure $\varphi^2(p)$ en faisant varier "le niveau" de polarisation des instances à partir des arguments choisis aux questions 1 et 2.

3 Distances et Mesures de Polarisation

Nous allons désormais étudier d'autres mesures de polarisation basées sur le concept de distance entre bulletins de vote. Soit X un ensemble non vide. Une application $d : X \times X \to \mathbb{R}_+$ est une *distance* sur X si elle vérifie les quatre propriétés suivantes :

1. *Positivité* : $\forall x, y \in X,\ d(x, y) \geq 0$,
2. *Séparation* : $\forall x, y \in X, d(x, y) = 0 \iff x = y$,
3. *Symétrie* : $\forall x, y \in X, d(x, y) = d(y, x)$,
4. et *Inégalité triangulaire* : $\forall x, y, z \in X, d(x, z) \leq d(x, y) + d(y, z)$.

Les distances permettent de formaliser la notion de *dissimilarité* entre des objets qui peuvent être de natures très distinctes. Nous présentons désormais une distance qui s'applique sur l'ensemble $\mathcal{A}$, et une distance qui s'applique sur l'ensemble $\mathcal{L}$.

La distance de Hamming. Soient $a_{v_k}, a_{v_l} \in \mathcal{A}$. La *distance de Hamming* est définie par : $d_H(a_{v_k}, a_{v_l}) = \sum_{i=1}^{m} \mathbb{1}_{a_{v_k}[i]\neq a_{v_l}[i]}$ où $\mathbb{1}_{a_{v_k}[i]\neq a_{v_l}[i]} = 1$ si $a_{v_k}[i] \neq a_{v_l}[i]$, et 0 sinon. Pour nos votes par approbations, il s'agit du nombre de candidates approuvées par l'une des votantes mais pas par l'autre.

Distance de Spearman. Soient $\prec_{v_k}$ et $\prec_{v_l}$ deux ordres totaux sur C, et r_{v_k}, r_{v_l} les fonctions de rang associées. La *distance de Spearman* est définie par : $d_S(\prec_{v_k}, \prec_{v_l}) = \sum_{i=1}^{m} |r_{v_k}(i) - r_{v_l}(i)|$. Pour nos votes par ordres totaux, cette distance agrége les distances de rang des différentes candidates quand on compare les rangements des votantes v_k et v_l.

Question 7. Montrer que d_S et d_H sont bien des fonctions de distance.

Question 8. Implémenter des méthodes permettant de calculer ces distances pour deux bulletins $a, a' \in \mathcal{A}$ ou pour deux bulletins $\succeq, \succeq' \in \mathcal{L}$.

Nous introduisons désormais les deux problèmes d'optimisation suivants.

— Étant donné un profil de votes p, trouver un bulletin $a \in \mathcal{A}$ qui minimise $\sum_{a'\in p} d_H(a, a')$. Dans le cas des préférences par ordres totaux, on remplacera d_H par d_S. On notera $u_1^*(p)$ la valeur optimale de ce problème. Ce problème consiste à trouver un bulletin de consensus pour l'ensemble des votantes. Plus $u_1^*(p)$ est petit est plus l'ensemble des votantes sont proches de ce consensus global.

— Étant donné un profil de votes p, trouver deux bulletins $a_1, a_2 \in \mathcal{A}$ qui ensemble minimisent $\sum_{a' \in p} \min\{d_H(a_1, a'), d_H(a_2, a')\}$. Dans le cas des préférences par ordres totaux, on remplacera d_H par d_S. On notera $u_2^*(p)$ la valeur optimale de ce problème. Ce problème consiste à trouver un bulletin de consensus pour deux clusters de votantes (un consensus par cluster), où chaque votante est affectée au cluster dont le bulletin de consensus est le plus proche. Plus $u_2^*(p)$ est petit est plus les votantes de chacun des deux clusters sont proches de leurs consensus respectuctifs.

Nous introduisons désormais les deux mesures de polarisation φ_{d_H} et φ_{d_S}, définies comme suit :

$$\varphi_{d_H}(p) = \frac{2}{n \times m}(u_1^*(p) - u_2^*(p)) \text{ et } \varphi_{d_S}(p) = \frac{4}{n \times m^2}(u_1^*(p) - u_2^*(p)).$$

Si $\varphi_{d_H}(p)$ est élevée alors les votantes se sentent lointaines d'un consensus global, mais forment des ensembles aux votes proches si on les autorisent à former deux clusters. Si $\varphi_{d_H}(p)$ est faible, alors le fait de faire un ou deux clusters de votantes ne change pas beaucoup la proximité des votantes à leurs consensus respectifs. Cela est bien cohérent avec la notion de polarisation.

Question 9. Pour mieux comprendre les propriétés de φ_{d_H} et φ_{d_S}, indiquer quels axiomes sont vérifiés par φ_{d_H}.

Question 10. Montrer comment $u_1^*(p)$ peut être calculée efficacement pour des votes par approbations.

Question 11. Montrer comment $u_1^*(p)$ peut être calculée efficacement pour des votes par ordres totaux. Indication : regarder le problème de couplage parfait de poids minimum dans un graphe biparti (le module python scipy permet de résoudre ce problème d'optimisation).

Question 12. Implémenter une méthode pour calculer $u_1^*(p)$ pour des votes par approbations et une méthode pour calculer $u_1^*(p)$ pour des votes par ordres totaux.

Pour calculer $u_2^*(p)$, nous utiliserons l'algorithme itératif k-means utilisé pour des tâches de *clustering*.

Algorithme 1 : k-means avec $k = 2$ adapté au cas des votes par approbations.

1 Entrée : $(a_1, \ldots, a_n) \in \mathcal{A}^n$.
2 Initialiser deux centroïdes aléatoirement $\tilde{a}_1, \tilde{a}_2 \in \mathcal{A}$. (un centroïde est ici un bulletin de consensus)
3 **repeat**
4 Affecter chaque bulletin de vote au cluster dont la distance au centroïde est la plus faible ; $\texttt{cluster}(a_i) = 1$ si $d_H(a_i, \tilde{a}_1) \leq d_H(a_i, \tilde{a}_2)$ et $\texttt{cluster}(a_i) = 2$ sinon, $\forall i \in \{1, \ldots, n\}$;
5 Trouver le centroïde de chaque cluster $\tilde{a}_1 \in arg\min_{a \in \mathcal{A}} \sum_{a_t : \texttt{cluster}(a_t)=1} d_H(a, a_t)$ et $\tilde{a}_2 \in arg\min_{a \in \mathcal{A}} \sum_{a_t : \texttt{cluster}(a_t)=2} d_H(a, a_t)$;
6 **until** *Tant qu'il y a des changements dans l'affectation des clusters*
7 **return** $\sum_{a_t : \texttt{cluster}(a_t)=1} d_H(\tilde{a}_1, a_t) + \sum_{a_t : \texttt{cluster}(a_t)=2} d_H(\tilde{a}_2, a_t)$

Notons que l'étape à la ligne 5 de l'algorithme demande d'utiliser votre réponse à la question 12. L'algorithme de k-means converge vers un optimum local. Pour trouver une valeur qui nous l'espérons sera souvent proche de l'optimum global, une approche consiste à relancer cette algorithme plusieurs fois. La valeur finale retenue sera le résultat minimum obtenu sur tous les lancés. L'algorithme des k-means retournera pour nous une estimation de $u_2^*(p)$ que nous noterons $\tilde{u}_2^*(p)$.

Question 13. Implémenter une méthode pour calculer $\tilde{u}_2^*(p)$ pour des votes par approbations et une méthode pour calculer $\tilde{u}_2^*(p)$ pour des votes par ordres totaux.

Question 14. Implémenter une méthode pour calculer $\varphi_{d_H}(p)$ pour des votes par approbations et une méthode pour calculer $\varphi_{d_S}(p)$ pour des votes par ordres totaux.

Question 15. En utilisant les méthodes des questions 1,2, et 14 tracer l'évolution des mésures $\varphi_{d_H}(p) = \frac{2}{n \times m}(u_1^*(p) - \tilde{u}_2^*(p))$ et $\varphi_{d_S}(p) = \frac{4}{n \times m^2}(u_1^*(p) - \tilde{u}_2^*(p))$ en faisant varier "le niveau" de polarisation des instances comme à la question 6.