Contrôler la distance au consensus de Kemeny

Yunlong Jiao* Anna Korba† Eric Sibony†

*Mines ParisTech, †LTCI, Telecom ParisTech/CNRS

Journée de la Chaire Machine Learning For Big Data, 10 Juin 2016

Plan

L'agrégation de rankings

Problème:

Comment agréger les préférences de plusieurs agents sur un ensemble fixé d'objets?

Entrée

- n objets: $[\![n]\!] := \{1, \ldots, n\}$.
- N agents qui donnent un classement de ces objets: $i_1 \succ i_2 \succ \cdots \succ i_n$

Sortie

Un ordre global ("consensus") σ^* sur les n objets.

Applications

Exemple 1: Elections

- Soit un ensemble de candidats $\{A, B, C, D\}$.
- Chaque électeur donne un classement complet des candidats, exemple: $B \succ D \succ A \succ C$
- L'ensemble des votes récoltés pour l'élection constitue un **dataset de rankings**.

$\Rightarrow$ Comment élire le (les) vainqueur(s)?

Débat Borda-Condorcet depuis le 18ème siècle

Jean-Charles de Borda

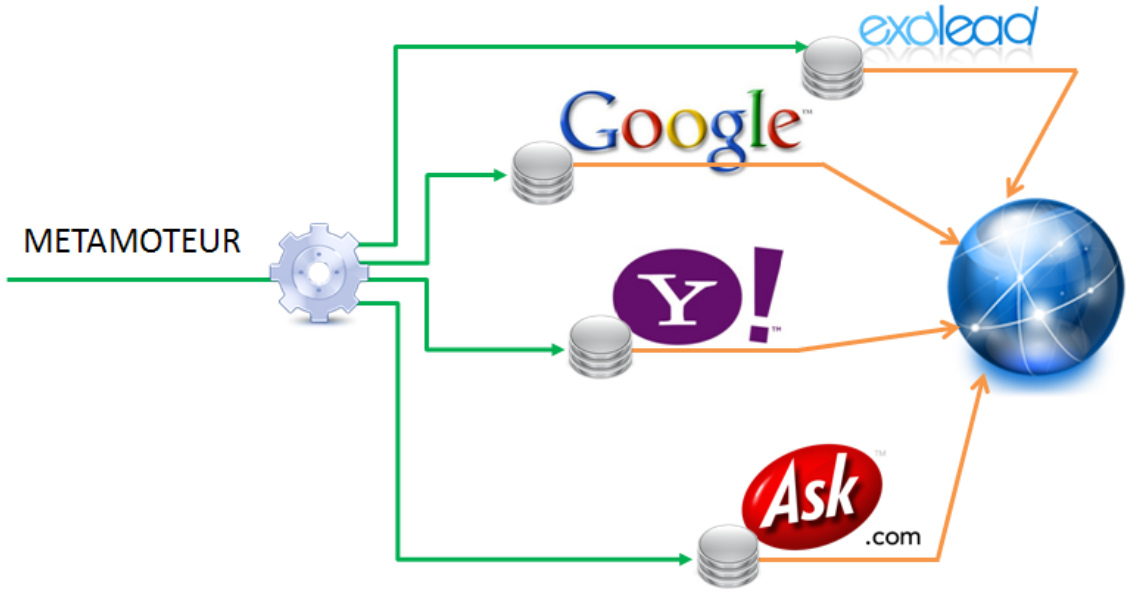
Nicolas de Condorcet

Applications

Exemple 2: Métamoteurs

Pour une requête q donnée, un métamoteur retourne les résultats de plusieurs moteurs de recherche.

$\Rightarrow$ Comment agréger les listes ordonnées de tous ces moteurs?

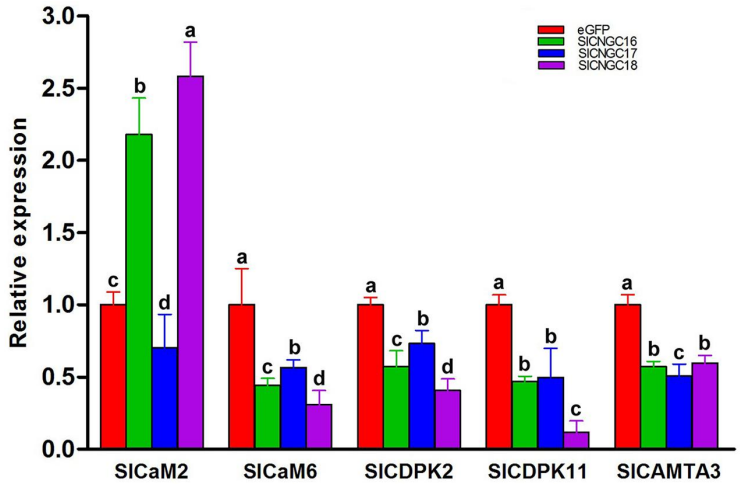


Applications

Exemple 3: Expression des gènes

- Développement des puces à ADN: mesure des niveaux d'expression simultanés de milliers de gènes ou protéines.
- Ces mesures peuvent grandement varier en échelle!
- Une possibilité est d'ordonner les gènes par leur niveau d'expression lors de chaque expérience.

$\Rightarrow$ Comment agréger ces observations?



L'agrégation de rankings

Ensemble d'objets: $[\![n]\!] := \{1, \ldots, n\}$
Ranking $i_1 \succ \cdots \succ i_n$ on $[\![n]\!] \Longleftrightarrow$ permutation σ sur $[\![n]\!]$ t.q. $\sigma(i_j) = j$.

Quelle permutation $\sigma^* \in \mathfrak{S}_n$ représente le mieux la collection de permutations $(\sigma_1, \ldots, \sigma_N) \in \mathfrak{S}_n^N$?

L'agrégation au sens de Kemeny (1)

Definition (Consensus ranking (Kemeny, 1959))

Une permutation $\sigma^* \in \mathfrak{S}_n$ est une meilleure représentation de la collection $(\sigma_1, \ldots, \sigma_N) \in \mathfrak{S}_n^N$ par rapport à la métrique d sur $\mathfrak{S}_n$ si c'est une solution de :

$$min_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{N} d(\sigma, \sigma_t).$$

L'agrégation au sens de Kemeny (1)

Definition (Consensus ranking (Kemeny, 1959))

Une permutation $\sigma^* \in \mathfrak{S}_n$ est une meilleure représentation de la collection $(\sigma_1, \ldots, \sigma_N) \in \mathfrak{S}_n^N$ par rapport à la métrique d sur $\mathfrak{S}_n$ si c'est une solution de :

$$min_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{N} d(\sigma, \sigma_t).$$

$\rightarrow$ Quelle distance sur les permutations?

L'agrégation au sens de Kemeny (2)

Definition

(Distance de Kendall)

$$d(\sigma, \pi) = \sum_{\{i,j\} \subset [\![n]\!]} \{\sigma \text{ and } \pi \text{ disagree on } \{i, j\}\}$$

L'agrégation au sens de Kemeny (2)

Definition

(Distance de Kendall)

$$d(\sigma, \pi) = \sum_{\{i,j\} \subset [\![n]\!]} \{\sigma \text{ and } \pi \text{ disagree on } \{i,j\}\}$$

Exemple

σ= 123 ($1 \succ 2 \succ 3$)
π= 231 ($2 \succ 3 \succ 1$)
$\rightarrow$ nombre de désaccords = sur 2 paires (12,13).

L'agrégation au sens de Kemeny (2)

Definition

(Distance de Kendall)

$$d(\sigma, \pi) = \sum_{\{i,j\} \subset [\![n]\!]} \{\sigma \text{ and } \pi \text{ disagree on } \{i,j\}\}$$

Exemple

σ= 123 ($1 \succ 2 \succ 3$)
π= 231 ($2 \succ 3 \succ 1$)
$\rightarrow$ nombre de désaccords = sur 2 paires (12,13).

$\rightarrow$ **agrégation au sens de Kemeny** = consensus de Kemeny avec la distance de Kendall

Les propriétés du consensus de Kemeny

- Justification en choix social: Satisfait de nombreuses propriétés, comme le critère de Condorcet: si un candidat gagne en duel contre chaque autre candidat, il est vainqueur [Young and Levenglick, 1978]
- Justification statistique: Correspond au paramètre qui maximise la vraisemblance sous le modèle de Mallows [Young, 1988]
- Inconvénient: Ce problème est NP-difficile en terme de nombre de votes N [Bartholdi et al., 1989] même pour $n = 4$ [Dwork et al., 2001].

Plan

Notre approche

Cadre:

- Catalogue d'objets $[\![n]\!] := \{1, \ldots, n\}$
- Un dataset de rankings $\mathcal{D}_N = (\sigma_1, \ldots, \sigma_N) \in \mathfrak{S}_n^N$
- Soit $\sigma \in \mathfrak{S}_n$ une permutation, typiquement la sortie d'une procédure approximative d'agrégation sur $\mathcal{D}_N$.

Notre approche

Cadre:

- Catalogue d'objets $[\![n]\!] := \{1, \ldots, n\}$
- Un dataset de rankings $\mathcal{D}_N = (\sigma_1, \ldots, \sigma_N) \in \mathfrak{S}_n^N$
- Soit $\sigma \in \mathfrak{S}_n$ une permutation, typiquement la sortie d'une procédure approximative d'agrégation sur $\mathcal{D}_N$.

Question: **Peut-on donner une borne supérieure sur la distance $d(\sigma, \sigma^*)$ entre σ et un consensus de Kemeny, en utilisant uniquement des quantités tractables?**

Plan

Kemeny embedding

Le Kemeny embedding est le mapping $\phi : \mathfrak{S}_n \to \mathbb{R}^{\binom{n}{2}}$ défini par:

$$\phi : \sigma \mapsto \begin{pmatrix} \vdots \\ signe(\sigma(i) - \sigma(j)) \\ \vdots \end{pmatrix}_{1 \leq i < j \leq n}$$

avec $signe(x) = 1$ si $x \geq 0$ et -1 sinon.

Exemple

$$123 \mapsto \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \begin{matrix} \to \text{ paire 12} \\ \to \text{ paire 13} \\ \to \text{ paire 23} \end{matrix}, 132 \mapsto \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \begin{matrix} \to \text{ paire 12} \\ \to \text{ paire 13} \\ \to \text{ paire 23} \end{matrix}$$

L'agrégation de Kemeny dans $\mathbb{R}^{\binom{n}{2}}$

On définit le barycentre de $(\sigma_1, \ldots, \sigma_N)$:

$$\phi(\mathcal{D}_N) := \frac{1}{N} \sum_{t=1}^{N} \phi(\sigma_t). \tag{1}$$

L'agrégation de Kemeny dans $\mathbb{R}^{\binom{n}{2}}$

On définit le barycentre de $(\sigma_1, \ldots, \sigma_N)$:

$$\phi(\mathcal{D}_N) := \frac{1}{N} \sum_{t=1}^{N} \phi(\sigma_t). \tag{1}$$

Proposition (Résultats préliminaires)

Pour tout $\sigma, \sigma' \in \mathfrak{S}_n$,

$$\|\phi(\sigma)\| = \sqrt{\frac{n(n-1)}{2}} \quad \textit{and} \quad \|\phi(\sigma) - \phi(\sigma')\|^2 = 4d(\sigma, \sigma'),$$

Pour tout dataset $\mathcal{D}_N = (\sigma_1, \ldots \sigma_N) \in \mathfrak{S}_n^N$, l'agrégation au sens de Kemeny est équivalente au problème de minimisation suivant:

$$\min_{\sigma \in \mathfrak{S}_n} \|\phi(\sigma) - \phi(\mathcal{D}_N)\|^2, \tag{2}$$

L'agrégation de Kemeny dans $\mathbb{R}^{\binom{n}{2}}$

L'agrégation de Kemeny se décompose en deux étapes:

- Calculer le barycentre (1) $\phi(\mathcal{D}_N) \in \mathbb{R}^{\binom{n}{2}}$ (complexité $O(Nn^2)$)
- Trouver le consensus σ^* solution du problème (2)

Illustration

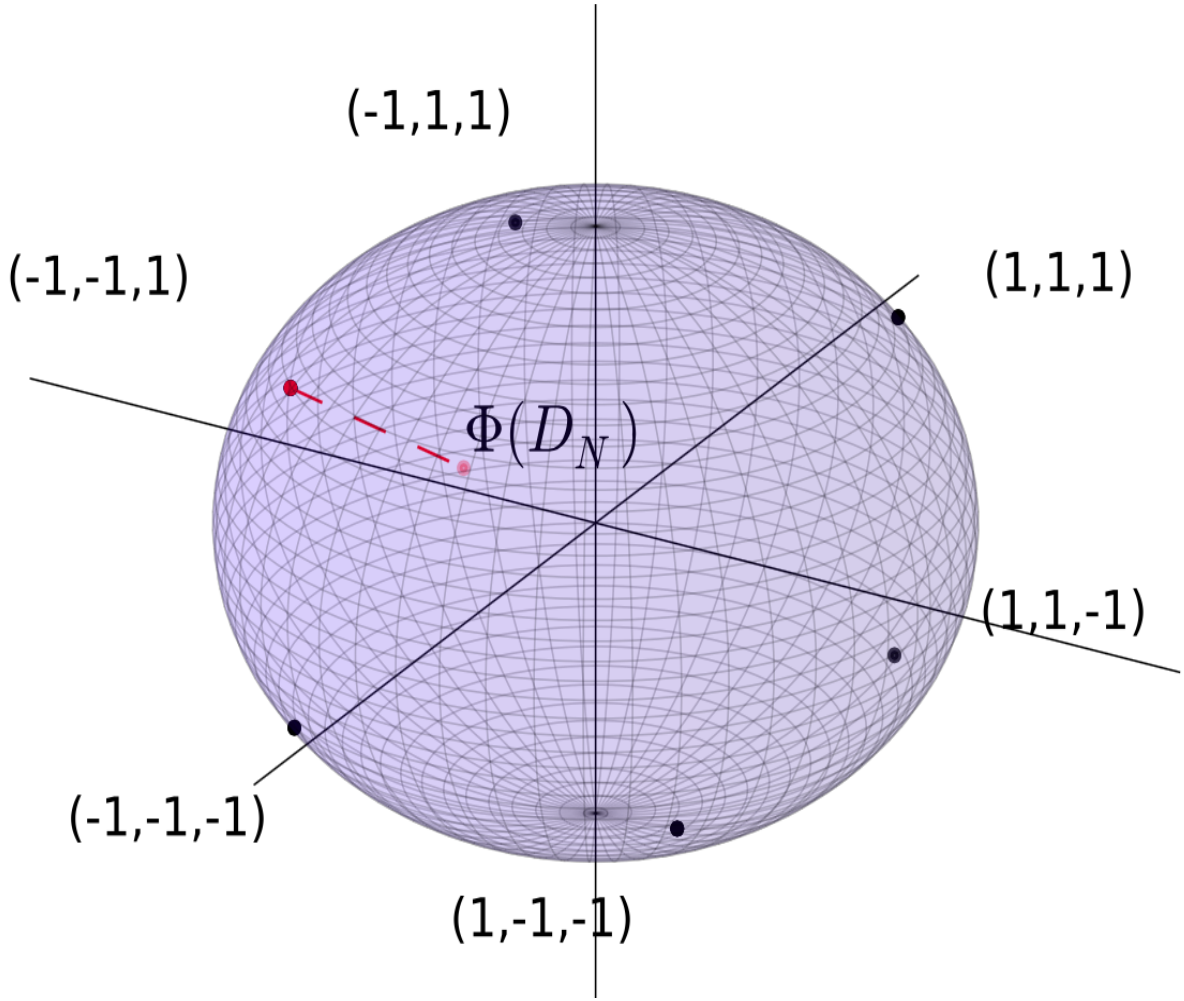


Figure: L'agrégation au sens de Kemeny pour $n = 3$.

Résultat principal

Pour $\sigma \in \mathfrak{S}_n$, nous définissons l'angle $\theta_N(\sigma)$ entre $\phi(\sigma)$ et $\phi(\mathcal{D}_N)$ par:

$$\cos(\theta_N(\sigma)) = \frac{\langle \phi(\sigma), \phi(\mathcal{D}_N) \rangle}{\|\phi(\sigma)\| \|\phi(\mathcal{D}_N)\|}, \tag{3}$$

avec $0 \leq \theta_N(\sigma) \leq \pi$.

Résultat principal

Pour $\sigma \in \mathfrak{S}_n$, nous définissons l'angle $\theta_N(\sigma)$ entre $\phi(\sigma)$ et $\phi(\mathcal{D}_N)$ par:

$$\cos(\theta_N(\sigma)) = \frac{\langle \phi(\sigma), \phi(\mathcal{D}_N) \rangle}{\|\phi(\sigma)\| \|\phi(\mathcal{D}_N)\|}, \tag{3}$$

avec $0 \leq \theta_N(\sigma) \leq \pi$.

Théorème: Soit $\mathcal{D}_N \in \mathfrak{S}_n^N$ un dataset et $\sigma \in \mathfrak{S}_n$ une permutation. Pour tout $k \in \{0, \ldots, \binom{n}{2} - 1\}$, nous avons l'implication suivante:

$$\cos(\theta_N(\sigma)) > \sqrt{1 - \frac{k+1}{\binom{n}{2}}} \quad \Rightarrow \quad \max_{\sigma^* \in \mathcal{K}_N} d(\sigma, \sigma^*) \leq k.$$

Méthode

Nous définissons:

$$k_{min}(\sigma; \mathcal{D}_N) = \left\lfloor \binom{n}{2} \sin^2(\theta_N(\sigma)) \right\rfloor . \tag{4}$$

le $k \in \{0, \ldots, \binom{n}{2} - 1\}$ minimal vérifiant la condition du théorème.

Méthode

Nous définissons:

$$k_{min}(\sigma; \mathcal{D}_N) = \left\lfloor \binom{n}{2} \sin^2(\theta_N(\sigma)) \right\rfloor . \tag{4}$$

le $k \in \{0, \ldots, \binom{n}{2} - 1\}$ minimal vérifiant la condition du théorème.

Deux étapes:

- Calculer $k_{min}(\sigma; \mathcal{D}_N)$ avec la formule (4).
- Puis en applicant le théorème, $d(\sigma, \sigma^*) \leq k_{min}(\sigma; \mathcal{D}_N)$ pour tout consensus de Kemeny $\sigma^* \in \mathcal{K}_N$.

Application sur le dataset Sushi

Table: Etude de cas sur la validité de la méthode avec le dataset Sushi ($N = 5000, n = 10$). Les lignes sont ordonnées par valeurs de k_{min} croissantes (ou cosinus décroissantes).

Voting rule	$\cos(\theta_N(\sigma))$	$k_{min}(\sigma)$
Borda	0.82	14
Copeland	0.82	14
QuickSort	0.82	14
Plackett-Luce	0.80	15
2-approval	0.74	20
1-approval	0.71	22
Pick-a-Perm	0.40	37
Pick-a-Random	0.28	41

Plan

Idée de la preuve

Agrégation de Kemeny:

$$\min_{\sigma \in \mathfrak{S}_n} C'_N(\sigma) = \|\phi(\sigma) - \phi(\mathcal{D}_N)\|^2.$$

Problème relaxé:

$$\min_{x \in \mathbb{S}} \mathcal{C}_N(x) := \|x - \phi(\mathcal{D}_N)\|^2.$$

Idée de la preuve

Agrégation de Kemeny:

$$\min_{\sigma \in \mathfrak{S}_n} C'_N(\sigma) = \|\phi(\sigma) - \phi(\mathcal{D}_N)\|^2.$$

Problème relaxé:

$$\min_{x \in \mathbb{S}} \mathcal{C}_N(x) := \|x - \phi(\mathcal{D}_N)\|^2.$$

Pour tout $x \in \mathbb{S}$, avec R rayon de $\mathbb{S}$, on a:

$$\mathcal{C}_N(x) = R^2 + \|\phi(\mathcal{D}_N)\|^2 - 2R\|\phi(\mathcal{D}_N)\| \cos(\theta_N(x)).$$

Les level sets de $\mathcal{C}_N$ sont donc de la forme $\{x \in \mathbb{S} \mid \theta_N(x) = \alpha\}$, pour $0 \leq \alpha \leq \pi$

Illustration

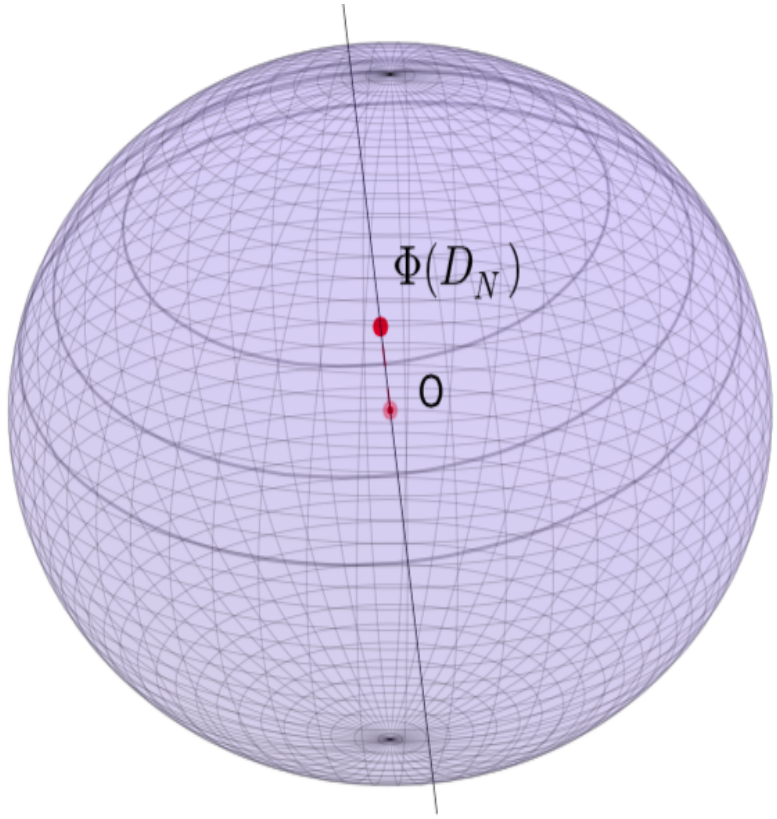


Figure: Level sets de $\mathcal{C}_N$

Lemmas

Lemme 1

Le consensus de Kemeny d'un dataset $\mathcal{D}_N$ est une permutation σ^* t.q:

$$\theta_N(\sigma^*) \leq \theta_N(\sigma) \qquad \text{for all } \sigma \in \mathfrak{S}_n.$$

Lemme 2

Pour $x \in \mathbb{S}$ et $r \geq 0$ on a:

$$\cos(\theta_N(x)) > \sqrt{1 - \frac{r^2}{4R^2}} \Rightarrow \min_{x' \in \mathbb{S} \setminus \mathcal{B}(x,r)} \theta_N(x') > \theta_N(x).$$

Lemme 3

Pour $\sigma \in \mathfrak{S}_n$ et $k \in \{0, \ldots, \binom{n}{2}\}$,

$$\phi\left(\mathfrak{S}_n \setminus B(\sigma, k)\right) \quad \subset \quad \mathbb{S} \setminus \mathcal{B}(\phi(\sigma), 2\sqrt{k+1})$$

Illustration

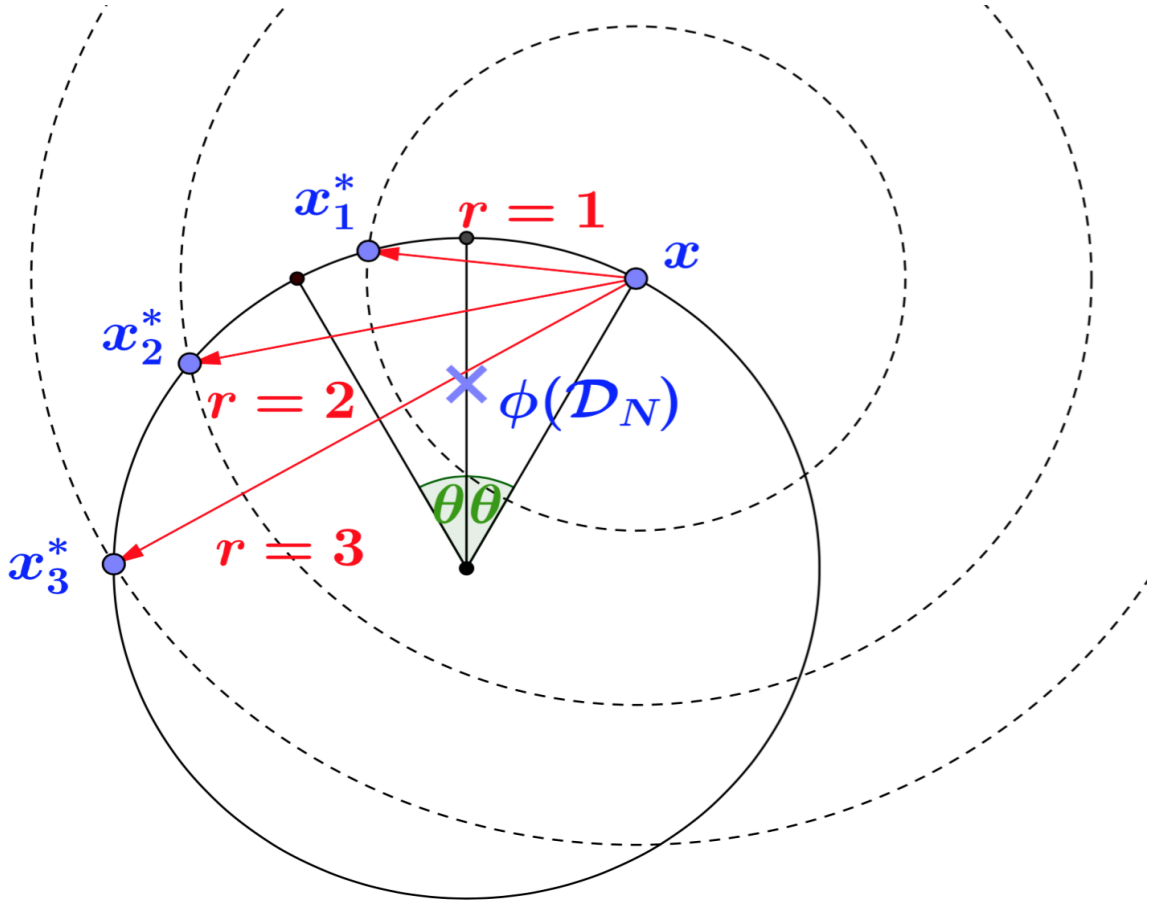


Figure: Illustration du Lemme 2 avec r prenant des valeurs entières (représentant les distances de Kendall possibles).

Plan

Applicabilité de la méthode

Nous notons:

- n le nombre d'objets
- $\mathcal{D}_N \in \mathfrak{S}_n^N$ le dataset
- r une règle de vote (quelconque), et $r(\mathcal{D}_N)$ le consensus sur $\mathcal{D}_N$ donné par r

On sait que:

$$d(r(\mathcal{D}_N), \mathcal{K}_N) \leq k_{min}\,.$$

On étudie la finesse de la borne avec:

$$s\,(r, \mathcal{D}_N, n) := k_{min} - d(r(\mathcal{D}_N), \mathcal{K}_N)\,.$$

Résultats

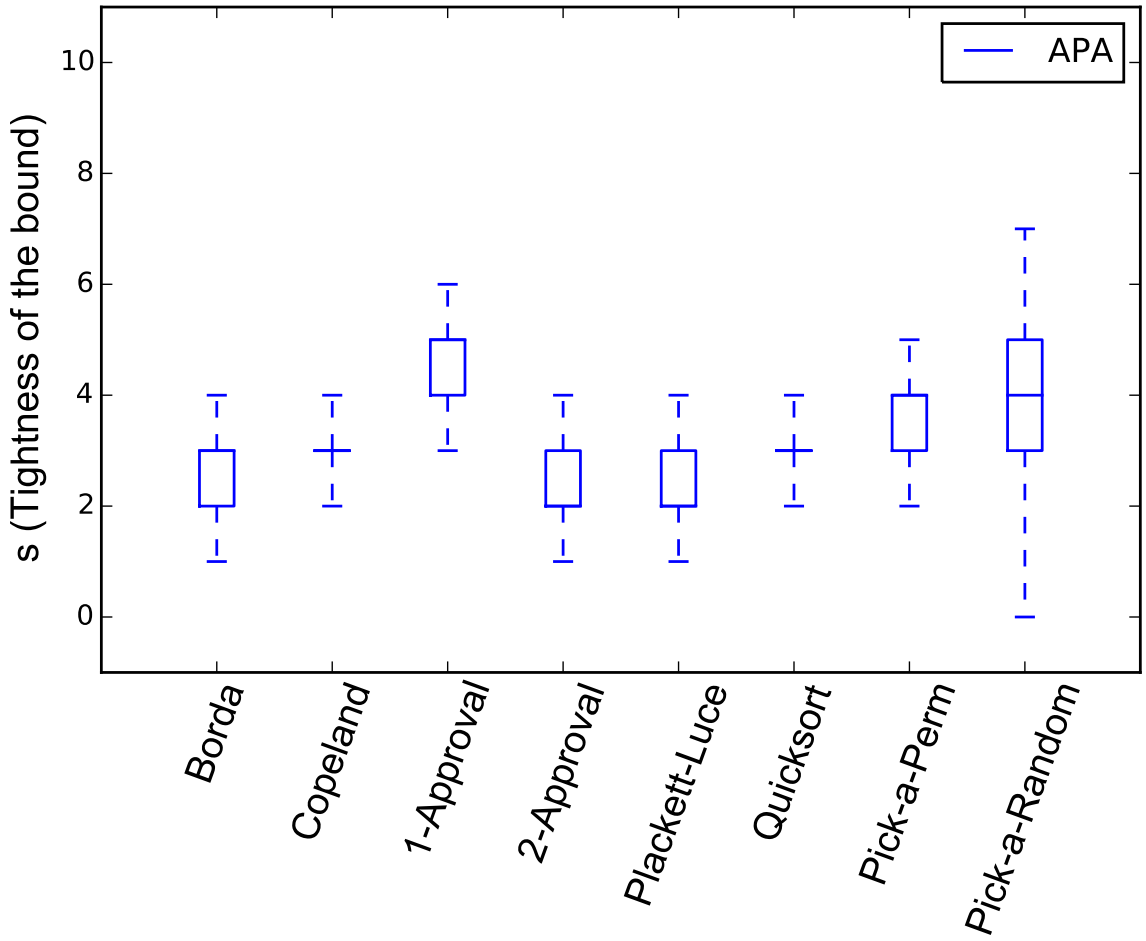


Figure: Boxplot de $s\left(r, \mathcal{D}_N, n\right)$, qui montre les effets selon la règle de vote r avec 500 échantillons bootstrappés du dataset APA ($n = 5, N = 5738$).

Prédictivité de la méthode

- Quand n grandit, caculer le(s) consensus de Kemeny σ^*, et donc $s\left(r, \mathcal{D}_N, n\right)$ devient vite impossible au niveau computationnel.

- Une fois que nous avons un consensus d'approximation $r(\mathcal{D}_N)$ et que k_{min} est identifié via notre méthode, le domaine de recherche pour le consensus de Kemeny peut être réduit aux permutations à une distance $\leq k_{min}$ de $r(\mathcal{D}_N)$.

- Le nombre de ces permutations dans $\mathfrak{S}_n$ est borné par $\binom{n+k_{min}-1}{k_{min}} << |\mathfrak{S}_n| = n!$ [Wang 2013].

Résultats

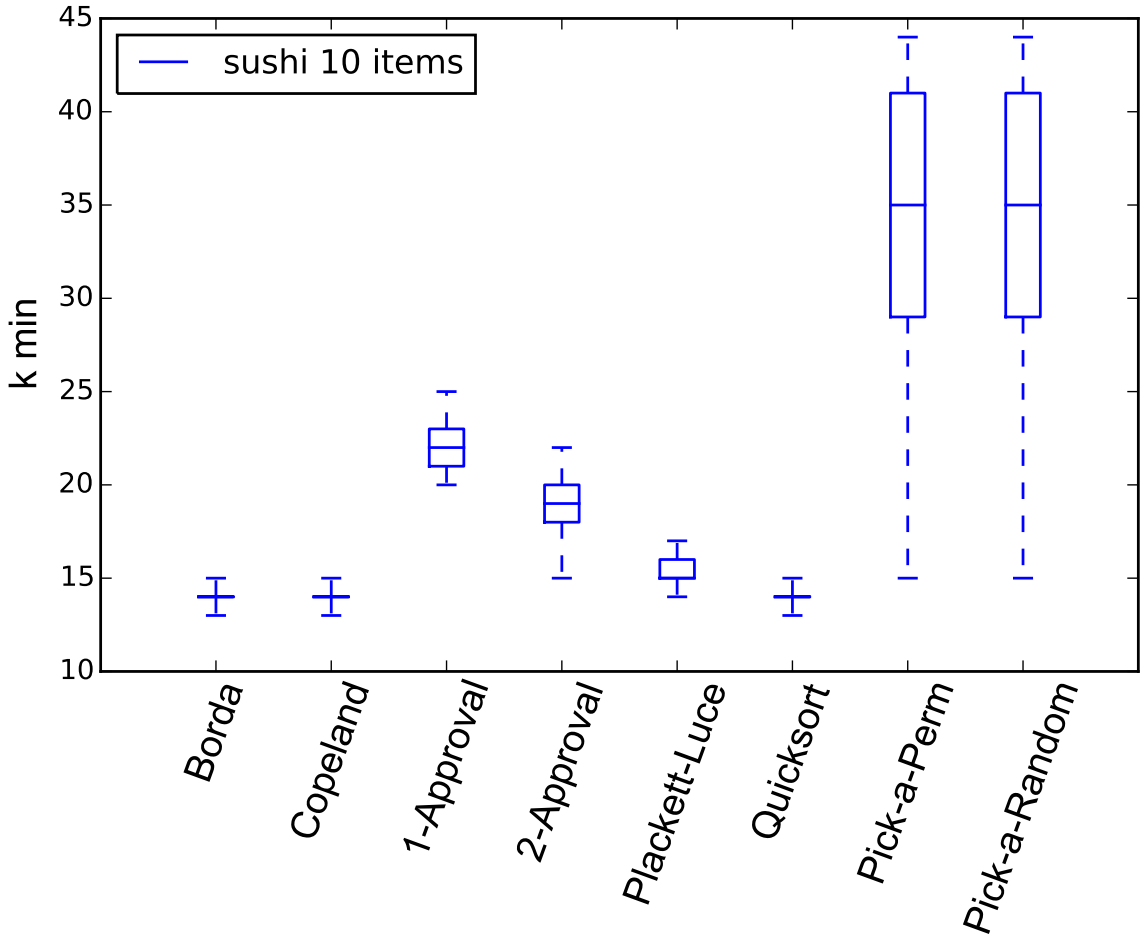


Figure: Boxplot de k_{min} sur 500 échantillons bootstrappés du dataset Sushi ($n = 10, N = 5000$).

Plan

Conclusion

- Nous avons établi un résultat théorique qui permet de contrôler la distance de Kendall entre une permutation et le consensus de Kemeny d'un dataset quelconque.

Conclusion

- Nous avons établi un résultat théorique qui permet de contrôler la distance de Kendall entre une permutation et le consensus de Kemeny d'un dataset quelconque.
- Nous présentons une méthode simple et générale, pour prédire, pour n'importe quelle procédure d'agrégation, à quel point son consensus sur le dataset est éloigné du consensus de Kemeny.

Directions futures

- Les propriétés géométriques de l'embedding de Kemeny sont riches et pourraient mener à de nombreux autres résultats.

Directions futures

- Les propriétés géométriques de l'embedding de Kemeny sont riches et pourraient mener à de nombreux autres résultats.
- On peut imaginer des procédures d'agrégation utilisant la diminution du domaine de recherche pour le consensus de Kemeny.

Directions futures

- Les propriétés géométriques de l'embedding de Kemeny sont riches et pourraient mener à de nombreux autres résultats.
- On peut imaginer des procédures d'agrégation utilisant la diminution du domaine de recherche pour le consensus de Kemeny.
- On peut également étendre notre résultat aux rankings incomplets.

Merci