

Une théorie statistique pour l'agrégation de classements

Journée de la chaire 2017

Anna Korba, Stéphan Cléménçon, Eric Sibony

May 5, 2017

Télécom ParisTech

Table des matières

1. L'agrégation de classements
2. Etude du risque théorique
3. Vitesses de convergence
4. Conclusion

L'agrégation de classements

L'agrégation de classements

Cadre

- n objets: $\{1, \dots, n\}$.
- N agents qui forment chacun un classement complet de ces objets, du premier au dernier objet dans l'ordre des préférences:
 $i_1 \succ i_2 \succ \dots \succ i_n$.

Exemple: $2 \succ 3 \succ 1 \succ 4$

L'agrégation de classements

Cadre

- n objets: $\{1, \dots, n\}$.
- N agents qui forment chacun un classement complet de ces objets, du premier au dernier objet dans l'ordre des préférences:
 $i_1 \succ i_2 \succ \dots \succ i_n$.

Exemple: $2 \succ 3 \succ 1 \succ 4$

Données

- $i_1 \succ \dots \succ i_n \iff$ permutation σ sur $\{1, \dots, n\}$ telle que $\sigma(i_j) = j$.
Exemple: $\sigma(2) = 1, \sigma(3) = 2 \dots$
- une collection de N classements est donc une collection de permutations $(\sigma_1, \dots, \sigma_N) \in \mathfrak{S}_n^N$.

L'agrégation de classements

Cadre

- n objets: $\{1, \dots, n\}$.
- N agents qui forment chacun un classement complet de ces objets, du premier au dernier objet dans l'ordre des préférences:
 $i_1 \succ i_2 \succ \dots \succ i_n$.

Exemple: $2 \succ 3 \succ 1 \succ 4$

Données

- $i_1 \succ \dots \succ i_n \iff$ permutation σ sur $\{1, \dots, n\}$ telle que $\sigma(i_j) = j$.
Exemple: $\sigma(2) = 1, \sigma(3) = 2 \dots$
- une collection de N classements est donc une collection de permutations $(\sigma_1, \dots, \sigma_N) \in \mathfrak{S}_n^N$.

Problème

On veut retrouver un ordre global ("consensus") σ^* sur les n objets.
 \implies Quelle permutation $\sigma^* \in \mathfrak{S}_n$ représente le mieux la collection de permutations $(\sigma_1, \dots, \sigma_N)$?

Exemple 1 - Elections

Soit un ensemble de candidats:  ,  ,  ,  . Chaque électeur peut donner par exemple:

- son candidat favori:



- un classement :



L'ensemble des votes récoltés pour l'élection constitue un **dataset de classements**.

Exemple 1 - Elections

Soit un ensemble de candidats:  ,  ,  ,  . Chaque électeur peut donner par exemple:

- son candidat favori:



- un classement :



L'ensemble des votes récoltés pour l'élection constitue un **dataset de classements**.

⇒ Comment élire le (les) vainqueur(s)?

Jean-Charles de Borda



Nicolas de Condorcet



Débat

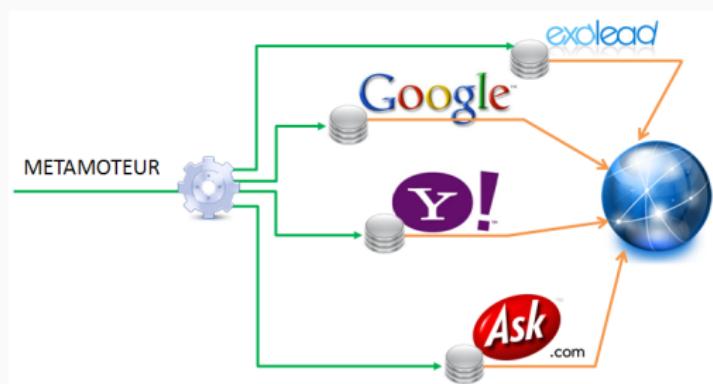
Borda-Condorcet au

18th siècle

Exemple 2: Métamoteurs

Comment dépasser le biais d'un moteur de recherche?

⇒ En agrégeant les listes ordonnées de plusieurs moteurs



Méthodes - Les règles de Borda et Copeland

Idée: calculer un score par objet et classer les objets selon leurs scores.

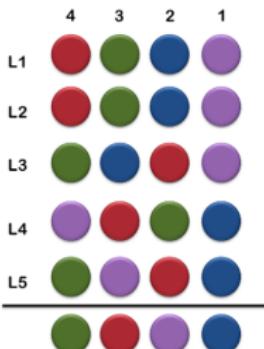
- score de Borda (méthode de vote pondéré):

$$s_B(i) = \frac{1}{N} \sum_{t=1}^N (n + 1 - \sigma_t(i))$$

- score de Copeland (méthode de Condorcet):

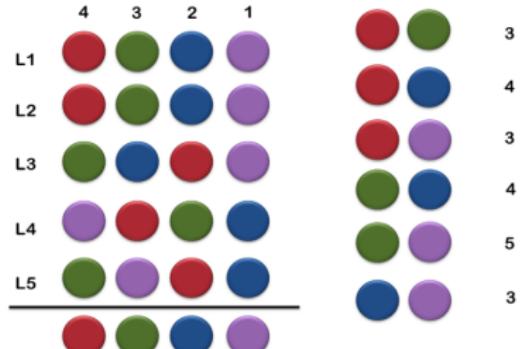
$$s_C(i) = \frac{1}{N} \sum_{t=1}^N \mathbb{I}[\sigma_t(i) < \sigma_t(j)]$$

Borda



$$\begin{aligned}4+4+2+3+2 &= 15 \\3+3+4+2+4 &= 16 \\2+2+3+1+1 &= 9 \\1+1+1+4+3 &= 10\end{aligned}$$

Copeland



Méthodes - La règle de Kemeny (1959)

Etant donné un dataset $(\sigma_1, \dots, \sigma_N) \in \mathfrak{S}_n^N$, une approche populaire pour agréger des classements est la règle de Kemeny, qui consiste à trouver la solution du problème suivant:

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

Méthodes - La règle de Kemeny (1959)

Etant donné un dataset $(\sigma_1, \dots, \sigma_N) \in \mathfrak{S}_n^N$, une approche populaire pour agréger des classements est la règle de Kemeny, qui consiste à trouver la solution du problème suivant:

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

où d est la distance de Kendall:

$$d(\sigma, \pi) = \sum_{\{i,j\} \subset [\![n]\!]} \{\sigma \text{ et } \pi \text{ ne concordent pas sur } \{i,j\}\}$$

Méthodes - La règle de Kemeny (1959)

Etant donné un dataset $(\sigma_1, \dots, \sigma_N) \in \mathfrak{S}_n^N$, une approche populaire pour agréger des classements est la règle de Kemeny, qui consiste à trouver la solution du problème suivant:

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

où d est la distance de Kendall:

$$d(\sigma, \pi) = \sum_{\{i,j\} \subset [\![n]\!]} \{\sigma \text{ et } \pi \text{ ne concordent pas sur } \{i,j\}\}$$

Exemple

$$\sigma = 1234 \quad (1 \succ 2 \succ 3 \succ 4)$$

$$\pi = 2413 \quad (2 \succ 4 \succ 1 \succ 3)$$

→ nombre de désaccords = sur 3 paires (12,14,34).

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].

Travaux - Partie 1 : Bornes sur les approximations

ICML 2016:

Etant donné un dataset $D_N = (\sigma_1, \dots, \sigma_N)$, σ le consensus d'une procédure d'approximation (exemple: Borda ou Copeland) et σ^* le consensus de Kemeny:

⇒ Nous donnons une borne supérieure sur la distance de Kendall entre σ et σ^* :
 $d(\sigma, \sigma^*) \leq$ (quantité calculée sur le dataset)

Travaux - Partie 2 : Reformulation statistique

Supposons que le dataset est composé de N copies i.i.d $\Sigma_1, \dots, \Sigma_N$ d'une variable aléatoire $\Sigma \sim P$. Une médiane de P selon d est une solution du problème suivant:

$$\min_{\sigma \in \mathfrak{S}_n} \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)],$$

où $L(\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)]$ est **le risque** de σ .

Travaux - Partie 2 : Reformulation statistique

Supposons que le dataset est composé de N copies i.i.d $\Sigma_1, \dots, \Sigma_N$ d'une variable aléatoire $\Sigma \sim P$. Une médiane de P selon d est une solution du problème suivant:

$$\min_{\sigma \in \mathfrak{S}_n} \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)],$$

où $L(\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)]$ est le **risque** de σ .

But de notre étude ?

Etudier la performance de **médianes de Kemeny empiriques**, i.e. solutions $\widehat{\sigma}_N$ de:

$$\min_{\sigma \in \mathfrak{S}_n} \widehat{L}_N(\sigma),$$

où $\widehat{L}_N(\sigma) = \frac{1}{N} \sum_{t=1}^N d(\Sigma_t, \sigma)$.

⇒ Lien avec Borda et Copeland

Etude du risque théorique

Rappel classification binaire - 1

On observe (X_i, Y_i) qui sont des réalisations i.i.d. d'une variable aléatoire (X, Y) , où X est une quantité aléatoire à valeurs dans un espace mesurable (\mathcal{X}, A) et Y est un label, i.e., une variable aléatoire binaire sur $\llbracket 0, 1 \rrbracket$.

On définit la fonction de régression par:

$$\eta(x) = \mathbb{E}[Y|X = x] = \mathbb{P}[Y = 1|X = x]$$

Rappel classification binaire - 1

On observe (X_i, Y_i) qui sont des réalisations i.i.d. d'une variable aléatoire (X, Y) , où X est une quantité aléatoire à valeurs dans un espace mesurable $(\mathcal{X}, \mathcal{A})$ et Y est un label, i.e., une variable aléatoire binaire sur $\llbracket 0, 1 \rrbracket$.

On définit la fonction de régression par:

$$\eta(x) = \mathbb{E}[Y|X = x] = \mathbb{P}[Y = 1|X = x]$$

Soit $h : \mathcal{X} \Rightarrow \llbracket 0, 1 \rrbracket$ un classifieur, i.e un prédicteur qui observant X , prédit le label $h(X)$. La performance du classifieur h est mesurée par l'erreur de classification :

$$R(h) = \mathbb{P}[Y \neq h(X)].$$

Rappel classification binaire - 2

Ce risque s'écrit:

$$R(h) = \mathbb{P}[Y \neq h(X)] = \int \mathbb{P}[Y \neq h(x)|X = x] P_X(dx).$$

avec:

$$\mathbb{P}[Y \neq h(x)|X = x] = \eta(x) \mathbb{I}[h(x) = 0] + (1-\eta(x)) \mathbb{I}[h(x) = 1]$$

Rappel classification binaire - 2

Ce risque s'écrit:

$$R(h) = \mathbb{P}[Y \neq h(X)] = \int \mathbb{P}[Y \neq h(x)|X = x] P_X(dx).$$

avec:

$$\mathbb{P}[Y \neq h(x)|X = x] = \eta(x) \mathbb{I}[h(x) = 0] + (1-\eta(x)) \mathbb{I}[h(x) = 1]$$

Il existe un classifieur qui atteint l'erreur minimale: $R^* = \min_h R(h)$.
C'est le classifieur de Bayes h^* défini par:

$$h^*(x) = \begin{cases} 1 & \text{si } \eta(x) > 1/2 \\ 0 & \text{si } \eta(x) \leq 1 \end{cases}$$

Interprétation de h^*

Donc si la probabilité du label 1 sachant l'entrée x est supérieure à 1/2, le classifieur de Bayes prédit 1.

Risque de l'agrégation de classement

Le risque d'une médiane σ est:

$$L(\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)]$$

où d est la distance de Kendall:

$$d(\sigma, \pi) = \sum_{\{i,j\} \subset [\![n]\!]} \{(\sigma(i) - \sigma(j))(\pi(i) - \pi(j)) < 0\}$$

Soit $p_{i,j} = \mathbb{P}[\Sigma(i) < \Sigma(j)]$ la probabilité que l'objet i soit préféré à j .

Risque de l'agrégation de classement

Le risque d'une médiane σ est:

$$L(\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)]$$

où d est la distance de Kendall:

$$d(\sigma, \pi) = \sum_{\{i,j\} \subset [\![n]\!]} \{(\sigma(i) - \sigma(j))(\pi(i) - \pi(j)) < 0\}$$

Soit $p_{i,j} = \mathbb{P}[\Sigma(i) < \Sigma(j)]$ la probabilité que l'objet i soit préféré à j .

On peut réécrire le risque:

$$L(\sigma) = \sum_{i < j} p_{i,j} \mathbb{I}\{\sigma(i) > \sigma(j)\} + \sum_{i < j} (1 - p_{i,j}) \mathbb{I}\{\sigma(i) < \sigma(j)\}.$$

Risque de l'agrégation de classement

Le risque d'une médiane σ est:

$$L(\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)]$$

où d est la distance de Kendall:

$$d(\sigma, \pi) = \sum_{\{i,j\} \subset [\![n]\!]} \{(\sigma(i) - \sigma(j))(\pi(i) - \pi(j)) < 0\}$$

Soit $p_{i,j} = \mathbb{P}[\Sigma(i) < \Sigma(j)]$ la probabilité que l'objet i soit préféré à j .

On peut réécrire le risque:

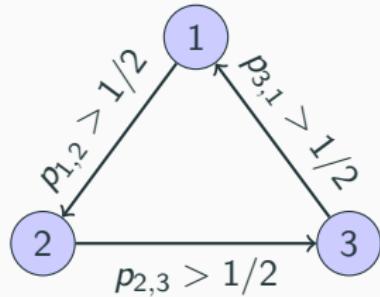
$$L(\sigma) = \sum_{i < j} p_{i,j} \mathbb{I}\{\sigma(i) > \sigma(j)\} + \sum_{i < j} (1 - p_{i,j}) \mathbb{I}\{\sigma(i) < \sigma(j)\}.$$

Donc s'il existe une permutation σ qui vérifie: $\forall i < j$ tel que $p_{i,j} \neq 1/2$,

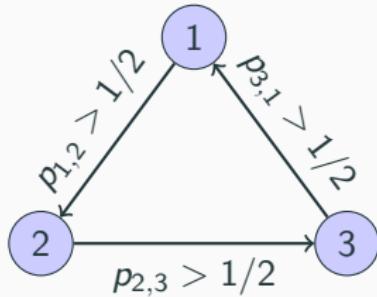
$$(\sigma(j) - \sigma(i)) \cdot (p_{i,j} - 1/2) > 0, \tag{2}$$

elle sera nécessairement une médiane pour P .

Cycles de préférences

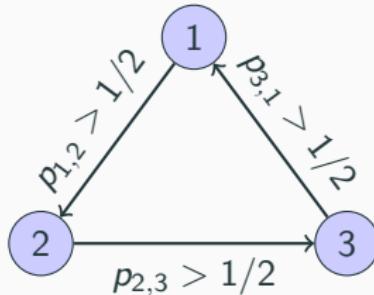


Cycles de préférences



→ Aucune permutation/aucun consensus ne peut satisfaire la condition pour chaque paire d'objets!

Cycles de préférences



→ Aucune permutation/aucun consensus ne peut satisfaire la condition pour chaque paire d'objets!

Définition - Transitivité stochastique

P sur \mathfrak{S}_n est stochastiquement transitive si : $\forall (i,j,k) \in [\![n]\!]^3$,

$$p_{i,j} \geq 1/2 \text{ and } p_{j,k} \geq 1/2 \Rightarrow p_{i,k} \geq 1/2.$$

De plus, si $p_{i,j} \neq 1/2$ pour tout $i < j$, P est strictement stochastiquement transitive.

Optimalité

Théorème

Si P est stochastiquement transitive:

- il existe $\sigma^* \in \mathfrak{S}_n$ telle que (2) est vraie
- l'ensemble des médianes de P est la classe d'équivalence de σ^* selon la relation d'équivalence:

$$\sigma \mathcal{R}_P \sigma' \Leftrightarrow (\sigma(j) - \sigma(i))(\sigma'(j) - \sigma'(i)) > 0 \text{ pour tout } i < j \text{ t.q. } p_{i,j} \neq 1/2.$$

Optimalité

Théorème

Si P est stochastiquement transitive:

- il existe $\sigma^* \in \mathfrak{S}_n$ telle que (2) est vraie
- l'ensemble des médianes de P est la classe d'équivalence de σ^* selon la relation d'équivalence:

$$\sigma \mathcal{R}_P \sigma' \Leftrightarrow (\sigma(j) - \sigma(i))(\sigma'(j) - \sigma'(i)) > 0 \text{ pour tout } i < j \text{ t.q. } p_{i,j} \neq 1/2.$$

De plus, le mapping s^* (**Copeland score**) qui pour chaque objet i vaut:

$$s^*(i) = 1 + \sum_{k \neq i} \mathbb{I}\{p_{i,k} < \frac{1}{2}\}$$

appartient à \mathfrak{S}_n et est l'unique médiane de P ssi P est strictement stochastiquement transitive.

Vitesses de convergence

Vitesses de convergence

Soit $\widehat{L}_N(\sigma) = \frac{1}{N} \sum_{t=1}^N d(\Sigma_t, \sigma)$ et $L(\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)]$.

Nous voulons comparer la perte/le risque de:

- la (une) solution empirique: $\widehat{\sigma}_N = \operatorname{argmin}_{\sigma \in \mathfrak{S}_n} \widehat{L}_N(\sigma)$
- la (une) solution optimale: $\sigma^* = \operatorname{argmin}_{\sigma \in \mathfrak{S}_n} L(\sigma)$

Donc $\mathbb{E}[L(\widehat{\sigma}_N) - L^*]$. Dans le cas général (pas d'hypothèses sur la distribution P), nous obtenons...

Vitesses de convergence

Soit $\widehat{L}_N(\sigma) = \frac{1}{N} \sum_{t=1}^N d(\Sigma_t, \sigma)$ et $L(\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)]$.

Nous voulons comparer la perte/le risque de:

- la (une) solution empirique: $\widehat{\sigma}_N = \operatorname{argmin}_{\sigma \in \mathfrak{S}_n} \widehat{L}_N(\sigma)$
- la (une) solution optimale: $\sigma^* = \operatorname{argmin}_{\sigma \in \mathfrak{S}_n} L(\sigma)$

Donc $\mathbb{E}[L(\widehat{\sigma}_N) - L^*]$. Dans le cas général (pas d'hypothèses sur la distribution P), nous obtenons...

Des vitesses de convergence classiques:

$\mathbb{E}[L(\widehat{\sigma}_N) - L^*]$ décroît à vitesse $\frac{1}{\sqrt{n}}$.

Vitesses de convergence rapides

Supposons que P vérifie:

- la transitivité stochastique:

$$p_{i,j} \geq 1/2 \text{ and } p_{j,k} \geq 1/2 \Rightarrow p_{i,k} \geq 1/2. \quad (3)$$

- la condition low-noise **NA(h)** pour un $h > 0$:

$$\min_{i < j} |p_{i,j} - 1/2| \geq h. \quad (4)$$

Vitesses de convergence rapides

Supposons que P vérifie:

- la **transitivité stochastique**:

$$p_{i,j} \geq 1/2 \text{ and } p_{j,k} \geq 1/2 \Rightarrow p_{i,k} \geq 1/2. \quad (3)$$

- la condition **low-noise NA(h)** pour un $h > 0$:

$$\min_{i < j} |p_{i,j} - 1/2| \geq h. \quad (4)$$

Vitesses rapides

Alors $\mathbb{E}[L(\hat{\sigma}_N) - L^*]$ décroît à vitesse exponentielle.

Vitesses de convergence rapides

Supposons que P vérifie:

- la **transitivité stochastique**:

$$p_{i,j} \geq 1/2 \text{ and } p_{j,k} \geq 1/2 \Rightarrow p_{i,k} \geq 1/2. \quad (3)$$

- la condition **low-noise NA(h)** pour un $h > 0$:

$$\min_{i < j} |p_{i,j} - 1/2| \geq h. \quad (4)$$

Vitesses rapides

Alors $\mathbb{E}[L(\widehat{\sigma}_N) - L^*]$ décroît à vitesse exponentielle.

Copeland score

Avec très grande probabilité, le score de **Copeland** empirique défini pour chaque objet $1 \leq i \leq n$ par:

$$\widehat{s}_N(i) = 1 + \sum_{k \neq i} \mathbb{I}\{\widehat{p}_{i,k} < \frac{1}{2}\}$$

appartient à \mathfrak{S}_n et est la solution de la règle de Kemeny empirique.

Conclusion

Conclusion

- Cadre statistique général pour l'agrégation de classements:
description des conditions d'optimalité , garanties théoriques pour la
généralisation d'un consensus de Kemeny empirique.

Conclusion

- Cadre statistique général pour l'agrégation de classements: description des conditions d'optimalité , garanties théoriques pour la généralisation d'un consensus de Kemeny empirique.
- Notre cadre est valable avec des classements complets sur les objets mais aussi avec des comparaisons par paires (plus courantes dans les applications modernes)

Conclusion

- Cadre statistique général pour l'agrégation de classements:
description des conditions d'optimalité , garanties théoriques pour la généralisation d'un consensus de Kemeny empirique.
- Notre cadre est valable avec des classements complets sur les objets mais aussi avec des comparaisons par paires (plus courantes dans les applications modernes)
- Analogie avec les vitesses rapides obtenues en classification
(Koltchinskii et Beznosova en 2005)

Conclusion

- Cadre statistique général pour l'agrégation de classements: description des conditions d'optimalité , garanties théoriques pour la généralisation d'un consensus de Kemeny empirique.
- Notre cadre est valable avec des classements complets sur les objets mais aussi avec des comparaisons par paires (plus courantes dans les applications modernes)
- Analogie avec les vitesses rapides obtenues en classification (Koltchinskii et Beznosova en 2005)
- Conséquence pratique du dernier résultat: sous la condition low-noise, la méthode de Copeland ($\mathcal{O}(N^{\binom{n}{2}})$) donne le consensus de Kemeny (NP-difficile) avec très grande probabilité.

Merci!