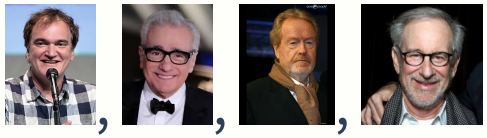
What is ranking data?

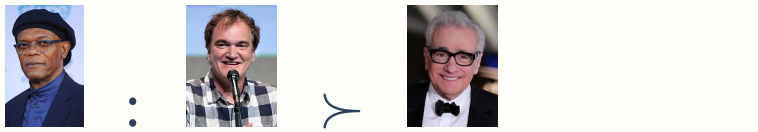
Consider a set of items $[\![n]\!] := \{1, \ldots, n\}$.

A ranking is an **ordered list** (of any size) **of items** of $[\![n]\!]$

Example:

$[\![4]\!] := \{1, 2, 3, 4\}$=  .

Ask an actor to rank/order them by preference ($\succ$):

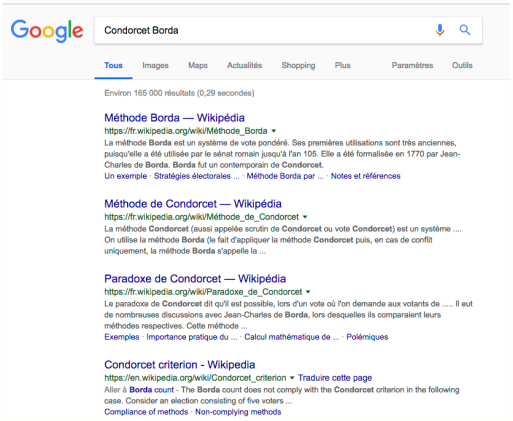


Many applications involve rankings/comparisons

- Modelling human preferences (elections, surveys, online implicit feedback)

$\Longrightarrow$ easier to rank than to rate

- Computer systems (search engines, recommendation systems)



- Others (competitions, biological data...)

Analysis of full rankings

Set of items $[\![n]\!] := \{1, \dots, n\}$. Ex: $\{1, 2, 3, 4\}$

- An individual expresses her preferences as a **full** ranking, i.e a strict order $\succ$ over the whole set $[\![n]\!]$:

$$a_1 \succ a_2 \succ \cdots \succ a_n$$

Other kind of rankings: **Top-k rankings**: $a_1, \dots, a_k \succ$ the rest, **Pairwise comparisons**: $a_1 \succ a_2$

Analysis of full rankings

Set of items $[\![n]\!] := \{1, \dots, n\}$. Ex: $\{1, 2, 3, 4\}$

- An individual expresses her preferences as a **full** ranking, i.e a strict order $\succ$ over the whole set $[\![n]\!]$:

$$a_1 \succ a_2 \succ \cdots \succ a_n$$

Other kind of rankings: **Top-k rankings**: $a_1, \dots, a_k \succ$ the rest, **Pairwise comparisons**: $a_1 \succ a_2$

A full ranking can be seen as the permutation σ that maps an item to its rank:

$$a_1 \succ \cdots \succ a_n \quad \Leftrightarrow \quad \sigma \in \mathfrak{S}_n \text{ such that } \sigma(a_i) = i$$

$$2 \succ 1 \succ 3 \succ 4 \quad \Leftrightarrow \quad \sigma = 2134 \quad (\sigma(2) = 1, \sigma(1) = 2, \dots)$$

Analysis of full rankings

Set of items $[\![n]\!] := \{1, \dots, n\}$. Ex: $\{1, 2, 3, 4\}$

- An individual expresses her preferences as a **full** ranking, i.e a strict order $\succ$ over the whole set $[\![n]\!]$:

$$a_1 \succ a_2 \succ \cdots \succ a_n$$

Other kind of rankings: **Top-k rankings**: $a_1, \dots, a_k \succ$ the rest, **Pairwise comparisons**: $a_1 \succ a_2$

A full ranking can be seen as the permutation σ that maps an item to its rank:

$$a_1 \succ \cdots \succ a_n \quad \Leftrightarrow \quad \sigma \in \mathfrak{S}_n \text{ such that } \sigma(a_i) = i$$

$$2 \succ 1 \succ 3 \succ 4 \quad \Leftrightarrow \quad \sigma = 2134 \quad (\sigma(2) = 1, \sigma(1) = 2, \dots)$$

Let $\mathfrak{S}_n$ be set of permutations of $[\![n]\!]$, the symmetric group.
Ex: $\mathfrak{S}_4 = 1234, 1324, 1423, \dots, 4321$

Common ranking problems

Consider N individuals expressing their preferences on $[\![n]\!]$:
$\Longrightarrow$ results in a dataset of N rankings/permutations

$$\mathcal{D}_N = (\sigma_1, \sigma_2, \ldots, \sigma_N) \in \mathfrak{S}_n^N$$

Common ranking problems

Consider N individuals expressing their preferences on $[\![n]\!]$:
$\Longrightarrow$ results in a dataset of N rankings/permutations

$$\mathcal{D}_N = (\sigma_1, \sigma_2, \ldots, \sigma_N) \in \mathfrak{S}_n^N$$

Considered here

- **Ranking aggregation:** Find a full ranking that "best summarizes" $\mathcal{D}_N$
  e.g. Ranking of a competition
- **Label ranking:** Predict a ranking given some information
  e.g. In a recommendation setting

Others

- **Top-1 recovery:** Find the "most preferred" item in $\mathcal{D}_N$
  e.g. Output of an election
- **Clustering:** Split $\mathcal{D}_N$ into clusters
  e.g. Segment customers based on their answers to a survey

Analysis of ranking data: main challenges

How to analyze a dataset of permutations $\mathcal{D}_N = (\sigma_1, \ldots, \sigma_N)$?

Analysis of ranking data: main challenges

How to analyze a dataset of permutations $\mathcal{D}_N = (\sigma_1, \ldots, \sigma_N)$?

- A random permutation $\Sigma \in \mathfrak{S}_n$ can be seen as a random vector $(\Sigma(1), \ldots, \Sigma(n)) \in \mathbb{R}^n$...

Analysis of ranking data: main challenges

How to analyze a dataset of permutations $\mathcal{D}_N = (\sigma_1, \ldots, \sigma_N)$?

- A random permutation $\Sigma \in \mathfrak{S}_n$ can be seen as a random vector $(\Sigma(1), \ldots, \Sigma(n)) \in \mathbb{R}^n$...

  but the random variables $\Sigma(1), \ldots, \Sigma(n)$ are highly dependent and the sum $\Sigma + \Sigma'$ is not a random permutation!

Analysis of ranking data: main challenges

How to analyze a dataset of permutations $\mathcal{D}_N = (\sigma_1, \ldots, \sigma_N)$?

- A random permutation $\Sigma \in \mathfrak{S}_n$ can be seen as a random vector $(\Sigma(1), \ldots, \Sigma(n)) \in \mathbb{R}^n$...

  but the random variables $\Sigma(1), \ldots, \Sigma(n)$ are highly dependent and the sum $\Sigma + \Sigma'$ is not a random permutation!

  $\Rightarrow$ No natural notion of mean and variance for Σ

Analysis of ranking data: main challenges

How to analyze a dataset of permutations $\mathcal{D}_N = (\sigma_1, \ldots, \sigma_N)$?

- A random permutation $\Sigma \in \mathfrak{S}_n$ can be seen as a random vector $(\Sigma(1), \ldots, \Sigma(n)) \in \mathbb{R}^n$...

  but the random variables $\Sigma(1), \ldots, \Sigma(n)$ are highly dependent and the sum $\Sigma + \Sigma'$ is not a random permutation!

  $\Rightarrow$ No natural notion of mean and variance for Σ

- The set of permutations $\mathfrak{S}_n$ is finite...

Analysis of ranking data: main challenges

How to analyze a dataset of permutations $\mathcal{D}_N = (\sigma_1, \ldots, \sigma_N)$?

- A random permutation $\Sigma \in \mathfrak{S}_n$ can be seen as a random vector $(\Sigma(1), \ldots, \Sigma(n)) \in \mathbb{R}^n$...

  but the random variables $\Sigma(1), \ldots, \Sigma(n)$ are highly dependent and the sum $\Sigma + \Sigma'$ is not a random permutation!

  $\Rightarrow$ No natural notion of mean and variance for Σ

- The set of permutations $\mathfrak{S}_n$ is finite...

  but it has exploding cardinality: $|\mathfrak{S}_n| = n!$

Analysis of ranking data: main challenges

How to analyze a dataset of permutations $\mathcal{D}_N = (\sigma_1, \ldots, \sigma_N)$?

- A random permutation $\Sigma \in \mathfrak{S}_n$ can be seen as a random vector $(\Sigma(1), \ldots, \Sigma(n)) \in \mathbb{R}^n$...

  but the random variables $\Sigma(1), \ldots, \Sigma(n)$ are highly dependent and the sum $\Sigma + \Sigma'$ is not a random permutation!

  $\Rightarrow$ No natural notion of mean and variance for Σ

- The set of permutations $\mathfrak{S}_n$ is finite...

  but it has exploding cardinality: $|\mathfrak{S}_n| = n!$

  $\Rightarrow$ Little statistical relevance

Main approaches 1 - Parametric

- Choose a predefined generative model on the data and analyze the data through that model
  ([Lu and Boutilier, 2014, Zhao et al., 2016, Szörényi et al., 2015])

Main approaches 1 - Parametric

- Choose a predefined generative model on the data and analyze the data through that model
  ([Lu and Boutilier, 2014, Zhao et al., 2016, Szörényi et al., 2015])
  - Mallows [Mallows, 1957]
    Parameterized by a central ranking $\sigma_0 \in \mathfrak{S}_n$ and a dispersion parameter $\gamma \in \mathbb{R}^+$

    $$P(\sigma) = Ce^{-\gamma d(\sigma_0, \sigma)} \qquad \text{with } d \text{ a distance on } \mathfrak{S}_n.$$

  - Plackett-Luce [Luce, 1959]
    Each item i is parameterized by w_i with $w_i \in \mathbb{R}^+$:

    $$P(\sigma) = \prod_{i=1}^{n} \frac{w_{\sigma^{-1}(i)}}{\sum_{j=i}^{n} w_{\sigma^{-1}(j)}}$$

    Ex: $2 \succ 1 \succ 3 = \frac{w_2}{w_1+w_2+w_3} \frac{w_1}{w_1+w_3}$

Main approaches 1 - Parametric

- Choose a predefined generative model on the data and analyze the data through that model
  ([Lu and Boutilier, 2014, Zhao et al., 2016, Szörényi et al., 2015])
  - Mallows [Mallows, 1957]
    Parameterized by a central ranking $\sigma_0 \in \mathfrak{S}_n$ and a dispersion parameter $\gamma \in \mathbb{R}^+$

    $$P(\sigma) = Ce^{-\gamma d(\sigma_0, \sigma)} \qquad \text{with } d \text{ a distance on } \mathfrak{S}_n.$$

  - Plackett-Luce [Luce, 1959]
    Each item i is parameterized by w_i with $w_i \in \mathbb{R}^+$:

    $$P(\sigma) = \prod_{i=1}^{n} \frac{w_{\sigma^{-1}(i)}}{\sum_{j=i}^{n} w_{\sigma^{-1}(j)}}$$

    Ex: $2 \succ 1 \succ 3 = \frac{w_2}{w_1+w_2+w_3} \frac{w_1}{w_1+w_3}$

- may fail to hold on real data (see for instance [Davidson and Marschak, 1959, Tversky, 1972] on decision making)

Main approaches 2 -“Non Parametric”

- Choose a structure on $\mathfrak{S}_n$ and analyze the data with respect to that structure
  - Harmonic analysis ([Kondor and Barbosa, 2010, Clémençon et al., 2011, Sibony et al., 2015]
  - Kernel density smoothing [Sun et al., 2012]
  - Modeling of pairwise comparisons ([Jiang et al., 2011, Rajkumar and Agarwal, 2014, Shah and Wainwright, 2017])
  - Kernel methods [Jiao and Vert, 2015]...

Presented contributions of the thesis

Two problems considered:

- Unsupervised learning: statistical framework and results for **ranking aggregation**
- Supervised learning: two family of methods for **label ranking**

Our approach (Non parametric)

- Introduce a rigorous statistical framework for each problem and establish theoretical guarantees
- Exploit the topology of the graph of pairwise preferences
  $2 \succ 1 \succ 3 \succ 4 \quad \Longrightarrow \quad 2 \succ 1, 2 \succ 3, 2 \succ 4, 1 \succ 3, 1 \succ 4, 3 \succ 4$
- Exploit the geometry of well-chosen feature maps for rankings
  $\sigma \quad \mapsto \quad \phi(\sigma) \in \mathbb{R}$

Outline

Ranking Aggregation

Consider a dataset of rankings/permutations:

$$\mathcal{D}_N = (\sigma_1, \dots, \sigma_N) \in \mathfrak{S}_n^N$$

We want to find a global order (*consensus*) on the n items:

$$\sigma^*_{\mathcal{D}_N} \in \mathfrak{S}_n$$

that **best represents** the dataset.

Ranking Aggregation

Consider a dataset of rankings/permutations:

$$\mathcal{D}_N = (\sigma_1, \ldots, \sigma_N) \in \mathfrak{S}_n^N$$

We want to find a global order (*consensus*) on the n items:

$$\sigma^*_{\mathcal{D}_N} \in \mathfrak{S}_n$$

that **best represents** the dataset.

- problem introduced in the study of elections systems in *social choice*
- modern applications (e.g. meta-search engines, label ranking)
- how to compute $\sigma^*_{\mathcal{D}_N}$?

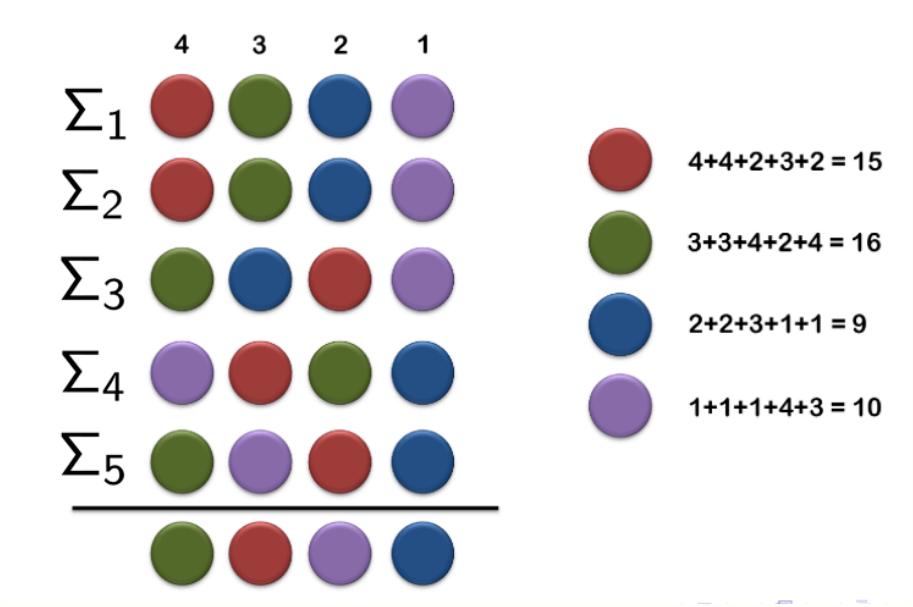
Scoring methods

Idea: compute a score for each item i and sort the items w.r.t. that score.

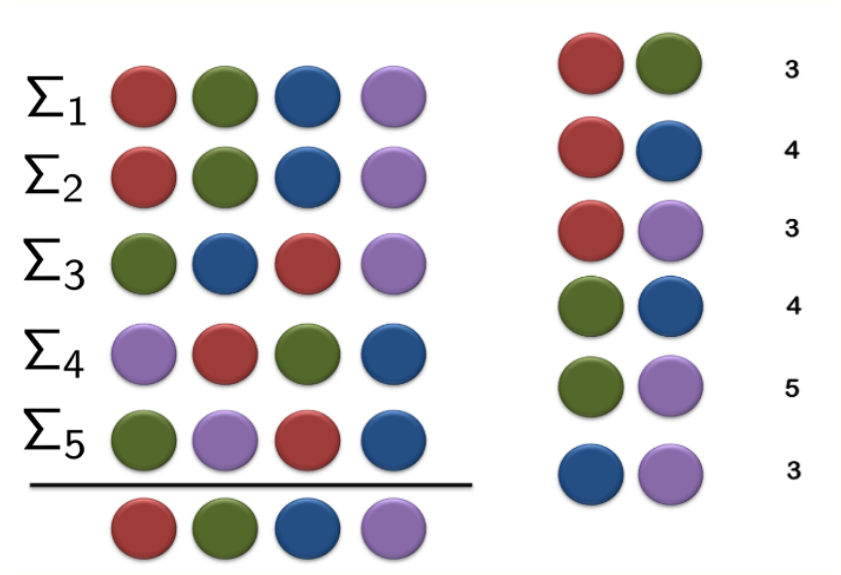
Scoring methods

Idea: compute a score for each item i and sort the items w.r.t. that score.

Borda [Borda, 1781]



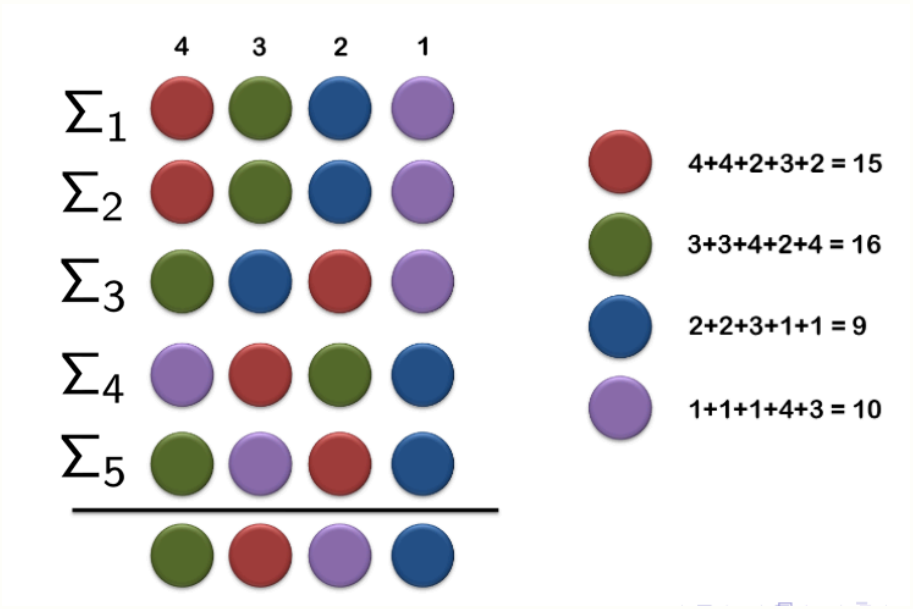
Copeland [Copeland, 1951]



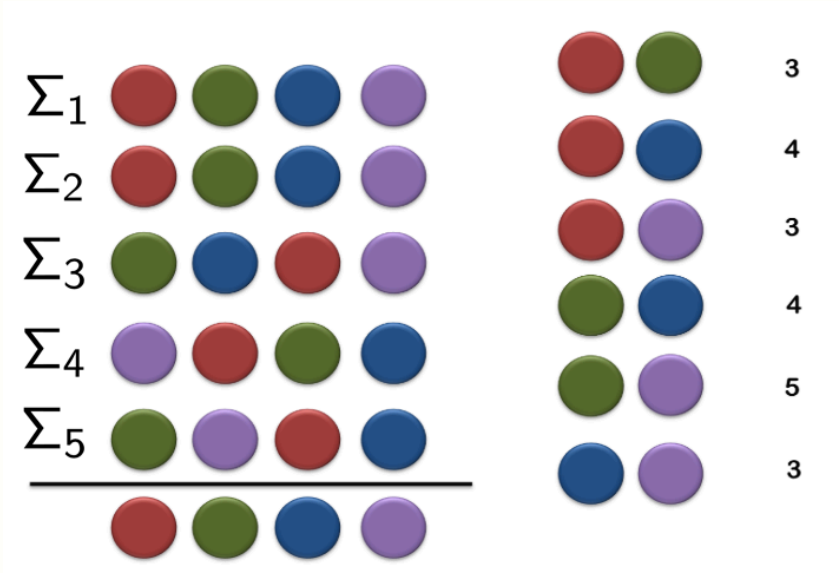
Scoring methods

Idea: compute a score for each item i and sort the items w.r.t. that score.

Borda [Borda, 1781]



Copeland [Copeland, 1951]

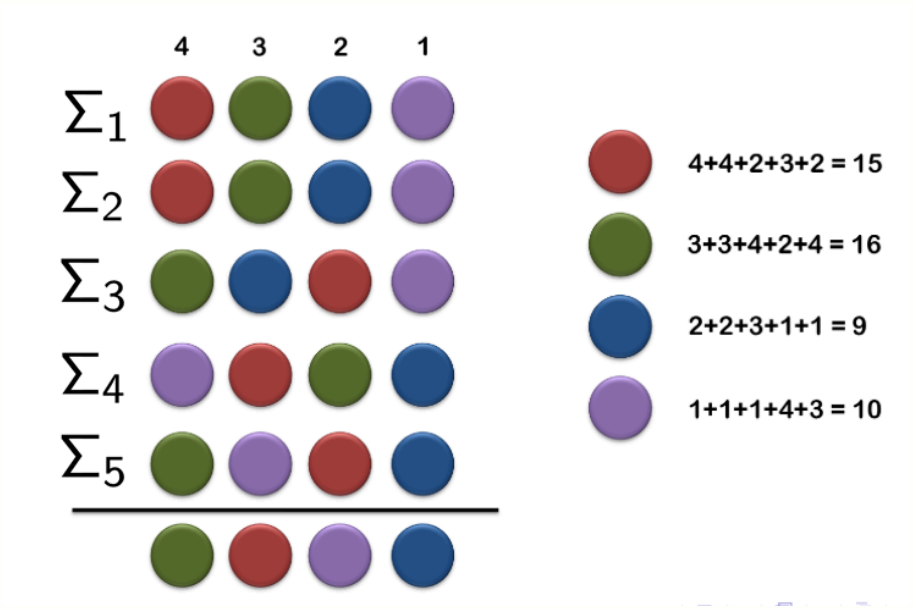


$$s_B(i) = \sum_{k=1}^{N} (n + 1 - \Sigma_k(i)) = \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{k=1}^{N} \mathbb{I}[\Sigma_k(i) < \Sigma_k(j)] + N$$

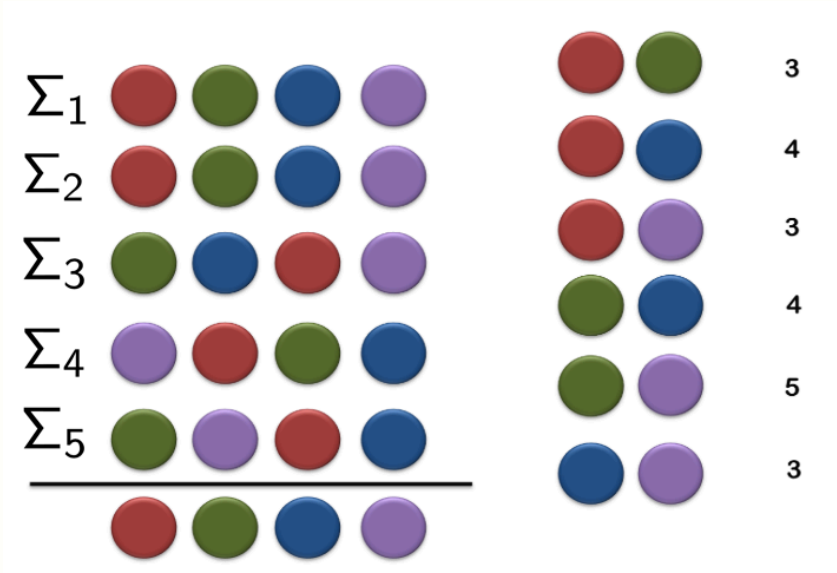
Scoring methods

Idea: compute a score for each item i and sort the items w.r.t. that score.

Borda [Borda, 1781]



Copeland [Copeland, 1951]



$$s_B(i) = \sum_{k=1}^{N} (n + 1 - \Sigma_k(i)) = \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{k=1}^{N} \mathbb{I}[\Sigma_k(i) < \Sigma_k(j)] + N$$

$$s_C(i) = \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathbb{I}\{ \sum_{k=1}^{N} \mathbb{I}[\Sigma_k(i) < \Sigma_k(j)] > \frac{N}{2} \}$$

Scoring methods

- Simple complexities (Borda $\mathcal{O}(nN)$, Copeland $\mathcal{O}(n^2N)$)
- But few guarantees on the output
  - does not necessarily output a full ranking (ties between scores)
  - **Borda** count does not satisfy the **Condorcet criterion**: if a candidate wins (in majority) against all the others in pairwise duels then it should be ranked first
  - **Copeland** ignores the relatives strengths of the pairwise comparisons (often many ties)

Scoring methods

- Simple complexities (Borda $\mathcal{O}(nN)$, Copeland $\mathcal{O}(n^2N)$)
- But few guarantees on the output
  - does not necessarily output a full ranking (ties between scores)
  - **Borda** count does not satisfy the **Condorcet criterion**: if a candidate wins (in majority) against all the others in pairwise duels then it should be ranked first
  - **Copeland** ignores the relatives strengths of the pairwise comparisons (often many ties)

$\Longrightarrow$ *Borda-Condorcet debate at the* 18^{th} *century*

Jean-Charles de Borda

Nicolas de Condorcet

Kemeny's rule [Kemeny, 1959]

Ranking aggregation as an optimization problem

$$\text{Solve} \quad \sigma^*_{\mathcal{D}_N} = \underset{\sigma \in \mathfrak{S}_n}{\text{argmin}} \sum_{k=1}^{N} d(\sigma, \Sigma_k)$$

Kemeny's rule [Kemeny, 1959]

Ranking aggregation as an optimization problem

$$\text{Solve} \quad \sigma^*_{\mathcal{D}_N} = \underset{\sigma \in \mathfrak{S}_n}{\text{argmin}} \sum_{k=1}^{N} d(\sigma, \Sigma_k)$$

where d_τ is the Kendall's τ distance, i.e. for $\sigma, \sigma' \in \mathfrak{S}_n$:

$$d_\tau(\sigma, \sigma') = \sum_{1 \leq i < j \leq n} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\},$$

Ex: σ= 1234, σ'= 2413 $\Rightarrow d_\tau(\sigma, \sigma') = 3$ (disagree on (12),(14),(34)).

Kemeny's rule [Kemeny, 1959]

Ranking aggregation as an optimization problem

$$\text{Solve} \quad \sigma^*_{\mathcal{D}_N} = \operatorname*{argmin}_{\sigma \in \mathfrak{S}_n} \sum_{k=1}^{N} d(\sigma, \Sigma_k)$$

where d_τ is the Kendall's τ distance, i.e. for $\sigma, \sigma' \in \mathfrak{S}_n$:

$$d_\tau(\sigma, \sigma') = \sum_{1 \leq i < j \leq n} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\},$$

Ex: σ= 1234, σ'= 2413 $\Rightarrow d_\tau(\sigma, \sigma') = 3$ (disagree on (12),(14),(34)).

- Solution always exists but may not be unique
- Natural loss d_τ when rankings represent preferences

Kemeny's rule properties

Social choice justification

- unique rule satisfying:
  - *Neutrality*: if i and j switch positions in $\Sigma_1, \dots \Sigma_N$, they should switch positions also in $\sigma^*_{\mathcal{D}_N}$
  - *Condorcet criterion*: any item which wins every other in pairwise simple majority voting should be ranked first
  - *Consistency*: if $\mathcal{D}_N$ is split is $\mathcal{D}^1_N$ and $\mathcal{D}^2_N$, and $i \succ j$ in $\sigma^*_{\mathcal{D}^1_N}$ and $\sigma^*_{\mathcal{D}^2_N} \Longrightarrow i \succ j$ in $\sigma^*_{\mathcal{D}_N}$

Statistical justification

Output the MLE estimator under the Mallows model:

$$\operatorname*{argmax}_{\sigma \in \mathfrak{S}_n} \prod_{k=1}^{N} \frac{e^{-\gamma d(\sigma, \Sigma_k)}}{Z} = \operatorname*{argmin}_{\sigma \in \mathfrak{S}_n} \sum_{k=1}^{N} d_\tau(\Sigma_k, \sigma)$$

Problem: **NP-hard** even for $N = 4$ ([Dwork et al., 2001]).

Statistical Ranking Aggregation [Korba et al., 2017]

Probabilistic Modeling

$$\mathcal{D}_N = (\Sigma_1, \ldots, \Sigma_N) \qquad \text{with} \qquad \Sigma_k \sim P$$

where P distribution on $\mathfrak{S}_n$.

Statistical Ranking Aggregation [Korba et al., 2017]

Probabilistic Modeling

$$\mathcal{D}_N = (\Sigma_1, \dots, \Sigma_N) \qquad \text{with} \qquad \Sigma_k \sim P$$

where P distribution on $\mathfrak{S}_n$.

Definition

A **Kemeny median** of P is solution of:

$$\sigma_P^* = \operatorname*{argmin}_{\sigma \in \mathfrak{S}_n} L_P(\sigma), \tag{1}$$

where $L_P(\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\sigma, \Sigma)]$ is **the risk** of σ.

$\Longrightarrow$ Statistical point of view $\neq$ from social choice: $\sigma^*_{\mathcal{D}_N}$ versus σ^*_P

Statistical Ranking Aggregation [Korba et al., 2017]

Probabilistic Modeling

$$\mathcal{D}_N = (\Sigma_1, \dots, \Sigma_N) \qquad \text{with} \qquad \Sigma_k \sim P$$

where P distribution on $\mathfrak{S}_n$.

Definition

A **Kemeny median** of P is solution of:

$$\sigma_P^* = \operatorname*{argmin}_{\sigma \in \mathfrak{S}_n} L_P(\sigma), \tag{1}$$

where $L_P(\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\sigma, \Sigma)]$ is **the risk** of σ.

$\Longrightarrow$ Statistical point of view $\neq$ from social choice: $\sigma^*_{\mathcal{D}_N}$ versus σ^*_P

Question: Can we exhibit some conditions on P so that solving (1) is tractable?

Conditions on the distribution

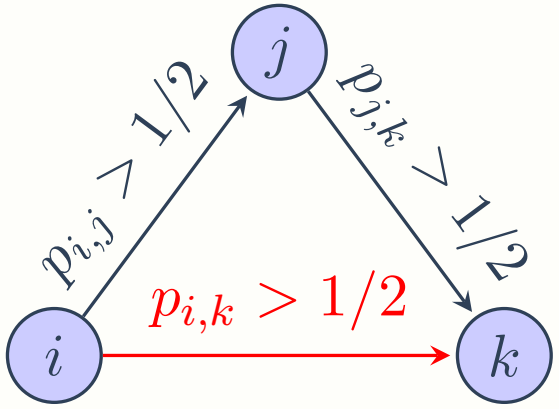
Let $p_{i,j} = \mathbb{P}[\Sigma(i) < \Sigma(j)]$ prob. that item $i \succ j$ (is preferred to).

Conditions on the distribution

Let $p_{i,j} = \mathbb{P}[\Sigma(i) < \Sigma(j)]$ prob. that item $i \succ j$ (is preferred to).

- Strict Stochastic Transitivity (**SST**): $(p_{i,j} \neq 1/2 \ \forall i, j)$

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

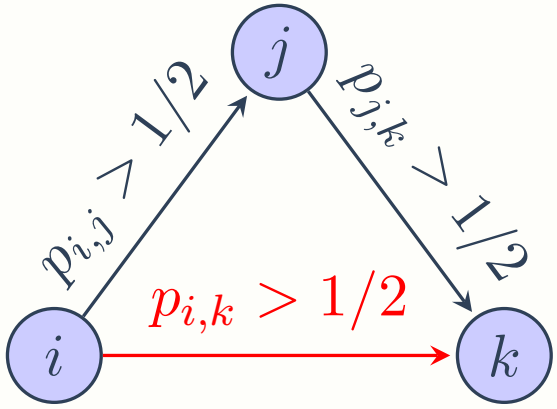


Conditions on the distribution

Let $p_{i,j} = \mathbb{P}[\Sigma(i) < \Sigma(j)]$ prob. that item $i \succ j$ (is preferred to).

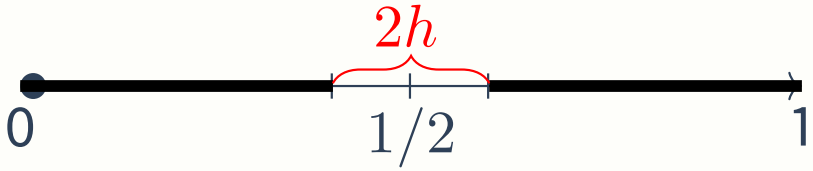
- Strict Stochastic Transitivity (**SST**): ($p_{i,j} \neq 1/2 \; \forall i, j$)

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



- Low-Noise/**NA**(h) for $h > 0$ ([Audibert and Tsybakov, 2007]):

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



Exact Solutions

Our result [Korba et al., 2017]

Suppose P satisfies **SST and NA**(h) for a given $h > 0$. Then with overwhelming probability $1 - \frac{n(n-1)}{4} e^{-\alpha_h N}$:

Exact Solutions

Our result [Korba et al., 2017]

Suppose P satisfies **SST and NA**(h) for a given $h > 0$. Then with overwhelming probability $1 - \frac{n(n-1)}{4} e^{-\alpha_h N}$: $\widehat{P}$ also verifies **SST**...

Exact Solutions

Our result [Korba et al., 2017]

Suppose P satisfies **SST and NA**(h) for a given $h > 0$. Then with overwhelming probability $1 - \frac{n(n-1)}{4} e^{-\alpha_h N}$: $\widehat{P}$ also verifies **SST**...and the Kemeny median of P is given by the empirical Copeland ranking:

$$\sigma_P^*(i) = 1 + \sum_{j \neq i} \mathbb{I}\{\widehat{p}_{i,j} < \frac{1}{2}\} \quad \text{for } 1 \leq i \leq n$$

$\alpha_h = \frac{1}{2} \log\left(1/(1-4h^2)\right)$

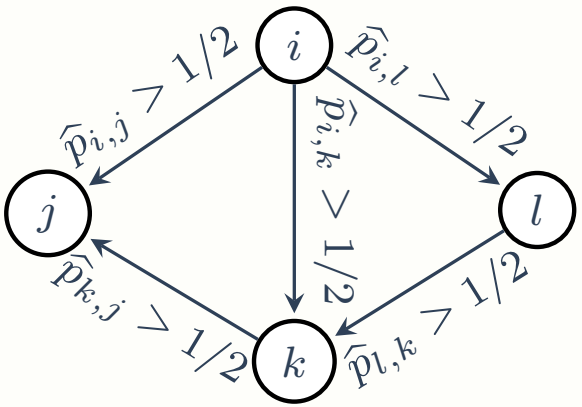
Exact Solutions

Our result [Korba et al., 2017]

Suppose P satisfies **SST and NA**(h) for a given $h > 0$. Then with overwhelming probability $1 - \frac{n(n-1)}{4} e^{-\alpha_h N}$: $\widehat{P}$ also verifies **SST**...and the Kemeny median of P is given by the empirical Copeland ranking:

$$\sigma^*_P(i) = 1 + \sum_{j \neq i} \mathbb{I}\{\widehat{p}_{i,j} < \frac{1}{2}\} \quad \text{for } 1 \leq i \leq n$$

$\alpha_h = \frac{1}{2} \log\left(1/(1-4h^2)\right)$



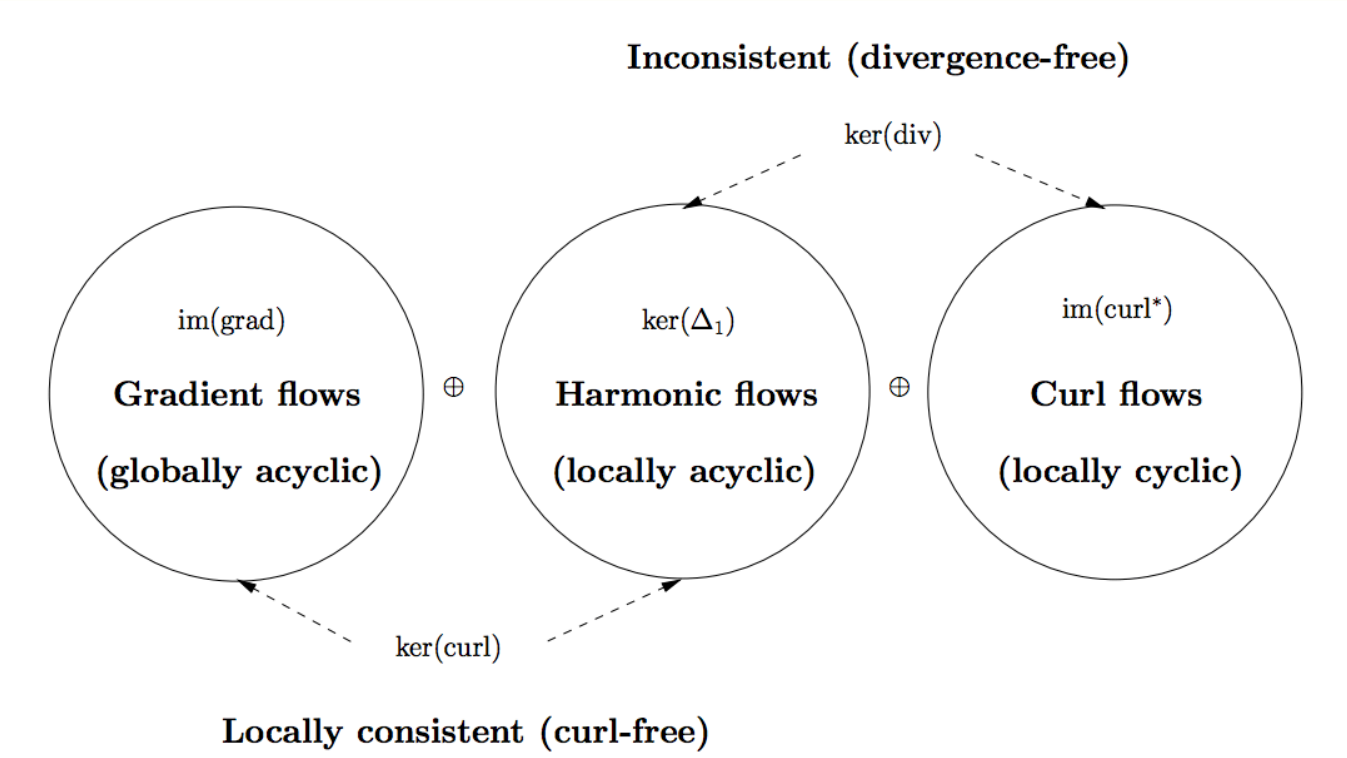
$\Rightarrow$ sort vertices by increasing input degree

Transitivity in practice

- If $\widehat{P}$ is SST, compute $\sigma^*_{\widehat{P}}$ with Copeland method based on the $\widehat{p}_{i,j}$'s

Transitivity in practice

- If $\widehat{P}$ is SST, compute $\sigma^*_{\widehat{P}}$ with Copeland method based on the $\widehat{p}_{i,j}$'s



Hodge decomposition of pairwise rankings

- Else, compute $\widetilde{\sigma}^*_{\widehat{P}}$ with empirical Borda count ([Jiang et al., 2011])

$$\widetilde{\sigma}^*_{\widehat{P}}(i) = \sigma^*_{proj_{im(grad)}(\widehat{P})}(i) = \frac{1}{N}\sum_{k=1}^{N}\Sigma_k(i) \quad \text{for } 1 \leq i \leq n$$

Outline

Label Ranking

Now $\mathcal{D}_N = (X_1, \Sigma_1), \ldots, (X_N, \Sigma_N)$ i.i.d. copies of $(X, \ \Sigma)$

Label Ranking

Now $\mathcal{D}_N = (X_1, \Sigma_1), \ldots, (X_N, \Sigma_N)$ i.i.d. copies of $(X,\ \Sigma)$

Ex: Users i with characteristics X_i and their produced rankings/preferences $\rightarrow \Sigma_i$.

Label Ranking

Now $\mathcal{D}_N = (X_1, \Sigma_1), \ldots, (X_N, \Sigma_N)$ i.i.d. copies of $(X,\ \Sigma)$
Ex: Users i with characteristics X_i and their produced rankings/preferences $\rightarrow \Sigma_i$.

- $X \sim \mu$, where μ is a distribution on some feature space $\mathcal{X}$
- $\Sigma \sim P_X$, where P_X is the conditional probability distribution (on $\mathfrak{S}_n$): $P_X(\sigma) = \mathbb{P}[\Sigma = \sigma | X]$

Label Ranking

Now $\mathcal{D}_N = (X_1, \Sigma_1), \ldots, (X_N, \Sigma_N)$ i.i.d. copies of $(X,\ \Sigma)$
Ex: Users i with characteristics X_i and their produced rankings/preferences $\rightarrow \Sigma_i$.

- $X \sim \mu$, where μ is a distribution on some feature space $\mathcal{X}$
- $\Sigma \sim P_X$, where P_X is the conditional probability distribution (on $\mathfrak{S}_n$): $P_X(\sigma) = \mathbb{P}[\Sigma = \sigma | X]$

Goal: Learn a predictive ranking rule :

$$\begin{array}{ccccc} s & : & \mathcal{X} & \to & \mathfrak{S}_n \\ & & x & \mapsto & s(x) \end{array}$$

which given a random vector X, predicts the permutation Σ on the n items.

Example: targeted advertising domain

Related Work

- Can be seen as an extension of multiclass and multilabel classification
- Many applications, e.g : document categorization, meta-learning
  - rank a set of topics relevant for a given document
  - rank a set of algorithms according to their suitability for a new dataset, based on the characteristics of the dataset
- A lot of approaches rely on parametric modelling [Cheng and Hüllermeier, 2009], [Cheng et al., 2010]

Related Work

- Can be seen as an extension of multiclass and multilabel classification
- Many applications, e.g : document categorization, meta-learning
  - rank a set of topics relevant for a given document
  - rank a set of algorithms according to their suitability for a new dataset, based on the characteristics of the dataset
- A lot of approaches rely on parametric modelling [Cheng and Hüllermeier, 2009], [Cheng et al., 2010]

$\Rightarrow$ We develop an approach free of any parametric assumptions (**local learning**) relying on results and framework developped for **ranking aggregation**.

Label Ranking and Ranking Aggregation

Performance: Measured by the risk:

$$\mathcal{R}(s) = \mathbb{E}_{X \sim \mu, \Sigma \sim P_X} \left[d_\tau \left(s(X), \Sigma \right) \right]$$

where d_τ is the Kendall's τ distance, i.e. for $\sigma, \sigma' \in \mathfrak{S}_n$:

$$d_\tau(\sigma, \sigma') = \sum_{1 \leq i < j \leq n} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\},$$

Label Ranking and Ranking Aggregation

Performance: Measured by the risk:

$$\mathcal{R}(s) = \mathbb{E}_{X \sim \mu, \Sigma \sim P_X} \left[d_\tau \left(s(X), \Sigma \right) \right]$$

where d_τ is the Kendall's τ distance, i.e. for $\sigma, \sigma' \in \mathfrak{S}_n$:

$$d_\tau(\sigma, \sigma') = \sum_{1 \leq i < j \leq n} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\},$$

Label ranking is an extension of ranking aggregation

$$\begin{aligned} \mathcal{R}(s) &= \mathbb{E}_{X \sim \mu, \Sigma \sim P_X} \left[d_\tau \left(s(X), \Sigma \right) \right] \\ &= \mathbb{E}_{X \sim \mu} \left[\mathbb{E}_{\Sigma \sim P_X} \left[d_\tau \left(s(X), \Sigma \right) \right] \right] \\ &= \mathbb{E}_{X \sim \mu} \left[L_{P_X}(s(X)) \right] \end{aligned}$$

Recall: $\sigma^*_{P_X} = \text{argmin}_{\sigma \in \mathfrak{S}_n} L_{P_X}(\sigma)$

Optimal Elements and Relaxation

Assumption

For $X \in \mathcal{X}$, P_X is **SST**: $\Rightarrow \sigma^*_{P_X} = \text{argmin}_{\sigma \in \mathfrak{S}_n} L_{P_X}(\sigma)$ is **unique**.

Optimal Elements and Relaxation

Assumption

For $X \in \mathcal{X}$, P_X is **SST**: $\Rightarrow \sigma^*_{P_X} = \text{argmin}_{\sigma \in \mathfrak{S}_n} L_{P_X}(\sigma)$ is **unique**.

Optimal elements

The best label ranking rules s^* (minimizing $\mathcal{R}(s)$) are the ones that maps any point $X \in \mathcal{X}$ to the **conditional** Kemeny median of P_X:

$$s^* = \underset{s:\mathcal{X} \to \mathfrak{S}_n}{\text{argmin}}\, \mathcal{R}(s) \quad \Leftrightarrow \quad s^*(X) = \sigma^*_{P_X}$$

Optimal Elements and Relaxation

Assumption

For $X \in \mathcal{X}$, P_X is **SST**: $\Rightarrow \sigma^*_{P_X} = \text{argmin}_{\sigma \in \mathfrak{S}_n} L_{P_X}(\sigma)$ is **unique**.

Optimal elements

The best label ranking rules s^* (minimizing $\mathcal{R}(s)$) are the ones that maps any point $X \in \mathcal{X}$ to the **conditional** Kemeny median of P_X:

$$s^* = \underset{s:\mathcal{X} \to \mathfrak{S}_n}{\text{argmin}}\, \mathcal{R}(s) \quad \Leftrightarrow \quad s^*(X) = \sigma^*_{P_X}$$

To minimize the risk $\mathcal{R}(s)$ approximately:

$$\sigma^*_{P_X} \text{ for any } X \xrightarrow{\text{is relaxed to}} s(X) = \sigma^*_{P_{\mathcal{C}}} \text{ for any } X \in \mathcal{C}$$

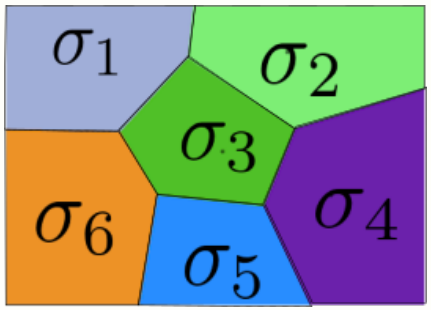
where $P_{\mathcal{C}}(\sigma) = \mathbb{P}[\Sigma = \sigma | X \in \mathcal{C}]$.

$\Rightarrow$ We develop Piecewise constant rules/ Local consensus methods.

Piecewise Constant Ranking Rules

Our approach: build a *piecewise constant* ranking rule $s_{\mathcal{P}} \in \mathcal{S}_{\mathcal{P}}$, ie: constant on each cell of a partition $\mathcal{P} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$ of the feature space $\mathcal{X}$.

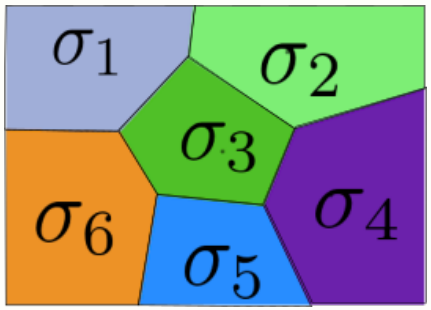
$\Longrightarrow s_{\mathcal{P}}(x) = \sigma_k \ (\sigma^*_{P_{\mathcal{C}_k}})$ iff $x \in \mathcal{C}_k$ for $k = 1, \ldots, K$.



$(\mathcal{X} = \mathbb{R}^2)$

Piecewise Constant Ranking Rules

Our approach: build a *piecewise constant* ranking rule $s_{\mathcal{P}} \in \mathcal{S}_{\mathcal{P}}$, ie: constant on each cell of a partition $\mathcal{P} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$ of the feature space $\mathcal{X}$.
$\Longrightarrow s_{\mathcal{P}}(x) = \sigma_k \ (\sigma^*_{P_{\mathcal{C}_k}})$ iff $x \in \mathcal{C}_k$ for $k = 1, \ldots, K$.



$(\mathcal{X} = \mathbb{R}^2)$

Approximation Error [Clémençon et al., 2018]

Suppose that $\exists M < \infty$ such that:

$\forall (x, x') \in \mathcal{X}^2, \ \sum_{i<j} |p_{i,j}(x) - p_{i,j}(x')| \leq M \cdot ||x - x'||$.

$$\Longrightarrow \inf_{s \in \mathcal{S}_{\mathcal{P}}} \mathcal{R}(s) - \mathcal{R}(s^*) \leq M.\delta_{\mathcal{P}}$$

where $\delta_{\mathcal{P}}$ is the max. diameter of $\mathcal{P}$'s cells.

Statistical Framework- ERM

Input: training data $\mathcal{D}_N = (X_1, \Sigma_1), \ldots, (X_N, \Sigma_N)$

Output: A partition $\mathcal{P}_N = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$ of $\mathcal{X}$ built from $\mathcal{D}_N$.

ERM: Optimize a statistical version of the theoretical risk based on the training data (X_k, Σ_k)'s:

$$\min_{s \in \mathcal{S}_\mathcal{P}} \widehat{\mathcal{R}}_N(s) = \frac{1}{N} \sum_{k=1}^{N} d_\tau(s(X_k),\ \Sigma_k)$$

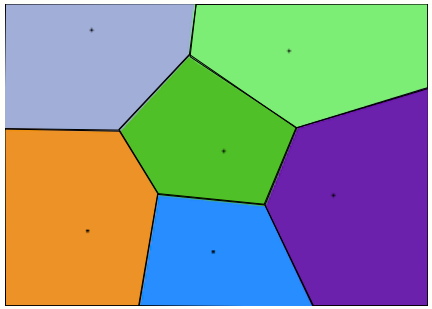
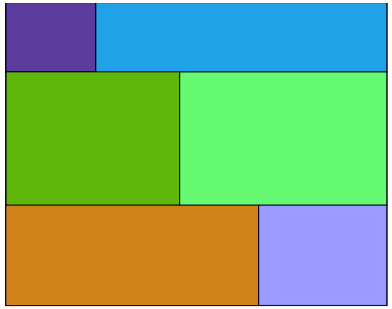
Rates of convergence [Clémençon et al., 2018]

- classical rates $\mathcal{O}(1/\sqrt{N})$ for ERM.
- fast rates $\mathcal{O}(1/N)$ under a "uniform" Low-Noise **NA**(h):

$$\inf_{x \in \mathcal{X}} \min_{i<j} |p_{i,j}(x) - 1/2| \geq h.$$

Partitioning Methods

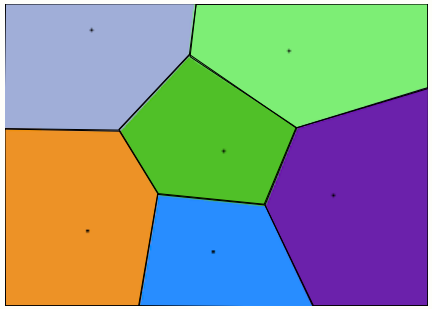
Two methods are investigated:

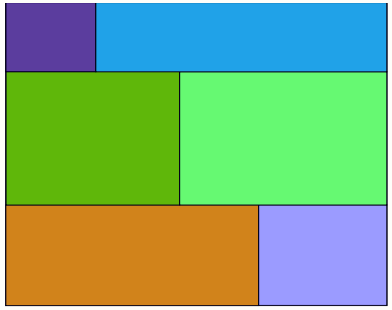
K-nearest neighbors (Voronoi partitioning)	*Decision tree* (Recursive partition)
	

Partitioning Methods

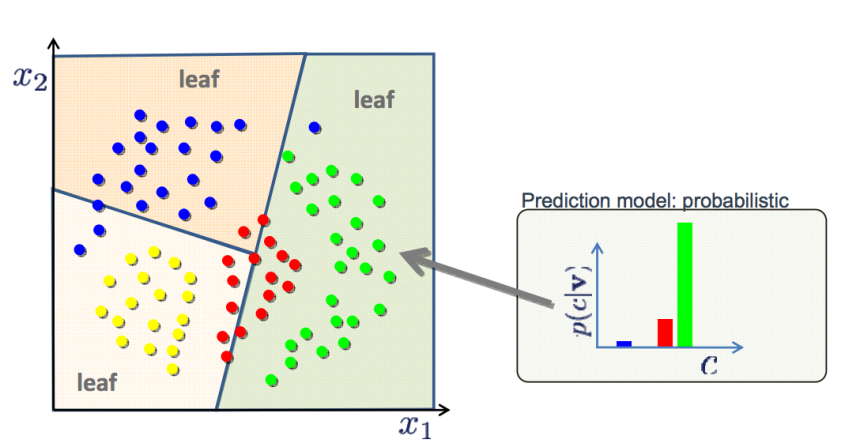
Two methods are investigated:

K-nearest neighbors (Voronoi partitioning)	*Decision tree* (Recursive partition)





For **classification**, the label of a cell (ex: a leaf) is the **majority** label among the training data which fall in this cell.



Compute Local Labels/Medians

Problem: Our training data are *permutations* Σ:

For a cell $\mathcal{C}$, if $\Sigma_1, \ldots, \Sigma_N \in \mathcal{C}$, how do we aggregate them into a final label σ^*?

$\Longrightarrow$ Ranking aggregation problem.

Compute Local Labels/Medians

Problem: Our training data are *permutations* Σ:

For a cell $\mathcal{C}$, if $\Sigma_1, \ldots, \Sigma_N \in \mathcal{C}$, how do we aggregate them into a final label σ^*?

$\Longrightarrow$ Ranking aggregation problem.

For $\mathcal{C} \in \mathcal{P}_N$, consider its empirical distribution:

$$\widehat{P}_{\mathcal{C}} = \frac{1}{N_{\mathcal{C}}} \sum_{k : X_k \in \mathcal{C}} \delta_{\Sigma_k}$$

and solve:

$$\min_{\sigma \in \mathfrak{S}_n} L_{\widehat{P}_{\mathcal{C}}}(\sigma)$$

$\Longrightarrow$ compute locally its Empirical Kemeny median $\widetilde{\sigma}^*_{\widehat{P}_{\mathcal{C}}}$ with the Copeland method) if $\widehat{P}_{\mathcal{C}}$ is **SST** or Borda otherwise.

K-Nearest Neigbors Algorithm

1. Fix $k \in \{1, \ldots, N\}$ and a query point $x \in \mathcal{X}$
2. Sort $(X_1, \Sigma_1), \ldots, (X_N, \Sigma_N)$ by increasing order of the distance to x : $\|X_{(1,N)} - x\| \leq \ldots \leq \|X_{(N,N)} - x\|$
3. Consider next the empirical distribution calculated using the k training points closest to x

$$\widehat{P}(x) = \frac{1}{k} \sum_{l=1}^{k} \delta_{\Sigma_{(l,N)}}$$

and compute the pseudo-empirical Kemeny median, yielding the k-NN prediction at x:

$$s_{k,N}(x) \overset{def}{=} \widetilde{\sigma}^*_{\widehat{P}(x)}.$$

$\Rightarrow$ We recover the classical bound $\mathcal{R}(s_{k,N}) - \mathcal{R}^* = \mathcal{O}(\frac{1}{\sqrt{k}} + \frac{k}{N})$

Decision Tree

Split recursively the feature space $\mathcal{X}$ by minimizing some impurity criterion.

Decision Tree

Split recursively the feature space $\mathcal{X}$ by minimizing some impurity criterion.

Gini criterion in multiclassification: m classes, f_i proportion of class $i \rightarrow I_G(\mathcal{C}) = \sum_{i=1}^{m} f_i(\mathcal{C})(1 - f_i(\mathcal{C}))$

Decision Tree

Split recursively the feature space $\mathcal{X}$ by minimizing some impurity criterion.

Gini criterion in multiclassification: m classes, f_i proportion of class $i \to I_G(\mathcal{C}) = \sum_{i=1}^{m} f_i(\mathcal{C})(1 - f_i(\mathcal{C}))$

Here, for a cell $\mathcal{C} \in \mathcal{P}_N$:

- Impurity [Alvo and Yu, 2014]:

$$\gamma_{\widehat{P}_{\mathcal{C}}} = \frac{1}{2} \sum_{1 \leq i < j \leq n} \widehat{p}_{i,j}(\mathcal{C}) \left(1 - \widehat{p}_{i,j}(\mathcal{C})\right)$$

  (ordering n elements can be seen as $\binom{n}{2}$ classification tasks) which is tractable and satisfies the double inequality

$$\widehat{\gamma}_{\widehat{P}_{\mathcal{C}}} \leq \min_{\sigma \in \mathfrak{S}_n} L_{\widehat{P}_{\mathcal{C}}}(\sigma) \leq 2\widehat{\gamma}_{\widehat{P}_{\mathcal{C}}}.$$

- Label of a cell/leaf : Compute the pseudo-empirical median of a cell $\mathcal{C}$:

$$s_{\mathcal{C}}(x) \overset{def}{=} \widetilde{\sigma}^*_{\widehat{P}_{\mathcal{C}}}.$$

Simulated Data

- We generate two explanatory variables, varying their nature (numerical, categorical) $\Rightarrow$ Setting 1/2/3
- We generate a partition of the feature space
- On each cell of the partition, a dataset of full rankings is generated, varying the distribution (constant, Mallows with $\neq$ dispersion): $D_0/D_1/D_2$

D_i	Setting 1			Setting 2			Setting 3		
	n=3	n=5	n=8	n=3	n=5	n=8	n=3	n=5	n=8
D_0	0.0698*	0.1290*	0.2670*	0.0173*	0.0405*	0.110*	0.0112*	0.0372*	0.0862*
	0.0473**	0.136**	0.324**	0.0568**	0.145**	0.2695**	0.099**	0.1331**	0.2188**
	(0.578)	(1.147)	(2.347)	(0.596)	(1.475)	(3.223)	(0.5012)	(1.104)	(2.332)
D_1	0.3475 *	0.569*	0.9405 *	0.306*	0.494*	0.784*	0.289*	0.457*	0.668*
	0.307**	0.529**	0.921**	0.308**	0.536**	0.862**	0.3374**	0.5714**	0.8544**
	(0.719)	(1.349)	(2.606)	(0.727)	(1.634)	(3.424)	(0.5254)	(1.138)	(2.287)
D_2	0.8656*	1.522*	2.503*	0.8305 *	1.447 *	2.359*	0.8105*	1.437*	2.189*
	0.7228**	1.322**	2.226**	0.723**	1.3305**	2.163**	0.7312**	1.3237**	2.252**
	(0.981)	(1.865)	(3.443)	(1.014)	(2.0945)	(4.086)	(0.8504)	(1.709)	(3.005)

Table: Empirical risk averaged on 50 trials on simulated data.

(): Clustering +PL, *: K-NN, **: Decision Tree