A Lecture on Statistical Ranking

Stéphan Clémençon

LTCI Telecom ParisTech, Paris Saclay
-
Institut Telecom

Bipartite Ranking

- (X, Y) random pair, valued in $\mathbb{R}^d \times \{-1, +1\}$ with $d >> 1$

Bipartite Ranking

- (X, Y) random pair, valued in $\mathbb{R}^d \times \{-1, +1\}$ with $d >> 1$
- **Observation:** sample $\mathcal{D}_n$ of i.i.d. copies of (X, Y)

$$(X_1, Y_1), \ \ldots, \ (X_n, Y_n)$$

Bipartite Ranking

- (X, Y) random pair, valued in $\mathbb{R}^d \times \{-1, +1\}$ with $d >> 1$
- **Observation:** sample $\mathcal{D}_n$ of i.i.d. copies of (X, Y)

$$(X_1, Y_1),\ \ldots,\ (X_n, Y_n)$$

- **Goal:** from labeled data $\mathcal{D}_n$, learn to **order** new data $X'_1,\ \ldots,\ X'_{n'}$

$$X'_7 \quad X'_{n'-2} \quad X'_3 \quad X'_6 \quad \ldots$$

Bipartite Ranking

- (X, Y) random pair, valued in $\mathbb{R}^d \times \{-1, +1\}$ with $d >> 1$
- **Observation:** sample $\mathcal{D}_n$ of i.i.d. copies of (X, Y)

$$(X_1, Y_1), \ \ldots, \ (X_n, Y_n)$$

- **Goal:** from labeled data $\mathcal{D}_n$, learn to **order** new data $X'_1, \ \ldots, \ X'_{n'}$

$$\begin{array}{ccccc} X'_7 & X'_{n'-2} & X'_3 & X'_6 & \ldots \\ + & + & - & + & \ldots \end{array}$$

in order to recover **positive instances on top of the list** with large probability

Goal of Bipartite Ranking

- Exactly the same setup as **binary classification**...

Goal of Bipartite Ranking

- Exactly the same setup as **binary classification**...
- ... except the nature of the problem is **global**!

Goal of Bipartite Ranking

- Exactly the same setup as **binary classification**...
- ... except the nature of the problem is **global**!
- **Applications:** credit-scoring, medical diagnosis, anomaly detection, information retrieval, *etc.*

Goal of Bipartite Ranking

- Exactly the same setup as **binary classification**...
- ... except the nature of the problem is **global**!
- **Applications:** credit-scoring, medical diagnosis, anomaly detection, information retrieval, *etc.*
- Our **agenda** for today:
  - "Ranking": a wide variety of problems motivated by numerous applications
  - Supervised ranking in its simplest form: bipartite ranking
  - ROC curves: a functional criterion for ranking performance

Goal of Bipartite Ranking

- Exactly the same setup as **binary classification**...
- ... except the nature of the problem is **global**!
- **Applications:** credit-scoring, medical diagnosis, anomaly detection, information retrieval, *etc.*
- Our **agenda** for today:
  - "Ranking": a wide variety of problems motivated by numerous applications
  - Supervised ranking in its simplest form: bipartite ranking
  - ROC curves: a functional criterion for ranking performance
  - Statistical learning theory and approximation theory

Goal of Bipartite Ranking

- Exactly the same setup as **binary classification**...
- ... except the nature of the problem is **global**!
- **Applications:** credit-scoring, medical diagnosis, anomaly detection, information retrieval, *etc.*
- Our **agenda** for today:
  - "Ranking": a wide variety of problems motivated by numerous applications
  - Supervised ranking in its simplest form: bipartite ranking
  - ROC curves: a functional criterion for ranking performance
  - Statistical learning theory and approximation theory
  - Ranking trees: ROC optimization through recursive partitioning

Goal of Bipartite Ranking

- Exactly the same setup as **binary classification**...
- ... except the nature of the problem is **global**!
- **Applications:** credit-scoring, medical diagnosis, anomaly detection, information retrieval, *etc.*
- Our **agenda** for today:
  - "Ranking": a wide variety of problems motivated by numerous applications
  - Supervised ranking in its simplest form: bipartite ranking
  - ROC curves: a functional criterion for ranking performance
  - Statistical learning theory and approximation theory
  - Ranking trees: ROC optimization through recursive partitioning
  - Limitations due to the global nature of the ranking problem

Goal of Bipartite Ranking

- Exactly the same setup as **binary classification**...
- ... except the nature of the problem is **global**!
- **Applications:** credit-scoring, medical diagnosis, anomaly detection, information retrieval, *etc.*
- Our **agenda** for today:
  - "Ranking": a wide variety of problems motivated by numerous applications
  - Supervised ranking in its simplest form: bipartite ranking
  - ROC curves: a functional criterion for ranking performance
  - Statistical learning theory and approximation theory
  - Ranking trees: ROC optimization through recursive partitioning
  - Limitations due to the global nature of the ranking problem
  - Aggregation in the context of Ranking? 'Ordinal' *vs.* 'metric-based'

Goal of Bipartite Ranking

- Exactly the same setup as **binary classification**...
- ... except the nature of the problem is **global**!
- **Applications:** credit-scoring, medical diagnosis, anomaly detection, information retrieval, *etc.*
- Our **agenda** for today:
    - "Ranking": a wide variety of problems motivated by numerous applications
    - Supervised ranking in its simplest form: bipartite ranking
    - ROC curves: a functional criterion for ranking performance
    - Statistical learning theory and approximation theory
    - Ranking trees: ROC optimization through recursive partitioning
    - Limitations due to the global nature of the ranking problem
    - Aggregation in the context of Ranking? 'Ordinal' *vs.* 'metric-based'
    - A computationally feasible consensus: median ranking trees
    - Ranking Forest: resampling + median computation
    - Extensions: multi-partite ranking

Bipartite setup -binary classification

- (X, Y) random pair with unknown distribution

Bipartite setup -binary classification

- (X, Y) random pair with unknown distribution
- $X \in \mathcal{X}$ observation with dist. $\mu(dx)$ and $Y \in \{-1, +1\}$ binary label

Bipartite setup -binary classification

- (X, Y) random pair with unknown distribution
- $X \in \mathcal{X}$ observation with dist. $\mu(dx)$ and $Y \in \{-1, +1\}$ binary label
- *A posteriori* probability $\sim$ **regression function**

$$\forall x \in \mathcal{X}, \quad \eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$$

Bipartite setup -binary classification

- (X, Y) random pair with unknown distribution
- $X \in \mathcal{X}$ observation with dist. $\mu(dx)$ and $Y \in \{-1, +1\}$ binary label
- *A posteriori* probability $\sim$ **regression function**

$$\forall x \in \mathcal{X}, \quad \eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$$

- $g : \mathcal{X} \to \{-1, +1\}$ prediction rule - **classifier**

Bipartite setup -binary classification

- (X, Y) random pair with unknown distribution
- $X \in \mathcal{X}$ observation with dist. $\mu(dx)$ and $Y \in \{-1, +1\}$ binary label
- *A posteriori* probability $\sim$ **regression function**

$$\forall x \in \mathcal{X}, \quad \eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$$

- $g : \mathcal{X} \to \{-1, +1\}$ prediction rule - **classifier**
- Performance measure = **classification error**

$$L(g) = \mathbb{P}\left\{g(X) \neq Y\right\} \quad \to \min_{g} L(g)$$

Bipartite setup -binary classification

- (X, Y) random pair with unknown distribution
- $X \in \mathcal{X}$ observation with dist. $\mu(dx)$ and $Y \in \{-1, +1\}$ binary label
- *A posteriori* probability $\sim$ **regression function**

$$\forall x \in \mathcal{X}, \quad \eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$$

- $g : \mathcal{X} \to \{-1, +1\}$ prediction rule - **classifier**
- Performance measure = **classification error**

$$L(g) = \mathbb{P}\{g(X) \neq Y\} \quad \to \min_g L(g)$$

- Solution: **Bayes classifier** $g^*(x) = 2\mathbb{I}\{\eta(x) > 1/2\} - 1$
- **Bayes error** $L^* = L(g^*) = 1/2 - \mathbb{E}[|2\eta(X) - 1|]/2$

Empirical Risk Minimization - Basics

- Sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ with i.i.d. copies of (X, Y)
- Class $\mathcal{G}$ of classifiers of a given **complexity**

Empirical Risk Minimization - Basics

- Sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ with i.i.d. copies of (X, Y)
- Class $\mathcal{G}$ of classifiers of a given **complexity**
- **Empirical Risk Minimization principle**

$$\hat{g}_n = \arg\min_{g \in \mathcal{G}} L_n(g)$$

with $L_n(g) \stackrel{def}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{g(X_i) \neq Y_i\}$

Empirical Risk Minimization - Basics

- Sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ with i.i.d. copies of (X, Y)
- Class $\mathcal{G}$ of classifiers of a given **complexity**
- **Empirical Risk Minimization principle**

$$\hat{g}_n = \arg\min_{g \in \mathcal{G}} L_n(g)$$

with $L_n(g) \stackrel{def}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{g(X_i) \neq Y_i\}$

- Mimic the best classifier among the class

$$\bar{g} = \arg\min_{g \in \mathcal{G}} L(g)$$

- **Bias-variance decomposition**

$$
\begin{aligned}
L(\hat{g}_n) - L^* &\leq (L(\hat{g}_n) - L_n(\hat{g}_n)) + (L_n(\bar{g}) - L(\bar{g})) + (L(\bar{g}) - L^*) \\
&\leq 2\left(\sup_{g \in \mathcal{G}} \mid L_n(g) - L(g) \mid\right) + \left(\inf_{g \in \mathcal{G}} L(g) - L^*\right)
\end{aligned}
$$

Empirical processes in classification

- **Bias-variance decomposition**

$$
\begin{aligned}
L(\hat{g}_n) - L^* &\leq (L(\hat{g}_n) - L_n(\hat{g}_n)) + (L_n(\bar{g}) - L(\bar{g})) + (L(\bar{g}) - L^*) \\
&\leq 2\left(\sup_{g \in \mathcal{G}} \mid L_n(g) - L(g) \mid\right) + \left(\inf_{g \in \mathcal{G}} L(g) - L^*\right)
\end{aligned}
$$

- **Concentration results**

  With probability $1 - \delta$:

$$
\sup_{g \in \mathcal{G}} \mid L_n(g) - L(g) \mid \leq \mathbb{E} \sup_{g \in \mathcal{G}} \mid L_n(g) - L(g) \mid + \sqrt{\frac{2 \log(1/\delta)}{n}}
$$

Main results in classification theory

1. Bayes risk consistency and rate of convergence

   Complexity control:

$$\mathbb{E} \sup_{g \in \mathcal{G}} \mid L_n(g) - L(g) \mid \leq C \sqrt{\frac{V}{n}}$$

   if $\mathcal{G}$ is a VC class with VC dimension V.

2. Fast rates of convergence

   Under variance control: rate faster than $n^{-1/2}$

3. Convex risk minimization

4. Oracle inequalities - Model selection

Main results in classification theory

1. Bayes risk consistency and rate of convergence

   Complexity control:

$$\mathbb{E} \sup_{g \in \mathcal{G}} \mid L_n(g) - L(g) \mid \leq C \sqrt{\frac{V}{n}}$$

   if $\mathcal{G}$ is a VC class with VC dimension V.

2. Fast rates of convergence

   Under variance control: rate faster than $n^{-1/2}$

3. Convex risk minimization
4. Oracle inequalities - Model selection

Main results in classification theory

1. Bayes risk consistency and rate of convergence

   Complexity control:

   $$\mathbb{E} \sup_{g \in \mathcal{G}} \mid L_n(g) - L(g) \mid \leq C \sqrt{\frac{V}{n}}$$

   if $\mathcal{G}$ is a VC class with VC dimension V.

2. Fast rates of convergence

   Under variance control: rate faster than $n^{-1/2}$

3. Convex risk minimization

4. Oracle inequalities - Model selection

Bipartite Ranking

- Same data, different questions:

  Classifying is a **local** task, while ranking is **global**!

- Ranking and scoring a set of instances

Bipartite Ranking

- Same data, different questions:

  Classifying is a **local** task, while ranking is **global**!

- Ranking and scoring a set of instances
  ... through a **scoring function** $s : \mathcal{X} \to \mathbb{R}$

- **Challenge:** develop theory and algorithms

- **Question:** are advances in classification theory/practice of any use for ranking?

Ranking - Rigorous problem statement

- **Data:** $(X_1, Y_1), \ldots, (X_n, Y_n) \in \left(\mathcal{X} \times \{-1, +1\}\right)^{\otimes n}$

Ranking - Rigorous problem statement

- **Data:** $(X_1, Y_1), \ldots, (X_n, Y_n) \in \left(\mathcal{X} \times \{-1, +1\}\right)^{\otimes n}$
- **Want to:** rank $X_1, \ldots, X_n$ through a scoring function $s : \mathcal{X} \to \mathbb{R}$, so that a large number instances with $+1$ labels appear on the top with high probability

Ranking - Rigorous problem statement

- **Data:** $(X_1, Y_1), \ldots, (X_n, Y_n) \in \left(\mathcal{X} \times \{-1, +1\}\right)^{\otimes n}$
- **Want to:** rank $X_1, \ldots, X_n$ through a scoring function $s : \mathcal{X} \to \mathbb{R}$, so that a large number instances with $+1$ labels appear on the top with high probability
- **Class of solutions:**

$$\mathcal{S}^* = \{ T \circ \eta \mid T : [0,1] \to \mathbb{R} \text{ increasing}\}$$

Ranking - Rigorous problem statement

- **Data:** $(X_1, Y_1), \ldots, (X_n, Y_n) \in \left(\mathcal{X} \times \{-1, +1\}\right)^{\otimes n}$
- **Want to:** rank $X_1, \ldots, X_n$ through a scoring function $s : \mathcal{X} \to \mathbb{R}$, so that a large number instances with $+1$ labels appear on the top with high probability
- **Class of solutions:**

$$\mathcal{S}^* = \{\, T \circ \eta \mid T : [0, 1] \to \mathbb{R} \text{ increasing}\}$$

- **Need to:** find an optimization criterion reflecting ranking performance

ROC Curve and AUC

ROC Curve and AUC

- True positive rate:

$$\mathrm{TPR}_s(x) = \mathbb{P}\left(s(X) \geq x \mid Y = 1\right)$$

- False positive rate:

$$\mathrm{FPR}_s(x) = \mathbb{P}\left(s(X) \geq x \mid Y = -1\right)$$

ROC Curve and AUC

- True positive rate:

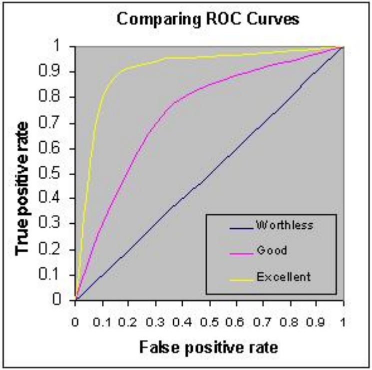
$$\mathrm{TPR}_s(x) = \mathbb{P}\left(s(X) \geq x \mid Y = 1\right)$$

- False positive rate:

$$\mathrm{FPR}_s(x) = \mathbb{P}\left(s(X) \geq x \mid Y = -1\right)$$

Receiving Operator Characteristic curve: $\quad x \mapsto \left(\mathrm{FPR}_s(x), \mathrm{TPR}_s(x)\right)$

ROC Curve and AUC



- True positive rate:

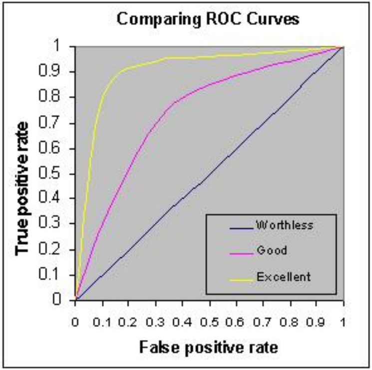
$$\mathrm{TPR}_s(x) = \mathbb{P}\left(s(X) \geq x \mid Y = 1\right)$$

- False positive rate:

$$\mathrm{FPR}_s(x) = \mathbb{P}\left(s(X) \geq x \mid Y = -1\right)$$

Receiving Operator Characteristic curve: $x \mapsto \left(\mathrm{FPR}_s(x), \mathrm{TPR}_s(x)\right)$

ROC Curve and AUC



- True positive rate:

$$\mathrm{TPR}_s(x) = \mathbb{P}\left(s(X) \geq x \mid Y = 1\right)$$

- False positive rate:

$$\mathrm{FPR}_s(x) = \mathbb{P}\left(s(X) \geq x \mid Y = -1\right)$$

Receiving Operator Characteristic curve: $\quad x \mapsto \left(\mathrm{FPR}_s(x), \mathrm{TPR}_s(x)\right)$

$$\text{AUC} = \text{Area Under an ROC Curve}$$

Connection to standard classification

Ranking = Classification of pairs of observations

Connection to standard classification

Ranking $=$ Classification of pairs of observations

- **Ranking vs. Classification**
  - same performance/risk measure
  - same raw data: $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.
  - different statistical model $(X, X', R) \in \mathcal{X} \times \mathcal{X} \times \{-1, +1\}$

Connection to standard classification

Ranking = Classification of pairs of observations

- **Ranking vs. Classification**
  - same performance/risk measure
  - same raw data: $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.
  - different statistical model $(X, X', R) \in \mathcal{X} \times \mathcal{X} \times \{-1, +1\}$
- **Empirical criterion for ranking:**

$$L_n(r) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}_{[(Y_i - Y_j) \cdot r(X_i, X_j) < 0]}$$

Connection to standard classification

Ranking = Classification of pairs of observations

- **Ranking vs. Classification**
  - same performance/risk measure
  - same raw data: $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.
  - different statistical model $(X, X', R) \in \mathcal{X} \times \mathcal{X} \times \{-1, +1\}$
- **Empirical criterion for ranking:** $Z_i = (X_i, Y_i)$

$$L_n(r) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}_{[(Y_i - Y_j) \cdot r(X_i, X_j) < 0]} = \frac{1}{n(n-1)} \sum_{i \neq j} q_r(Z_i, Z_j)$$

Connection to standard classification

Ranking = Classification of pairs of observations

- **Ranking vs. Classification**
  - same performance/risk measure
  - same raw data: $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.
  - different statistical model $(X, X', R) \in \mathcal{X} \times \mathcal{X} \times \{-1, +1\}$

- **Empirical criterion for ranking:** $Z_i = (X_i, Y_i)$

$$L_n(r) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}_{[(Y_i - Y_j) \cdot r(X_i, X_j) < 0]} = \frac{1}{n(n-1)} \sum_{i \neq j} q_r(Z_i, Z_j)$$

- **But:** the pairs $\{(Z_i, Z_j)\}_{1 \leq i < j \leq n}$ are not independent!

U-statistics

- $Z_1, ..., Z_n$ i.i.d.
- $q : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a symmetric real-valued function.

Definition

The statistic

$$U_n(Z_1, ..., Z_n) = \frac{1}{n(n-1)} \sum_{i \neq j} q(Z_i, Z_j)$$

is a *U-statistic* of order 2 with kernel q.

U-statistics

- $Z_1, \ldots, Z_n$ i.i.d.
- $q : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a symmetric real-valued function.

Definition

The U-statistic U_n is *degenerate* if $\mathbb{E}(q(z, Z_1)) = 0, \forall z \in \mathcal{Z}$.

U-statistics

- $Z_1, ..., Z_n$ i.i.d.
- $q : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a symmetric real-valued function.

Definition

The statistic

$$U_n(Z_1, ..., Z_n) = \frac{1}{n(n-1)} \sum_{i \neq j} q(Z_i, Z_j)$$

is a *U-statistic* of order 2 with kernel q.

References: Halmos (1946), Hoeffding (1948), Serfling (1980), de la Peña and Giné (1999)

Two representations of U-statistics

- **Average of 'sums-of-i.i.d.' blocks:**

$$U_n = \frac{1}{n!} \sum_{\pi} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} q\big(Z_{\pi(i)}, Z_{\pi(\lfloor n/2 \rfloor + i)}\big)$$

where π permutations of $\{1, \ldots, n\}$

- **Hoeffding's decomposition**

$$U_n = \mathbb{E}(U_n) + 2T_n + W_n$$

with T_n empirical average and W_n degenerate U-statistic.

Two representations of U-statistics

- **Average of 'sums-of-i.i.d.' blocks:**

$$U_n = \frac{1}{n!}\sum_{\pi}\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} q\left(Z_{\pi(i)}, Z_{\pi(\lfloor n/2\rfloor+i)}\right)$$

where π permutations of $\{1,\ldots,n\}$

- **Hoeffding's decomposition**

$$U_n = \mathbb{E}(U_n) + 2T_n + W_n$$

with T_n empirical average and W_n degenerate U-statistic.

First-order analysis

Theorem

Set

- $\mathcal{R}$ *class of ranking rules of* VC *dimension* V
- *Empirical risk:* $L_n(r) = \dfrac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}_{[(Y_i - Y_j) \cdot r(X_i, X_j) < 0]}$
- *Empirical risk minimizer:* $r_n = \arg\min_{r \in \mathcal{R}} L_n(r)$

First-order analysis

Theorem

Set

- $\mathcal{R}$ *class of ranking rules of* VC *dimension* V
- *Empirical risk:* $L_n(r) = \dfrac{1}{n(n-1)} \displaystyle\sum_{i \neq j} \mathbb{I}_{[(Y_i - Y_j)\cdot r(X_i, X_j) < 0]}$
- *Empirical risk minimizer:* $r_n = \arg\min_{r \in \mathcal{R}} L_n(r)$

Then, with probability larger than $1 - \delta$*:*

$$L(r_n) - \inf_{r \in \mathcal{R}} L(r) \leq c\sqrt{\frac{V}{n}} + 2\sqrt{\frac{\log(1/\delta)}{n-1}} \ .$$

Structure of a U-statistic

- **Hoeffding's decomposition:**

$$U_n = \mathbb{E}(U_n) + 2T_n + W_n = \frac{1}{n(n-1)} \sum_{i \neq j} q(Z_i, Z_j)$$

with $T_n = \frac{1}{n} \sum_{i=1}^{n} h(Z_i)$ and $W_n = \frac{1}{n(n-1)} \sum_{i \neq j} \widehat{h}(Z_i, Z_j)$

Structure of a U-statistic

- **Hoeffding's decomposition:**

$$U_n = \mathbb{E}(U_n) + 2T_n + W_n = \frac{1}{n(n-1)} \sum_{i \neq j} q(Z_i, Z_j)$$

with $T_n = \dfrac{1}{n} \sum_{i=1}^{n} h(Z_i)$ and $W_n = \dfrac{1}{n(n-1)} \sum_{i \neq j} \widehat{h}(Z_i, Z_j)$

where

- $h(z) = \mathbb{E}q(Z, z) - \mathbb{E}U_n$,
- $\widehat{h}(Z_i, Z_j) = q(Z_i, Z_j) - \mathbb{E}U_n - h(Z_i) - h(z_j)$

Insights

- leading term T_n is an empirical process

Insights

- leading term T_n is an empirical process
  - handled by Talagrand's concentration inequality

Insights

- leading term T_n is an empirical process
  - handled by Talagrand's concentration inequality
  - involves "standard" complexity measures:
    e.g. Localized Rademacher Averages

Insights

- leading term T_n is an empirical process
  - handled by Talagrand's concentration inequality
  - involves "standard" complexity measures:
    e.g. Localized Rademacher Averages

$\Rightarrow$ Variance control involves the function h

Insights

- leading term T_n is an empirical process
  - handled by Talagrand's concentration inequality
  - involves "standard" complexity measures:
    e.g. Localized Rademacher Averages

$\Rightarrow$ Variance control involves the function h

- W_n requires an exponential inequality for degenerate U-processes

Insights

- leading term T_n is an empirical process
  - handled by Talagrand's concentration inequality
  - involves "standard" complexity measures:
    e.g. Localized Rademacher Averages

$\Rightarrow$ Variance control involves the function h

- W_n requires an exponential inequality for degenerate U-processes
  - VC classes - exponential inequality by Arcones and Giné (AoP1993)

Insights

- leading term T_n is an empirical process
  - handled by Talagrand's concentration inequality
  - involves "standard" complexity measures:
    e.g. Localized Rademacher Averages

$\Rightarrow$ Variance control involves the function h

- W_n requires an exponential inequality for degenerate U-processes
  - VC classes - exponential inequality by Arcones and Giné (AoP1993)
  - general case - a new moment inequality

Insights

- leading term T_n is an empirical process
  - handled by Talagrand's concentration inequality
  - involves "standard" complexity measures:
    e.g. Localized Rademacher Averages

$\Rightarrow$ Variance control involves the function h

- W_n requires an exponential inequality for degenerate U-processes
  - VC classes - exponential inequality by Arcones and Giné (AoP1993)
  - general case - a new moment inequality

$\Rightarrow$ additional complexity measures

Fast rates - Notations

- **Kernel:**

$$q_r((x,y),(x',y')) = \mathbb{I}_{[(y-y')\cdot r(x,x')<0]} - \mathbb{I}_{[(y-y')\cdot r^*(x,x')<0]}$$

Fast rates - Notations

- **Kernel:**

$$q_r((x,y),(x',y')) = \mathbb{I}_{[(y-y')\cdot r(x,x')<0]} - \mathbb{I}_{[(y-y')\cdot r^*(x,x')<0]}$$

- **Excess risk:**

$$\Lambda(r) = L(r) - L^*$$

Fast rates - Notations

- **Kernel:**

$$q_r((x,y),(x',y')) = \mathbb{I}_{[(y-y')\cdot r(x,x')<0]} - \mathbb{I}_{[(y-y')\cdot r^*(x,x')<0]}$$

- **Excess risk:**

$$\Lambda(r) = L(r) - L^* = \mathbb{E}q_r((X,Y),(X',Y'))$$

Fast rates - Notations

- **Kernel:**

$$q_r((x,y),(x',y')) = \mathbb{I}_{[(y-y')\cdot r(x,x')<0]} - \mathbb{I}_{[(y-y')\cdot r^*(x,x')<0]}$$

- **Excess risk:**

$$\Lambda(r) = L(r) - L^* = \mathbb{E} q_r((X,Y),(X',Y'))$$

- ***U*-process indexed by ranking rule $r \in \mathcal{R}$**

$$\Lambda_n(r) - \Lambda(r) = \frac{1}{n(n-1)} \sum_{i \neq j} q_r((X_i,Y_i),(X_j,Y_j)),$$

Fast rates - Notations

- **Kernel:**

$$q_r((x,y),(x',y')) = \mathbb{I}_{[(y-y')\cdot r(x,x')<0]} - \mathbb{I}_{[(y-y')\cdot r^*(x,x')<0]}$$

- **Excess risk:**

$$\Lambda(r) = L(r) - L^* = \mathbb{E} q_r((X,Y),(X',Y'))$$

- ***U*-process indexed by ranking rule** $r \in \mathcal{R}$

$$\Lambda_n(r) - \Lambda(r) = \frac{1}{n(n-1)} \sum_{i \neq j} q_r((X_i,Y_i),(X_j,Y_j)),$$

- **Key quantity:**

$$h_r(x,y) = \mathbb{E} q_r((x,y),(X',Y')) - \Lambda(r)$$

(function in the empirical average part)

Fast rates - VC case

Theorem

Assume we have:

- *The class $\mathcal{R}$ of ranking rules has finite VC dimension V.*
- *for all $r \in \mathcal{R}$,*

$$\mathbb{V}(h_r(X, Y)) \leq c\,\Lambda(r)^\alpha \qquad \textbf{(V)}$$

with some constants $c > 0$ and $\alpha \in [0, 1]$.

Fast rates - VC case

Theorem

Assume we have:

- *The class $\mathcal{R}$ of ranking rules has finite VC dimension V.*
- *for all $r \in \mathcal{R}$,*
$$\mathbb{V}(h_r(X, Y)) \leq c\,\Lambda(r)^{\alpha} \qquad \textbf{(V)}$$
with some constants $c > 0$ and $\alpha \in [0, 1]$.

Then, with probability larger than $1 - \delta$:

$$L(r_n) - L^* \leq 2\left(\inf_{r\in\mathcal{R}} L(r) - L^*\right) + C\left(\frac{V\log(n/\delta)}{n}\right)^{1/(2-\alpha)}$$

Comments

Proof uses:

- Hoeffding's decomposition of the empirical excess risk
- A new moment inequality
- Excess risk bound for approximate empirical risk minimizers by Massart (LNSF, 2006)
  (check also Bartlett and Mendelson (PTRF, 2006))

Comments

Proof uses:

- Hoeffding's decomposition of the empirical excess risk
- A new moment inequality
- Excess risk bound for approximate empirical risk minimizers by Massart (LNSF, 2006)
  (check also Bartlett and Mendelson (PTRF, 2006))

Question

Sufficient condition for Assumption **(V)**:

$$\forall r \in \mathcal{R}, \quad \mathbb{V}(h_r(X,Y)) \leq c\,\Lambda(r)^{\alpha} \qquad ?$$

Comments

Proof uses:

- Hoeffding's decomposition of the empirical excess risk
- A new moment inequality
- Excess risk bound for approximate empirical risk minimizers by Massart (LNSF, 2006)
  (check also Bartlett and Mendelson (PTRF, 2006))

Question

Sufficient condition for Assumption **(V)**:

$$\forall r \in \mathcal{R}, \quad \mathbb{V}(h_r(X, Y)) \leq c\,\Lambda(r)^{\alpha} \qquad ?$$

Noise assumptions on $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$?

Example: bipartite ranking

Noise Assumption (NA)

There exist constants $c > 0$ and $\alpha \in [0,1]$ such that :

$$\forall x \in \mathcal{X}, \quad \mathbb{E}(|\eta(x) - \eta(X)|^{-\alpha}) \leq c\,.$$

Example: bipartite ranking

Noise Assumption (NA)

There exist constants $c > 0$ and $\alpha \in [0, 1]$ such that :

$$\forall x \in \mathcal{X}, \quad \mathbb{E}(|\eta(x) - \eta(X)|^{-\alpha}) \leq c\,.$$

Discussion:

- Compare to: $\forall x, x' \in \mathcal{X}, \quad |\eta(x) - \eta(x')|^{-1} \leq c$ (when splitting the sample)
- $\alpha = 0$: no restriction.

Example: bipartite ranking

Noise Assumption (NA)

There exist constants $c > 0$ and $\alpha \in [0, 1]$ such that :

$$\forall x \in \mathcal{X}, \quad \mathbb{E}(|\eta(x) - \eta(X)|^{-\alpha}) \leq c\,.$$

Discussion:

- Compare to: $\forall x, x' \in \mathcal{X}, \quad |\eta(x) - \eta(x')|^{-1} \leq c$ (when splitting the sample)
- $\alpha = 0$: no restriction.
- $\alpha = 1$: too restrictive.

Sufficient condition for (NA) with $\alpha < 1$

$\eta(X)$ absolutely continuous on $[0, 1]$ with bounded density

Additional complexity measures

Degenerate U-process

We have

$$W_n = \sup_{r \in \mathcal{R}} \left| \sum_{i,j} \widehat{h}_r((X_i, Y_i), (X_j, Y_j)) \right|$$

where $\widehat{h}_r((x,y),(x',y')) = q_r((x,y),(x',y')) - \Lambda(r) - h_r(x,y) - h_r(x',y')$

Additional complexity measures

Degenerate U-process

Set

$$W_n = \sup_{f \in \mathcal{F}} \left| \sum_{i,j} f(Z_i, Z_j) \right|$$

where $\mathcal{F}$ is a class of degenerate kernels

Additional complexity measures

Degenerate U-process

Set

$$W_n = \sup_{f \in \mathcal{F}} \left| \sum_{i,j} f(Z_i, Z_j) \right|$$

where $\mathcal{F}$ is a class of degenerate kernels

Complexity measures:

$$(1)\ Z_\epsilon = \sup_{f \in \mathcal{F}} \left| \textstyle\sum_{i,j} \epsilon_i \epsilon_j f(Z_i, Z_j) \right|$$

$$(2)\ U_\epsilon = \sup_{f \in \mathcal{F}} \sup_{\alpha : \|\alpha\|_2 \leq 1} \sum_{i,j} \epsilon_i \alpha_j f(Z_i, Z_j)$$

$$(3)\ M_\epsilon = \sup_{f \in \mathcal{F}} \max_{k=1 \dots n} \left| \sum_{i=1}^{n} \epsilon_i f(Z_i, Z_k) \right|$$

A Moment Inequality

Theorem

If W_n is a degenerate U-process, then there exists a universal constant $C > 0$ such that for all n and $q \geq 2$,

$$(\mathbb{E}W_n^q)^{1/q} \leq C\left(\mathbb{E}Z_\epsilon + q^{1/2}\mathbb{E}U_\epsilon + q(\mathbb{E}M_\epsilon + n) + q^{3/2}n^{1/2} + q^2\right).$$

A Moment Inequality

Theorem

If W_n is a degenerate U-process, then there exists a universal constant $C > 0$ such that for all n and $q \geq 2$,

$$(\mathbb{E}W_n^q)^{1/q} \leq C\left(\mathbb{E}Z_\epsilon + q^{1/2}\mathbb{E}U_\epsilon + q(\mathbb{E}M_\epsilon + n) + q^{3/2}n^{1/2} + q^2\right).$$

- **Main tools:** symmetrization, decoupling and concentration inequalities
- **Sources:** de la Peña and Giné (1999), Boucheron, Bousquet, Lugosi and Massart (AoP, 2005)

A Moment Inequality

Theorem

If W_n is a degenerate U-process, then there exists a universal constant $C > 0$ such that for all n and $q \geq 2$,

$$(\mathbb{E}W_n^q)^{1/q} \leq C\left(\mathbb{E}Z_\epsilon + q^{1/2}\mathbb{E}U_\epsilon + q(\mathbb{E}M_\epsilon + n) + q^{3/2}n^{1/2} + q^2\right).$$

- **Main tools:** symmetrization, decoupling and concentration inequalities
- **Sources:** de la Peña and Giné (1999), Boucheron, Bousquet, Lugosi and Massart (AoP, 2005)

Related work

Adamczak (AoP, to appear), Arcones and Giné (AoP, 1993), Giné, Latala and Zinn (HDP II, 2000), Houdré and Reynaud-Bouret (SIA, 2003)

Control of the degenerate part

Corollary

With probability $1-\delta$,

$$W_n \leq C\left(\frac{\mathbb{E}Z_\epsilon}{n^2} + \frac{\mathbb{E}U_\epsilon\sqrt{\log(1/\delta)}}{n^2} + \frac{\mathbb{E}M_\epsilon \log(1/\delta)}{n^2} + \frac{\log(1/\delta)}{n}\right)$$

Control of the degenerate part

Corollary

With probability $1-\delta$,

$$W_n \leq C\left(\frac{\mathbb{E}Z_\epsilon}{n^2} + \frac{\mathbb{E}U_\epsilon\sqrt{\log(1/\delta)}}{n^2} + \frac{\mathbb{E}M_\epsilon \log(1/\delta)}{n^2} + \frac{\log(1/\delta)}{n}\right)$$

VC case

$$\mathbb{E}Z_\epsilon \leq CnV \;, \qquad \mathbb{E}U_\epsilon \leq Cn\sqrt{V} \;, \qquad \mathbb{E}_\epsilon M_\epsilon \leq C\sqrt{Vn}$$

Hence, with probability $1-\delta$

$$W_n \leq \frac{1}{n}\left(V + \log(1/\delta)\right)$$

Summary

Have seen...

- A framework for ranking
- Connection to AUC criterion
- Interpretation as pairwise classification
- Consistency, excess risk bounds and fast rates
- U-statistics improve on splitting the sample through weaker noise assumption
- A new moment inequality for degenerate U-processes
- Additional complexity measures: Rademacher averages and Rademacher chaoses

What's next?

Summary

Have seen...

- A framework for ranking
- Connection to AUC criterion
- Interpretation as pairwise classification
- Consistency, excess risk bounds and fast rates
- U-statistics improve on splitting the sample through weaker noise assumption
- A new moment inequality for degenerate U-processes
- Additional complexity measures: Rademacher averages and Rademacher chaoses

What's next?

... Optimizing the ROC curve in the sup norm sense

A functional criterion for measuring ranking performance

- **Notations:**

$$\begin{aligned}
\mathcal{S} &= \{s : \mathcal{X} \subset \mathbb{R}^d \to | \text{ borelian}\} \text{ set of scoring functions,} \\
H(dx) &= \mathcal{L}(X \mid Y = -1) \text{ and } G(dx) = \mathcal{L}(X \mid Y = +1), \\
H_s(dt) &= \mathcal{L}(s(X) \mid Y = -1) \text{ and } G_s(dt) = \mathcal{L}(s(X) \mid Y = +1).
\end{aligned}$$

Definition

The ROC curve of the scoring function is the curve:

$$t \in \mathbb{R} \mapsto (1 - H_s(z), 1 - G_s(z)) .$$

When G_s and H_s are continuous, it is the plot of the mapping:

$$\mathrm{ROC}(s, .) : \alpha \in [0, 1] \mapsto 1 - G_s \circ H_s(1 - \alpha).$$

By convention, jumps are connected by line segments.

A functional criterion for measuring ranking performance

- **Properties:**
  - $\mathrm{ROC}(s, .)$ increasing, connects $(0, 0)$ to $(1, 1)$

A functional criterion for measuring ranking performance

- **Properties:**
  - $\mathrm{ROC}(s,.)$ increasing, connects $(0,0)$ to $(1,1)$
  - $\mathrm{ROC}(s,.)$ is invariant by strictly increasing transforms of s

A functional criterion for measuring ranking performance

- **Properties:**
  - $\mathrm{ROC}(s, .)$ increasing, connects $(0,0)$ to $(1,1)$
  - $\mathrm{ROC}(s, .)$ is invariant by strictly increasing transforms of s
  - $\mathrm{ROC}(s, .)$ is concave iff dG_s/dH_s is monotone

A functional criterion for measuring ranking performance

- **Properties:**
  - $\mathrm{ROC}(s,.)$ increasing, connects $(0,0)$ to $(1,1)$
  - $\mathrm{ROC}(s,.)$ is invariant by strictly increasing transforms of s
  - $\mathrm{ROC}(s,.)$ is concave iff dG_s/dH_s is monotone
  - If dG_s/dH_s is constant on some interval in the range of $s(X)$, $\mathrm{ROC}(s,.)$ is linear on the corresponding domain

A functional criterion for measuring ranking performance

- **Properties:**
  - $\mathrm{ROC}(s,.)$ increasing, connects $(0,0)$ to $(1,1)$
  - $\mathrm{ROC}(s,.)$ is invariant by strictly increasing transforms of s
  - $\mathrm{ROC}(s,.)$ is concave iff dG_s/dH_s is monotone
  - If dG_s/dH_s is constant on some interval in the range of $s(X)$, $\mathrm{ROC}(s,.)$ is linear on the corresponding domain

A partial order on $\mathcal{S}$

$$s_1 \text{ is better than } s_2 \Leftrightarrow \forall \alpha \in (0,1),\ \ \mathrm{ROC}(s_1,\alpha) \geq \mathrm{ROC}(s_2,\alpha)$$

A functional criterion for measuring ranking performance

- **Neyman-Pearson theory:**
  - $\mathrm{ROC}(s,.)$ is the **power curve** of the test statistic $s(X)$ for discriminating between $\mathcal{H}_0 : X \sim H(dx)$ *vs.* $\mathcal{H}_1 : X \sim G(dx)$
  - The likelihood ratio $\phi(X)$ yields a **uniformly most powerful** test

$$\phi(X) = \frac{dG}{dH}(X) = \frac{1-p}{p} \times \frac{\eta(X)}{1-\eta(X)}.$$

  - $\mathcal{S}^*$ forms the set of optimal scoring functions w.r.t. the ROC criterion:

$$\forall (s^*, s) \in \mathcal{S}^* \times \mathcal{S},\ \forall \alpha \in [0,1] : \ \ \mathrm{ROC}(s,\alpha) \leq \mathrm{ROC}^*(\alpha) \overset{def}{=} \mathrm{ROC}(s^*,\alpha).$$

- **Additional notations**

Q(s(X),α)	:	(1-α)-quantile of $s(X)$ given $Y=-1$
Q$^*(\alpha)$	:	(1-α)-quantile of $\eta(X)$ given $Y=-1$

$$R^*_\alpha = \{x \in \mathcal{X} \mid \eta(x) > Q^*(\alpha)\},\ R_{s,\alpha} = \{x \in \mathcal{X} \mid s(x) > Q(s(X),\alpha)\}$$

A functional criterion for measuring ranking performance

- **Assumptions:**

(A1) The distributions G and H are *equivalent*. In addition, the likelihood ratio $\phi(X)$ is supposed to be bounded, *i.e. ess sup* $\eta(X) < 1$.

(A2) The distribution of $\eta(X)$ is continuous.

Pointwise difference (Clémençon & Vayatis (2008b))

For any $s \in \mathcal{S}$, we have:

$$\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(s,\alpha) = \frac{\mathbb{E}(|\eta(X) - Q^*(\alpha)|\ \mathbb{I}\{X \in R^*_\alpha \Delta R_{s,\alpha}\})}{p(1 - Q^*(\alpha))} + \frac{1-p}{p}\frac{Q^*(\alpha)}{1-Q^*(\alpha)}\left(\alpha - 1 + H_s(Q(s(X),\alpha))\right),$$

A functional criterion for measuring ranking performance

- **Assumptions:**
  - (A1) The distributions G and H are *equivalent*. In addition, the likelihood ratio $\phi(X)$ is supposed to be bounded, *i.e. ess sup* $\eta(X) < 1$.
  - (A2) The distribution of $\eta(X)$ is continuous.

Pointwise difference (Clémençon & Vayatis (2008b))

For any $s \in \mathcal{S}$, we have:

$$\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(s, \alpha) = \frac{\mathbb{E}(|\eta(X) - Q^*(\alpha)|\ \mathbb{I}\{X \in R^*_\alpha \Delta R_{s,\alpha}\})}{p(1 - Q^*(\alpha))} + \frac{1-p}{p} \frac{Q^*(\alpha)}{1 - Q^*(\alpha)} \left(\alpha - 1 + H_s(Q(s(X), \alpha))\right),$$

- Ranking boils down to recover **all** level sets of η...

A functional criterion for measuring ranking performance

- **Assumptions:**
  - (A1) The distributions G and H are *equivalent*. In addition, the likelihood ratio $\phi(X)$ is supposed to be bounded, *i.e. ess sup* $\eta(X) < 1$.
  - (A2) The distribution of $\eta(X)$ is continuous.

Pointwise difference (Clémençon & Vayatis (2008b))

For any $s \in \mathcal{S}$, we have:

$$\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(s, \alpha) = \frac{\mathbb{E}(|\eta(X) - Q^*(\alpha)|\ \mathbb{I}\{X \in R^*_\alpha \Delta R_{s,\alpha}\})}{p(1 - Q^*(\alpha))} + \frac{1-p}{p} \frac{Q^*(\alpha)}{1 - Q^*(\alpha)} \left(\alpha - 1 + H_s(Q(s(X), \alpha))\right),$$

- Ranking boils down to recover **all** level sets of η...
  ... not only $\{\eta(x) > 1/2\}$ (in contrast to classification)

Ranking performance - The AUC summary criterion

- The L_1-**metric** is a convenient distance in the ROC space:

$$\min_s \int_{\alpha=0}^{1} \{\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(s,\alpha)\}\, d\alpha = \mathrm{AUC}^* - \max_s \mathrm{AUC}(s),$$

where the **area under the** ROC **curve** is defined by

$$\mathrm{AUC}(s) = \int_{\alpha=0}^{1} \mathrm{ROC}(s,\alpha)d\alpha$$

and $\mathrm{AUC}^* = \mathrm{AUC}(s^*)$ for $s \in \mathcal{S}^*$.

- The L_1-**metric** is a convenient distance in the ROC space:

$$\min_s \int_{\alpha=0}^{1} \{\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(s,\alpha)\}\, d\alpha = \mathrm{AUC}^* - \max_s \mathrm{AUC}(s),$$

where the **area under the** ROC **curve** is defined by

$$\mathrm{AUC}(s) = \int_{\alpha=0}^{1} \mathrm{ROC}(s,\alpha) d\alpha$$

and $\mathrm{AUC}^* = \mathrm{AUC}(s^*)$ for $s \in \mathcal{S}^*$.

- **Probabilistic interpretation:** If $s(X)$ is a continuous r.v., then

$$\begin{aligned} \mathrm{AUC}(s) &= \mathbb{P}\{s(X) > s(X') \mid Y = 1, Y' = -1\} \\ &= \frac{1}{2p(1-p)} \mathbb{P}\{(s(X) - s(X'))(Y - Y') > 0\}\ . \end{aligned}$$

A stronger measure of ranking performance

- Consider the metric induced by the *sup-norm* in the ROC space:

$$||\mathrm{ROC}^* - \mathrm{ROC}(s, .)||_\infty = \sup_{\alpha \in (0,1)} \{\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(s, \alpha)\}$$

- Equivalent to statistical recovery of the **continuum** of η's level sets

A stronger measure of ranking performance

- Consider the metric induced by the *sup-norm* in the ROC space:

$$||\mathrm{ROC}^* - \mathrm{ROC}(s,.)||_\infty = \sup_{\alpha\in(0,1)} \{\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(s,\alpha)\}$$

- Equivalent to statistical recovery of the **continuum** of η's level sets
- No simple empirical counterpart to minimize...

A stronger measure of ranking performance

- Consider the metric induced by the *sup-norm* in the ROC space:

$$||\mathrm{ROC}^* - \mathrm{ROC}(s,.)||_\infty = \sup_{\alpha\in(0,1)} \{\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(s,\alpha)\}$$

- Equivalent to statistical recovery of the **continuum** of η's level sets
- No simple empirical counterpart to minimize...
- ... need to discretize the learning using **Approximation theory**

A stronger measure of ranking performance

- Consider the metric induced by the *sup-norm* in the ROC space:

$$||\mathrm{ROC}^* - \mathrm{ROC}(s,.)||_\infty = \sup_{\alpha\in(0,1)} \{\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(s,\alpha)\}$$

- Equivalent to statistical recovery of the **continuum** of η's level sets
- No simple empirical counterpart to minimize...
- ... need to discretize the learning using **Approximation theory**
- Let $\widetilde{\mathrm{ROC}}$ an (adaptive) **approximant** of ROC^* **described by a finite number of well-chosen level sets**

  $\Rightarrow$ the objective is now:

$$\min_{s\in\mathcal{S}_0} ||\widetilde{\mathrm{ROC}}^* - \mathrm{ROC}(s,.)||_\infty$$

ROC Optimization through Recursive Partitioning

- Perform ROC optimization over the set $\mathcal{S}_N$ of

  piecewise constant scoring functions with N pieces
- D-**representation:**

$$s_N(x) = \sum_{j=1}^{N} a_j \, \mathbb{I}\{x \in C_j\},$$

where $(a_j)_{j \geq 1}$ decreasing, $\mathcal{C}_N = (C_j)_{1 \leq j \leq N}$ partition

ROC Optimization through Recursive Partitioning

- Perform ROC optimization over the set $\mathcal{S}_N$ of

  piecewise constant scoring functions with N pieces

- *D*-**representation:**

$$s_N(x) = \sum_{j=1}^{N} a_j \, \mathbb{I}\{x \in C_j\},$$

  where $(a_j)_{j \geq 1}$ decreasing, $\mathcal{C}_N = (C_j)_{1 \leq j \leq N}$ partition

- *I*-**representation:** taking $a_j = N - j + 1$, $R_1 = C_1$, $C_i = R_i \setminus R_{i-1}$

$$s_N(x) = \sum_{j=1}^{N} \mathbb{I}\{x \in R_j\}.$$

ROC Optimization through Recursive Partitioning

- $\mathrm{ROC}(s_N)$ is the **broken line** that connects $\{\alpha(R_j), \beta(R_j)\}_{0 \leq j \geq N}$, where

$$\begin{aligned} \alpha(C) &= \mathbb{P}\{X \in C \mid Y = -1\}, \\ \beta(C) &= \mathbb{P}\{X \in C \mid Y = +1\} \ . \end{aligned}$$

- $\mathrm{ROC}(s_N)$ is the **broken line** that connects $\{\alpha(R_j), \beta(R_j)\}_{0 \leq j \geq N}$, where

$$\begin{aligned} \alpha(C) &= \mathbb{P}\{X \in C \mid Y = -1\}, \\ \beta(C) &= \mathbb{P}\{X \in C \mid Y = +1\} \; . \end{aligned}$$

- **"Concavification":** $s_{N,\sigma}(x) = \sum_{j=1}^{N} (N - j + 1)\mathbb{I}\{x \in C_{\sigma(j)}\}$ with

$$\frac{\beta(C_{\sigma(1)})}{\alpha(C_{\sigma(1)})} \geq \frac{\beta(C_{\sigma(2)})}{\alpha(C_{\sigma(2)})} \geq \ldots \geq \frac{\beta(C_{\sigma(N)})}{\alpha(C_{\sigma(N)})}.$$

has maximum AUC among all scoring functions based on the C_j's (voir Clémençon & Vayatis (2009a)), as the *plug-in* scoring function

$$\tilde{\eta}(x) = \sum_{j=1}^{N} \frac{p}{(p + (1 - p)\alpha(C_j)/\beta(C_j))} \cdot \mathbb{I}\{x \in C_j\}$$

Proposition, Clémençon & Vayatis (2008a)

Assume $(A1)-(A2)$ and that there exists $c>0$ such that $H^{*\prime}(u)\geq c$ for any $u\in \text{supp}(H^{*\prime})$, where $\text{supp}(H^{*\prime})$ is the support of $H^{*\prime}$. Then, ROC^* is twice differentiable on $[0,1]$ with bounded derivatives: $\forall \alpha\in[0,1]$,

$$\begin{aligned}
\frac{d}{d\alpha}\mathrm{ROC}^*(\alpha) &= \frac{1-p}{p}\cdot\frac{Q^*(\alpha)}{1-Q^*(\alpha)}\ , \\
\frac{d^2}{d\alpha^2}\mathrm{ROC}^*(\alpha) &= \frac{1-p}{p}\cdot\frac{Q^{*\prime}(\alpha)}{(1-Q^*(\alpha))^2}\ ,
\end{aligned}$$

where $Q^{*\prime}(\alpha)=-1/H^{*\prime}(Q^*(\alpha))$, $H^*=H_\eta$.

ROC Optimization through Recursive Partitioning

Proposition, Clémençon & Vayatis (2008a)

Assume $(A1) - (A2)$ and that there exists $c > 0$ such that $H^{*\prime}(u) \geq c$ for any $u \in \text{supp}(H^{*\prime})$, where $\text{supp}(H^{*\prime})$ is the support of $H^{*\prime}$. Then, ROC^* is twice differentiable on $[0,1]$ with bounded derivatives: $\forall \alpha \in [0,1]$,

$$\begin{aligned}
\frac{d}{d\alpha}\mathrm{ROC}^*(\alpha) &= \frac{1-p}{p} \cdot \frac{Q^*(\alpha)}{1-Q^*(\alpha)} , \\
\frac{d^2}{d\alpha^2}\mathrm{ROC}^*(\alpha) &= \frac{1-p}{p} \cdot \frac{Q^{*\prime}(\alpha)}{(1-Q^*(\alpha))^2} ,
\end{aligned}$$

where $Q^{*\prime}(\alpha) = -1/H^{*\prime}(Q^*(\alpha))$, $H^* = H_\eta$.

- There exists $s_N \in \mathcal{S}_N$ such that:

$$d_\infty(s^*, s_N) \leq C \cdot N^{-2} ,$$

where the constant C depends only on the distribution.

Adaptive recursive piecewise linear approximation of ROC^*

- **Initialization:** main diagonal of the ROC space, connect the knots

$$(\alpha^*_{0,0}, \beta^*_{0,0}) = (0,0) \text{ and } (\alpha^*_{0,1}, \beta^*_{0,1}) = (1,1).$$

Adaptive recursive piecewise linear approximation of ROC^*

- **Initialization:** main diagonal of the ROC space, connect the knots

$$(\alpha^*_{0,0}, \beta^*_{0,0}) = (0,0) \text{ and } (\alpha^*_{0,1}, \beta^*_{0,1}) = (1,1).$$

- **First step:** break the line in order to maximize AUC:
add the knot $(\alpha^*, \mathrm{ROC}^*(\alpha^*))$ in order to maximize

$$\mathrm{AUC} = 1/2 + \{(\alpha^*_{0,1} - \alpha^*_{0,0})\mathrm{ROC}^*(\alpha) - (\beta^*_{0,1} - \beta^*_{0,0})\alpha\}/2$$

Adaptive recursive piecewise linear approximation of ROC^*

- AUC is maximum when:

$$\mathrm{ROC}^{*\prime}(\alpha) \quad = \quad \frac{\beta^*_{1,0}}{\alpha^*_{1,0}} = 1$$

- Max. is attained at $\alpha^*_{1,1}$ such that:

$$Q^*(\alpha^*_{1,1}) \quad = \quad p$$

Adaptive recursive piecewise linear approximation of ROC^*

- AUC is maximum when:

$$\mathrm{ROC}^{*\prime}(\alpha) = \frac{\beta^*_{1,0}}{\alpha^*_{1,0}} = 1$$

- Max. is attained at $\alpha^*_{1,1}$ such that:

$$Q^*(\alpha^*_{1,1}) = p$$

Adaptive recursive piecewise linear approximation of ROC^*

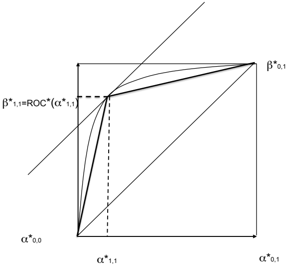
- AUC is maximum when:

$$\mathrm{ROC}^{*\prime}(\alpha) = \frac{\beta^*_{1,0}}{\alpha^*_{1,0}} = 1$$

- Max. is attained at $\alpha^*_{1,1}$ such that:

$$Q^*(\alpha^*_{1,1}) = p$$

Adaptive recursive piecewise linear approximation of ROC^*

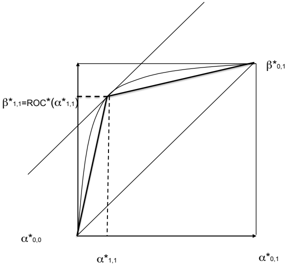


- AUC is maximum when:

$$\mathrm{ROC}^{*\prime}(\alpha) \quad = \quad \frac{\beta^*_{1,0}}{\alpha^*_{1,0}} = 1$$

- Max. is attained at $\alpha^*_{1,1}$ such that:

$$Q^*(\alpha^*_{1,1}) \quad = \quad p$$



- AUC is maximum when:

$$\mathrm{ROC}^{*\prime}(\alpha) = \frac{\beta^*_{1,0}}{\alpha^*_{1,0}} = 1$$

- Max. is attained at $\alpha^*_{1,1}$ such that:

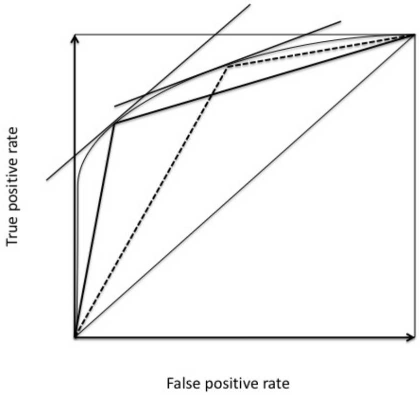
$$Q^*(\alpha^*_{1,1}) = p$$

Get the ROC curve of $s_1^*(x) = 2\mathbb{I}\{x \in C^*_{1,0}\} + \mathbb{I}\{x \in C^*_{1,1}\}$
Split $\mathcal{X}$ into $C^*_{1,0} \bigcup C^*_{1,1}$ where:

$$C^*_{1,0} = \{x \in \mathcal{X} : \eta(x) > p\} = \{x \in \mathcal{X} : \Phi(x) > 1\}$$

We have $\alpha(C^*_{1,0}) = \alpha^*_{1,1}$ and $\beta(C^*_{1,0}) = \beta^*_{1,1}$

Ranking through a binary scoring function $\neq$ Classification



Optimal binary scoring function (solid broken line) *vs.* Bayes classifier (dotted broken line) in a situation where $p > 1/2$

Adaptive recursive piecewise linear approximation of ROC^*

- **Update:** set $\alpha^*_{1,0} = \alpha^*_{0,1}$ and $\beta^*_{1,2} = \beta^*_{0,1}$.
- L_∞**-metric:** best broken line with two pieces in the L_∞ sense too
- **Iterate** the splitting/breaking rule:
  - Recursively, get a **tree-structured adaptive subdivision** of $[0,1]$:

$$\alpha^*_{D,k}, \quad k = 0, \ldots,\ 2^D.$$

  - Form a **concave piecewise linear approximant/interpolant** of ROC^*:

$$\text{connect the knots } \{(\alpha^*_{D,k}, \beta^*_{D,k}) : \quad k = 0, \ldots,\ 2^D\}$$

  - In parallel, get a **tree-structured recursive partition** of the space $\mathcal{X}$:

$$\mathcal{X} = C^*_{D,0} \bigcup \ldots \bigcup C^*_{D,2^D-1}$$

  where $C^*_{D,k} = \{x \in \mathcal{X} :\ \Delta^*_{d,k} < \eta(x) \leq \Delta^*_{d,k+1}\}$

- **Piecewise constant rule**: $s^*_D(x) = \sum_{k=0}^{2^D-1} (2^D - k + 1)\mathbb{I}\{x \in C^*_{D,k}\}$

Recursive Approximation Scheme

- The curve $\mathrm{ROC}(s_D^*)$ as a piecewise linear approximant of ROC^*:

Theorem (Clémençon & Vayatis 2008a, 2008b)

For $i \in \{1, \infty\}$, we have: $\forall D \geq 1$,

$$d_i(s_D^*, s^*) \leq C \cdot 2^{-2D}$$

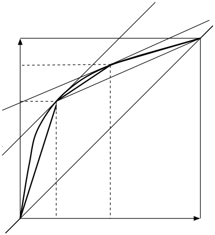
- It is the best scoring function that may be built from the $C_{D,k}^*$'s:

$$\mathrm{AUC}(s_D^*) \geq \mathrm{AUC}(s^\sigma),$$

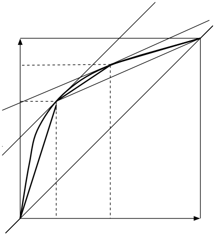
where $s^\sigma(x) = \sum_{k=0}^{2^D-1}(2^D - k + 1)\mathbb{I}\{x \in C_{D,\sigma(k)}^*\}$, for all σ in the symmetric group of $\{0, \ldots, 2^D - 1\}$

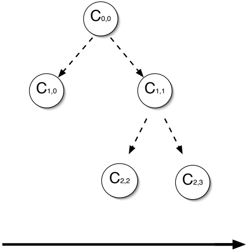
- TreeRank: statistical version based on empirical counterparts

Tree-structured approximation scheme



Tree-structured approximation scheme





Left-right oriented tree: read ranks at the bottom

The TreeRank algorithm

1. **Initialization.** Set $C_{0,0} = \mathcal{X}$.
2. **Iterations.** For $d = 0, \ldots, D-1$ and $k = 0, \ldots, 2^d - 1$:
    1. (Optimization step.) Set the entropic measure:
    $$\Lambda_{d,k+1}(C) = (\alpha_{d,k+1} - \alpha_{d,k})\hat{\beta}(C) - (\beta_{d,k+1} - \beta_{d,k})\hat{\alpha}(C) .$$
    Find the best subset $C_{d+1,2k}$ of rectangle $C_{d,k}$ in the AUC sense:
    $$C_{d+1,2k} = \underset{C \in \mathcal{C},\ C \subset C_{d,k}}{\arg\max} \Lambda_{d,k+1}(C) .$$
    Then, set $C_{d+1,2k+1} = C_{d,k} \setminus C_{d+1,2k}$.
    2. (Update.) Set
    $$\alpha_{d+1,2k+1} = \alpha_{d,k} + \hat{\alpha}(C_{d+1,2k}) \text{ and } \beta_{d+1,2k+1} = \beta_{d,k} + \hat{\beta}(C_{d+1,2k})$$
    $$\alpha_{d+1,2k+2} = \alpha_{d,k+1} \text{ and } \beta_{d+1,2k+2} = \beta_{d,k+1}.$$
3. **Output.** After D iterations, get the scoring function:
$$s_D(x) = \sum^{2^D-1} (2^D - k)\, \mathbb{I}\{x \in C_{D,k}\},$$

TreeRank's output

- Tree-structured ranking rule

- Reading the ranks:
  at the bottom, from the left to the right
- Empirical ROC and AUC estimates

TreeRank's output

- Tree-structured ranking rule

- Reading the ranks:
  at the bottom, from the left to the right

- Empirical ROC and AUC estimates

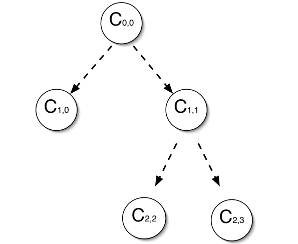
- Tree-structured ranking rule

- Reading the ranks:
  at the bottom, from the left to the right

- Empirical ROC and AUC estimates

TreeRank's output

- Tree-structured ranking rule



- Reading the ranks:
  at the bottom, from the left to the right
- Empirical ROC and AUC estimates

Theoretical Results

- If the class of subset candidates $\mathcal{C}$ is *union stable*, then $\widehat{\mathrm{ROC}}(s_D, .)$ is **concave**
- **Rate bounds** Suppose that $\mathcal{C}$ is of *VC* dimension $V < \infty$ and contains the $C^*_{d,k}$'s

Theorem (Clémençon & Vayatis '08)

For all $\delta \in (0,1)$ we have with prob. at least $1-\delta$: $\forall D \geq 1$, $i \in \{1, \infty\}$

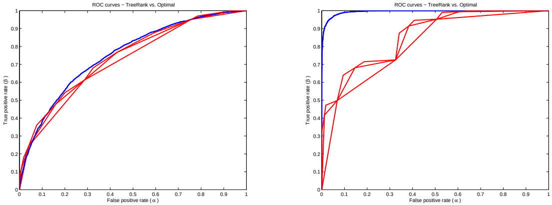
$$d_i(s_D, s_D^*) \leq c_0^D \left\{ \left(\frac{c_1^2 V}{n} \right)^{1/2(D+1)} + \left(\frac{c_2^2 \log(1/\delta)}{n} \right)^{1/2(D+1)} \right\}$$

If one chooses: $D_n \sim \sqrt{\log n}$, the rate is of order $e^{-\kappa \log(n)}$

- The same rate applies to the ROC curve estimate

Empirical Results

- Drawbacks due to the hierarchical structure: *instability* and *lack of smoothness*
- Even worse because of the **global** nature of the ranking problem: mistakes cannot be corrected by growing the tree deeper...
- Splitting rule must be **flexible** in order to mimic $\eta(x)$'s bilevel sets $C^*_{d,k}$'s, *cf* TreeRank's optimization step



TreeRank's optimization step: a data-dependent cost-sensitive classification problem

- **Cost-sensitive classification error** with asymmetry factor $\omega \in (0, 1)$

$$\mathcal{L}_\omega(C) = 2p(1-\omega)\ (1-\beta(C)) + 2(1-p)\omega\ \alpha(C)\ ,$$

TreeRank's optimization step: a data-dependent cost-sensitive classification problem

- **Cost-sensitive classification error** with asymmetry factor $\omega \in (0,1)$

$$\mathcal{L}_\omega(C) = 2p(1-\omega)\ (1-\beta(C)) + 2(1-p)\omega\ \alpha(C)\ ,$$

Theorem (Clémençon & Vayatis, 2008c)

The optimal set is $C^*_\omega = \{x\ :\ \eta(x) > \omega\}$. For all $C \subset \mathcal{X}$:

$$\mathcal{L}_\omega(C^*_\omega) \leq \mathcal{L}_\omega(C)\ .$$

The excess risk for an arbitrary set C can be written:

$$\mathcal{L}_\omega(C) - \mathcal{L}_\omega(C^*_\omega) = 2\mathbb{E}\left[\mid \eta(X) - \omega \mid \cdot \mathbb{I}\{X \in C \Delta C^*_\omega\}\right]\ .$$

The optimal error is $\mathcal{L}_\omega(C^*_\omega) = 2\mathbb{E}[\min\{\omega(1-\eta(X)), (1-\omega)\eta(X)\}]$

TreeRank's optimization step: a data-dependent cost-sensitive classification problem

- For $\omega = p$, recover the target subset $C^*_{1,0} = \{x \in \mathcal{X} : \ \eta(x) > p\}$
- Replacing p (unknown) by n_+/n, minimize the empirical version

$$\widehat{\mathcal{L}}_{\hat{p}}(C) = 4\hat{p}(1-\hat{p})\left\{1 - \widehat{\mathrm{AUC}}(s)\right\}.$$

- The optimization step is a **cost-sensitive classification problem with data-dependent cost**
- The (local) cost is the **empirical rate of positive instances within the node to split**
- **Any classification algorithm may be adapted for "solving" the Optimization step**

Example: Optimization using a data-dependent cost-sensitive version of CART

LEAFRANK ALGORITHM

1. (INPUT.) Data $\{(X_i, Y_i):\ 1 \leq i \leq n\}$ in the region $\mathcal{X}$, depth $d \geq 1$.
2. (GROWING STEP.) Run TREERANK with a naive splitting rule at depth d, yielding a ranking tree with terminal leaves: $C_{d,k},\ \ k = 0, \ldots, 2^d - 1$.
3. ("CONCAVIFICATION" STEP.) Compute $\sigma \in S(\{0, \ldots, 2^d - 1\}$ s.t.

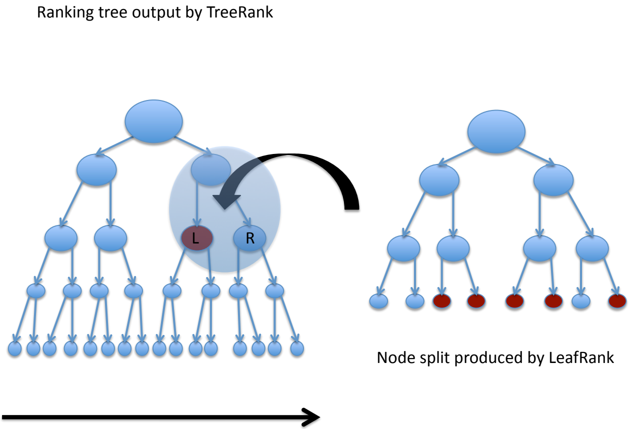
$$\frac{\widehat{\beta}(C_{d,\sigma(0)})}{\widehat{\alpha}(C_{d,\sigma(0)})} \geq \ldots \geq \frac{\widehat{\beta}(C_{d,\sigma(2^d-1)})}{\widehat{\alpha}(C_{d,\sigma(2^d-1)})}$$

4. (MERGING STEP.) $\forall k \in \{0, \ldots, 2^d - 1\}$, set $L_k = \bigcup_{l \leq k} C_{d,\sigma(l)}$ and compute the entropic measure $\widehat{\Lambda}(k) = \widehat{\beta}(L_k) - \widehat{\alpha}(L_k)$. Let

$$k^* = \arg\max_{1 \leq k \leq K} \left\{ \widehat{\beta}(L_k) - \widehat{\alpha}(L_k) \right\}.$$

5. (OUTPUT.) Form the leaves $L = L_{k^*}$ and $R = L \setminus \mathcal{X}$.

A recursive implementation of a data-dependent cost-sensitive version of CART

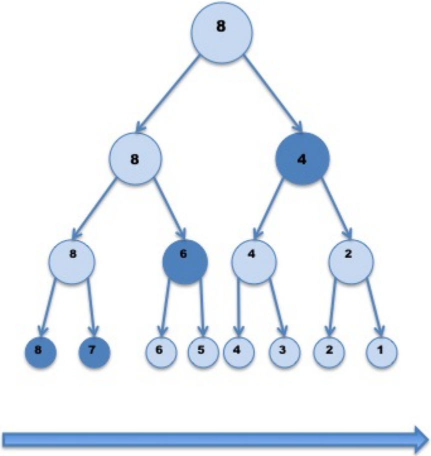


Pruning ranking trees - Merging cells

- **Model selection:** choose the "right size" for the ranking tree
- Grow first a **Master ranking tree** $\mathcal{T}$ at depth D and then select a sub- ranking tree
- **Admissible sub-tree** $\mathcal{T}(\omega)$**:** determined by $\{\omega(C_{d,k})\}$ such that:
  1. (Keep-or-kill) For all $d \in \{0, \ldots, D\}$ and $k \in \{0, \ldots, 2^D - 1\}$, the weight $\omega(C_{d,k})$ belongs to $\{0, 1\}$.
  2. (Heredity) If $\omega(C_{d,k}) = 1$, then for each cell $C_{d',k'}$ such that $C_{d,k} \subset C_{d',k'}$, we have $\omega(C_{d',k'}) = 1$.
- $C_{d,k}$ is a **terminal leave** if $\omega(C_{d,k}) = 1$ and $\forall C_{d',k'} \subset C_{d,k}$, $\omega(C_{d',k'}) = 0$
- $\mathcal{P}(\mathcal{T}(\omega)) = \{C_{d,k}$ terminal$\}$ forms a partition of $\mathcal{X}$

$$S_{\mathcal{P}(\mathcal{T}(\omega))}(x) = \sum_{C_{d,k} \in \mathcal{P}(\mathcal{T}(\omega))} (2^D - 2^{D-d}k) \cdot \mathbb{I}\{x \in C_{d,k}\}.$$

Pruning ranking trees - Merging cells



Pruning ranking trees - Merging cells

- Find the **best admissible subtree**:

$$\omega^* = \arg\max_{\omega} \mathrm{AUC}(s_{\mathcal{P}(\mathcal{T}(\omega))})$$

Pruning ranking trees - Merging cells

- Find the **best admissible subtree**:

$$\omega^* = \arg\max_{\omega} \mathrm{AUC}(S_{\mathcal{P}(\mathcal{T}(\omega))})$$

- **Cross-validation based approach**:
  - Linear complexity penalty

$$\widehat{\mathrm{CPAUC}}(S_{\mathcal{P}(\mathcal{T}(\omega))}, \lambda) = \widehat{\mathrm{AUC}}(S_{\mathcal{P}(\mathcal{T}(\omega))}) - \lambda \cdot \#\mathcal{P}(\mathcal{T}(\omega))$$

  - Choose the best λ by K-fold cross validation

Pruning ranking trees - Merging cells

- Find the **best admissible subtree**:

$$\omega^* = \arg\max_{\omega} \mathrm{AUC}(S_{\mathcal{P}(\mathcal{T}(\omega))})$$

- **Cross-validation based approach**:
  - Linear complexity penalty

$$\widehat{\mathrm{CPAUC}}(S_{\mathcal{P}(\mathcal{T}(\omega))}, \lambda) = \widehat{\mathrm{AUC}}(S_{\mathcal{P}(\mathcal{T}(\omega))}) - \lambda \cdot \#\mathcal{P}(\mathcal{T}(\omega))$$

  - Choose the best λ by K-fold cross validation
- **Structural** AUC **maximization**:
  - $\widetilde{\mathrm{CPAUC}}(S_{\mathcal{P}(\mathcal{T}(\omega))}) = \widehat{\mathrm{AUC}}(S_{\mathcal{P}(\mathcal{T}(\omega))}) - \mathrm{pen}\left(\#\mathcal{P}(\mathcal{T}(\omega)), n\right),$
  - Choice of the penalty driven by a **distribution-free bound** for

$$\mathbb{E}\left[\sup_{\omega:\ \#\mathcal{P}(\mathcal{T}(\omega))=K} |\widehat{\mathrm{AUC}}(S_{\mathcal{P}(\mathcal{T}(\omega))}) - \mathrm{AUC}(S_{\mathcal{P}(\mathcal{T}(\omega))})|\right]$$

Pruning ranking trees - Example

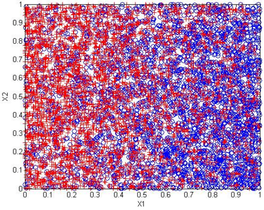
- Suppose that LeafRank is implemented with **at most** k **perpendicular cuts** and $p \in [\underline{p}, \bar{p}] \subset]0, 1[$

Pruning ranking trees - Example

- Suppose that LEAFRANK is implemented with **at most** k **perpendicular cuts** and $p \in [\underline{p}, \bar{p}] \subset]0,1[$
- Set the penalty as

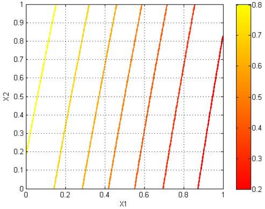
$$\mathrm{pen}(K,n) = \frac{1}{\underline{p}(1-\bar{p})}\sqrt{32\frac{\log\left(16((n+1)q)^{2Kk}\right)+K}{n}}$$

TreeRank in action - Example



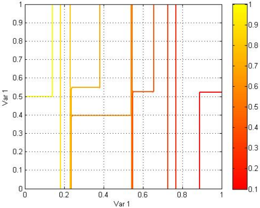
a.

positives in red, negatives in blue.



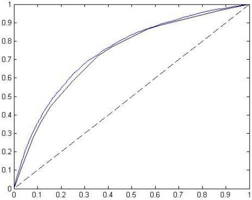
b.

Ideal ordered partition.



c.

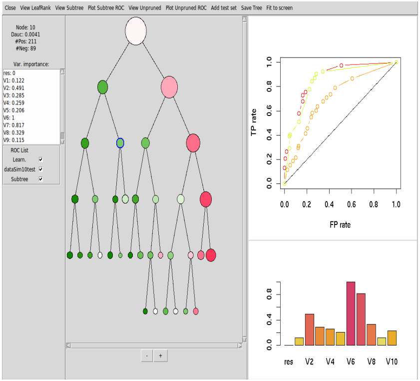
Ordered partition learnt from the training dataset.

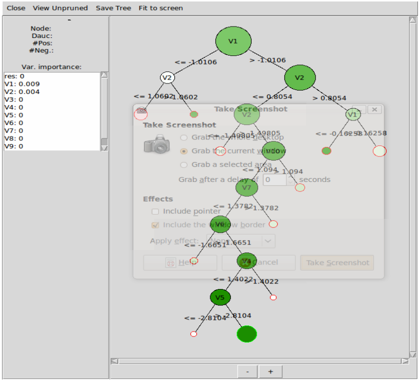


d.

Optimal (blue) and test (black) ROC curves.

TREERANK available at http://treerank.sourceforge.net





Extending the 'aggregation paradigm' to ranking

- In (binary) classification, aggregation boils down to a **(possibly weighted) majority voting scheme**:

$$C_{agg}(X) = sign\left(\sum_{k=1}^{K} \omega_k C_k(X)\right).$$

- **B**ootstrap **agg**regat**ing** techniques, Random Forests, Boosting, *etc.*
- In ranking, the prediction rule is a **linear (pre)order** $\preceq_s$ on $\mathcal{X}$:

$$\forall (x, x') \in \mathcal{X}^2, \;\; x \preceq_s x' \Leftrightarrow s(x) \leq s(x').$$

- Given K preorders on a set $\mathcal{Z}$, $\preceq_1, \ldots, \preceq_K$, how to define a **barycentric preorder**?

Aggregation of binary relations on a finite set

- **A very old issue ...**

Aggregation of binary relations on a finite set

- **A very old issue ...**
  - Voting/social choice theory (18-th century: Condorcet, *etc.*)

- **A very old issue ...**
  - Voting/social choice theory (18-th century: Condorcet, *etc.*)
  - No **ideal** solution (*cf* Arrow's impossibility theorem) to the "consensus problem"

Aggregation of binary relations on a finite set

- **A very old issue ...**
  - Voting/social choice theory (18-th century: Condorcet, *etc.*)
  - No **ideal** solution (*cf* Arrow's impossibility theorem) to the "consensus problem"
  - The **ordinal** view: Condorcet's tournaments, the Hare system, *etc.*

Aggregation of binary relations on a finite set

- **A very old issue ...**
  - Voting/social choice theory (18-th century: Condorcet, *etc.*)
  - No **ideal** solution (*cf* Arrow's impossibility theorem) to the "consensus problem"
  - The **ordinal** view: Condorcet's tournaments, the Hare system, *etc.*
  - The **metric-based** approach: Kemeny, Cailey, Kendall, Spearman, *etc.*

Aggregation of binary relations on a finite set

- **A very old issue ...**
  - Voting/social choice theory (18-th century: Condorcet, *etc.*)
  - No **ideal** solution (*cf* Arrow's impossibility theorem) to the "consensus problem"
  - The **ordinal** view: Condorcet's tournaments, the Hare system, *etc.*
  - The **metric-based** approach: Kemeny, Cailey, Kendall, Spearman, *etc.*
  - **Probabilistic models on** $\mathfrak{S}_K$: Mallows ('57), Fligner Verducci ('86), Lebanon Lafferty ('02), *etc.*
- **... revitalized by new problems**:

Aggregation of binary relations on a finite set

- **A very old issue ...**
  - Voting/social choice theory (18-th century: Condorcet, *etc.*)
  - No **ideal** solution (*cf* Arrow's impossibility theorem) to the "consensus problem"
  - The **ordinal** view: Condorcet's tournaments, the Hare system, *etc.*
  - The **metric-based** approach: Kemeny, Cailey, Kendall, Spearman, *etc.*
  - **Probabilistic models on $\mathfrak{S}_K$**: Mallows ('57), Fligner Verducci ('86), Lebanon Lafferty ('02), *etc.*
- **... revitalized by new problems**:
  - Collaborative filtering

Aggregation of binary relations on a finite set

- **A very old issue ...**
  - Voting/social choice theory (18-th century: Condorcet, *etc.*)
  - No **ideal** solution (*cf* Arrow's impossibility theorem) to the "consensus problem"
  - The **ordinal** view: Condorcet's tournaments, the Hare system, *etc.*
  - The **metric-based** approach: Kemeny, Cailey, Kendall, Spearman, *etc.*
  - **Probabilistic models on** $\mathfrak{S}_K$: Mallows ('57), Fligner Verducci ('86), Lebanon Lafferty ('02), *etc.*
- **... revitalized by new problems**:
  - Collaborative filtering
  - Meta-search engines

Aggregation of binary relations on a finite set

- **A very old issue ...**
  - Voting/social choice theory (18-th century: Condorcet, *etc.*)
  - No **ideal** solution (*cf* Arrow's impossibility theorem) to the "consensus problem"
  - The **ordinal** view: Condorcet's tournaments, the Hare system, *etc.*
  - The **metric-based** approach: Kemeny, Cailey, Kendall, Spearman, *etc.*
  - **Probabilistic models on** $\mathfrak{S}_K$: Mallows ('57), Fligner Verducci ('86), Lebanon Lafferty ('02), *etc.*
- **... revitalized by new problems**:
  - Collaborative filtering
  - Meta-search engines
  - Spam-fighting

- Let $\mathcal{Z} = \{z_1, \ldots, z_K\}$ and $\preceq$ a preorder on $\mathcal{Z}$
- Denote by $\mathcal{R}_{\preceq}(z_k)$ the rank of z_k (*mid-rank* convention)
- Many ways of measuring concordance/agreement between two rankings $\preceq$ and $\preceq'$
  1. **Spearman footrule distance.**

$$d_1(\preceq, \preceq') = \sum_{i=1}^{K} |\mathcal{R}_{\preceq}(z_i) - \mathcal{R}_{\preceq'}(z_i)|.$$

Metric-based aggregation of binary relations on a finite set

- Let $\mathcal{Z} = \{z_1, \ldots, z_K\}$ and $\preceq$ a preorder on $\mathcal{Z}$
- Denote by $\mathcal{R}_{\preceq}(z_k)$ the rank of z_k (*mid-rank* convention)
- Many ways of measuring concordance/agreement between two rankings $\preceq$ and $\preceq'$

1. **Spearman footrule distance.**

$$d_1(\preceq, \preceq') = \sum_{i=1}^{K} |\mathcal{R}_{\preceq}(z_i) - \mathcal{R}_{\preceq'}(z_i)|.$$

2. **Spearman rank-order correlation distance.**

$$d_2(\preceq, \preceq') = \sum_{i=1}^{K} \left(\mathcal{R}_{\preceq}(z_i) - \mathcal{R}_{\preceq'}(z_i)\right)^2$$

Metric-based aggregation of binary relations on a finite set

- Let $\mathcal{Z} = \{z_1, \ldots, z_K\}$ and $\preceq$ a preorder on $\mathcal{Z}$
- Denote by $\mathcal{R}_{\preceq}(z_k)$ the rank of z_k (*mid-rank* convention)
- Many ways of measuring concordance/agreement between two rankings $\preceq$ and $\preceq'$
  1. **Spearman footrule distance.**

$$d_1(\preceq, \preceq') = \sum_{i=1}^{K} |\mathcal{R}_{\preceq}(z_i) - \mathcal{R}_{\preceq'}(z_i)|.$$

  2. **Spearman rank-order correlation distance.**

$$d_2(\preceq, \preceq') = \sum_{i=1}^{K} (\mathcal{R}_{\preceq}(z_i) - \mathcal{R}_{\preceq'}(z_i))^2$$

  3. Kemeny top k-lists, word-metrics on $\mathfrak{S}_K$, ... see Deza Deza ('09)

- Count the number of discording pairs:

$$d_\tau(\preceq, \preceq') = \sum_{i<j} U_{i,j}(\preceq, \preceq'),$$

with

$$\begin{aligned} U_{i,j}(\preceq, \preceq') = \mathbb{I}\{(\mathcal{R}_{\preceq}(z_i) - \mathcal{R}_{\preceq}(z_j))(\mathcal{R}_{\preceq'}(z_i) - \mathcal{R}_{\preceq'}(z_j)) < 0\} \\ + \frac{1}{2}\mathbb{I}\{\mathcal{R}_{\preceq}(z_i) = s_{\preceq}(z_j),\ \mathcal{R}_{\preceq'}(z_i) \neq \mathcal{R}_{\preceq'}(z_j)\} \\ + \frac{1}{2}\mathbb{I}\{\mathcal{R}_{\preceq'}(z_i) = \mathcal{R}_{\preceq'}(z_j),\ \mathcal{R}_{\preceq}(z_i) \neq \mathcal{R}_{\preceq}(z_j)\} \end{aligned}$$

Kendall τ distance

- Count the number of discording pairs:

$$d_\tau(\preceq, \preceq') = \sum_{i<j} U_{i,j}(\preceq, \preceq'),$$

with

$$\begin{aligned} U_{i,j}(\preceq, \preceq') = \mathbb{I}\{(\mathcal{R}_\preceq(z_i) - \mathcal{R}_\preceq(z_j))(\mathcal{R}_{\preceq'}(z_i) - \mathcal{R}_{\preceq'}(z_j)) < 0\} \\ + \frac{1}{2}\mathbb{I}\{\mathcal{R}_\preceq(z_i) = s_\preceq(z_j),\ \mathcal{R}_{\preceq'}(z_i) \neq \mathcal{R}_{\preceq'}(z_j)\} \\ + \frac{1}{2}\mathbb{I}\{\mathcal{R}_{\preceq'}(z_i) = \mathcal{R}_{\preceq'}(z_j),\ \mathcal{R}_\preceq(z_i) \neq \mathcal{R}_\preceq(z_j)\} \end{aligned}$$

- Can be computed in $O((K \log K)/\log\log K)$ time
- Equivalent to the Spearman footrule distance

Median rankings

- Let $\preceq_1, \ldots, \preceq_K$ be a *profile* of rankings on $\mathcal{Z}$

Median rankings

- Let $\preceq_1, \ldots, \preceq_K$ be a *profile* of rankings on $\mathcal{Z}$
- Let $d(.,.)$ be a distance between rankings on $\mathcal{Z}$

Median rankings

- Let $\preceq_1, \ldots, \preceq_K$ be a *profile* of rankings on $\mathcal{Z}$
- Let $d(.,.)$ be a distance between rankings on $\mathcal{Z}$
- A **median ranking** is any ranking $\preceq_{med}$ is any ranking s.t.

$$\sum_{k=1}^{K} d(\preceq_{med}, \preceq_k) = \min_{\preceq} \sum_{k=1}^{K} d(\preceq, \preceq_k)$$

Median rankings

- Let $\preceq_1, \ldots, \preceq_K$ be a *profile* of rankings on $\mathcal{Z}$
- Let $d(.,.)$ be a distance between rankings on $\mathcal{Z}$
- A **median ranking** is any ranking $\preceq_{med}$ is any ranking s.t.

$$\sum_{k=1}^{K} d(\preceq_{med}, \preceq_k) = \min_{\preceq} \sum_{k=1}^{K} d(\preceq, \preceq_k)$$

- **Non uniqueness** in general (ex: $\mathcal{Z} = 1, 2$)

Median rankings

- Let $\preceq_1, \ldots, \preceq_K$ be a *profile* of rankings on $\mathcal{Z}$
- Let $d(.,.)$ be a distance between rankings on $\mathcal{Z}$
- A **median ranking** is any ranking $\preceq_{med}$ is any ranking s.t.

$$\sum_{k=1}^{K} d(\preceq_{med}, \preceq_k) = \min_{\preceq} \sum_{k=1}^{K} d(\preceq, \preceq_k)$$

- **Non uniqueness** in general (ex: $\mathcal{Z} = 1, 2$)
- If $\#\mathcal{Z} = N$, there are

$$\sum_{k=1}^{N} (-1)^k \sum_{m=1}^{k} (-1)^m \binom{k}{m} m^N$$

rankings on Z.

Median rankings

- Let $\preceq_1, \ldots, \preceq_K$ be a *profile* of rankings on $\mathcal{Z}$
- Let $d(.,.)$ be a distance between rankings on $\mathcal{Z}$
- A **median ranking** is any ranking $\preceq_{med}$ is any ranking s.t.

$$\sum_{k=1}^{K} d(\preceq_{med}, \preceq_k) = \min_{\preceq} \sum_{k=1}^{K} d(\preceq, \preceq_k)$$

- **Non uniqueness** in general (ex: $\mathcal{Z} = 1, 2$)
- If $\#\mathcal{Z} = N$, there are

$$\sum_{k=1}^{N} (-1)^k \sum_{m=1}^{k} (-1)^m \binom{k}{m} m^N$$

  rankings on Z.
- NP-hard problems, require use of **meta-heuristics**

Agreggation of ranking trees

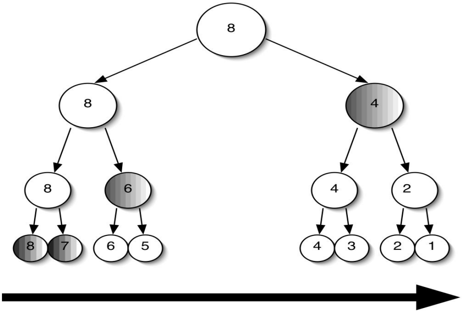
- Discrete maths *vs.* continuous maths ...

Agreggation of ranking trees

- Discrete maths *vs.* continuous maths ...
- In a general setup, existence of a median is an open problem

Agreggation of ranking trees

- Discrete maths *vs.* continuous maths ...
- In a general setup, existence of a median is an open problem
- For a ranking tree, the preorder on $\mathcal{X}$ is induced by an ordering of the terminal leaves (left-right orientation)



Agreggation of ranking trees

- Consider an ensemble of ranking trees $\mathcal{T}_1, \ldots, \mathcal{T}_B$

Agreggation of ranking trees

- Consider an ensemble of ranking trees $\mathcal{T}_1, \ldots, \mathcal{T}_B$
- The terminal leaves of $\mathcal{T}_b$ form a partition $\mathcal{P}_b$ of $\mathcal{X}$

Agreggation of ranking trees

- Consider an ensemble of ranking trees $\mathcal{T}_1, \ldots, \mathcal{T}_B$
- The terminal leaves of $\mathcal{T}_b$ form a partition $\mathcal{P}_b$ of $\mathcal{X}$
- Consider the largest subpartition $\mathcal{P}_B^*$ of the $\mathcal{P}_b$'s

Agreggation of ranking trees

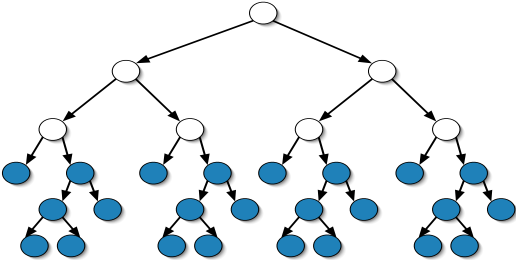
- Consider an ensemble of ranking trees $\mathcal{T}_1, \ldots, \mathcal{T}_B$
- The terminal leaves of $\mathcal{T}_b$ form a partition $\mathcal{P}_b$ of $\mathcal{X}$
- Consider the largest subpartition $\mathcal{P}_B^*$ of the $\mathcal{P}_b$'s
  - Cells of $\mathcal{P}_B^*$ are of the form $\bigcap_{b=1}^{B} \mathcal{C}_b$ with $\mathcal{C}_b \in \mathcal{P}_b$

Agreggation of ranking trees

- Consider an ensemble of ranking trees $\mathcal{T}_1, \ldots, \mathcal{T}_B$
- The terminal leaves of $\mathcal{T}_b$ form a partition $\mathcal{P}_b$ of $\mathcal{X}$
- Consider the largest subpartition $\mathcal{P}_B^*$ of the $\mathcal{P}_b$'s
    - Cells of $\mathcal{P}_B^*$ are of the form $\bigcap_{b=1}^{B} \mathcal{C}_b$ with $\mathcal{C}_b \in \mathcal{P}_b$
    - It $\exists(\mathcal{C}_b, \mathcal{C}) \in \mathcal{P}_b \times \mathcal{P}_B^*$ s.t. $\mathcal{C}_b \subset \mathcal{C}$, then $\mathcal{C}_b = \mathcal{C}$

Agreggation of ranking trees

- Consider an ensemble of ranking trees $\mathcal{T}_1, \ldots, \mathcal{T}_B$
- The terminal leaves of $\mathcal{T}_b$ form a partition $\mathcal{P}_b$ of $\mathcal{X}$
- Consider the largest subpartition $\mathcal{P}_B^*$ of the $\mathcal{P}_b$'s
  - Cells of $\mathcal{P}_B^*$ are of the form $\bigcap_{b=1}^{B} \mathcal{C}_b$ with $\mathcal{C}_b \in \mathcal{P}_b$
  - It $\exists(\mathcal{C}_b, \mathcal{C}) \in \mathcal{P}_b \times \mathcal{P}_B^*$ s.t. $\mathcal{C}_b \subset \mathcal{C}$, then $\mathcal{C}_b = \mathcal{C}$
  - From a computational angle, bind less and less complex ranking trees as one goes along



Agreggation of ranking trees

- Each ranking tree $\mathcal{T}_B$ defines:
  1. a preorder on $\mathcal{P}_B^*$, $\preceq_b$ say

Agreggation of ranking trees

- Each ranking tree $\mathcal{T}_B$ defines:
  1. a preorder on $\mathcal{P}_B^*$, $\preceq_b$ say
  2. a preorder on $\mathcal{X}$, $\preccurlyeq_{s_b}$ say

- Let $\mathcal{C} \neq \mathcal{C}'$ in $\mathcal{P}_B^*$ and $(x, x') \in \mathcal{C} \times \mathcal{C}'$, we have:

$$x \preccurlyeq_{s_b} x' \Leftrightarrow \mathcal{C} \preceq_b \mathcal{C}'$$

- This permits us to define "distances" between $\preccurlyeq_{s_b}$ and $\preccurlyeq_{s_{b'}}$

$$\tilde{d}(\preccurlyeq_{s_b}, \preccurlyeq_{s_{b'}}) \stackrel{def}{=} d(\preceq_b, \preceq_{b'})$$

Probabilistic measures of scoring agreement

- Most agreement measures between rankings arise from nonparametric testing procedures

Probabilistic measures of scoring agreement

- Most agreement measures between rankings arise from nonparametric testing procedures
- Kendall τ between two r.v.'s Z_1 and Z_2: $\widetilde{\tau}(Z_1, Z_2) = 1 - 2d_{\tilde{\tau}}(Z_1, Z_2)$, with:

$$d_{\tilde{\tau}}(Z_1, Z_2) = \mathbb{P}\{(Z_1 - Z_1') \cdot (Z_2 - Z_2') < 0\} + \frac{1}{2}\mathbb{P}\{Z_1 = Z_1',\ Z_2 \neq Z_2'\} + \frac{1}{2}\mathbb{P}\{Z_1 \neq Z_1',\ Z_2 = Z_2'\}.$$

- $\mathrm{AUC}(s)$ and Kendall τ of $(s(X), Y)$ are related:

$$\frac{1}{2}\left(1 - \tilde{\tau}(s(X), Y)\right) = 2p(1-p)\left(1 - \mathrm{AUC}(s)\right) + \frac{1}{2}\mathbb{P}\{s(X) \neq s(X'),\ Y = Y'\}.$$

Probabilistic Kendall τ distance

- Consider $d_{\tilde{\tau}}(s_b(X), s_{b'}(X)) = d_{\tau_X}(\preccurlyeq_{s_b}, \preccurlyeq_{s_{b'}})$. We have:

$$d_{\tau_X}(\preccurlyeq_{s_b}, \preccurlyeq_{s_{b'}}) = 2\sum_{k<l} \mu(\mathcal{C}_k^*)\mu(\mathcal{C}_l^*)U_{k,l}(\preceq_b, \preceq_{b'}),$$

where $\mathcal{P}_B^* = \{\mathcal{C}_k^*\}$ and $\mu(dx)$ denotes X's marginal distribution.
$\Rightarrow$ "weighted rate of discording pairs

Probabilistic Kendall τ distance

- Consider $d_{\tilde{\tau}}(s_b(X), s_{b'}(X)) = d_{\tau_X}(\preccurlyeq_{s_b}, \preccurlyeq_{s_{b'}})$. We have:

$$d_{\tau_X}(\preccurlyeq_{s_b}, \preccurlyeq_{s_{b'}}) = 2\sum_{k<l} \mu(\mathcal{C}_k^*)\mu(\mathcal{C}_l^*)U_{k,l}(\preceq_b, \preceq_{b'}),$$

  where $\mathcal{P}_B^* = \{\mathcal{C}_k^*\}$ and $\mu(dx)$ denotes X's marginal distribution.
  $\Rightarrow$ "weighted rate of discording pairs

- Analogous relationships for Spearman's distances

Probabilistic Kendall τ distance and AUC criterion

Scoring functions close in Kendall sense have close AUC:

Lemma (Clémençon, 2010)

Let $p = \mathbb{P}\{Y = +1\} \in (0, 1)$. For any scoring functions s_1 and s_2 on $\mathcal{X}$:

$$|\mathrm{AUC}(s_1) - \mathrm{AUC}(s_2)| \leq \frac{1 - \tau_X(\preccurlyeq_{s_1}, \preccurlyeq_{s_2})}{4p(1-p)}.$$

The reverse assertion is not true. However...

Lemma (Clémençon, 2010)

Assume that $\eta(X)$ is continuous and $\epsilon \in (0, 1/2)$ s.t. $\epsilon \leq \eta(X) \leq 1 - \epsilon$ a.s., and $c < \infty$ and $a \in (0, 1)$ s.t. $\forall x \in \mathcal{X}$, $\mathbb{E}\left[|\eta(X) - \eta(x)|^{-a}\right] \leq c$. Then, we have for all (s, s^*):

$$1 - \tau_X(\preccurlyeq_{s^*}, \preccurlyeq_s) \leq C \cdot \left(\mathrm{AUC}^* - \mathrm{AUC}(s)\right)^{a/(1+a)},$$

with $C = 2 \cdot \max\{c^{1/(1+a)},\ p(1-p)/\epsilon^2\}$.

Statistical version of the probabilistic Kendall τ distance

- Based on a sample of i.i.d. copies of X, simply replace the $\mu(\mathcal{C}_k^*)$'s by their empirical counterparts $\Rightarrow \widehat{d}_{\tau_X}(\preccurlyeq_{s_1}, \preccurlyeq_{s_2})$

Statistical version of the probabilistic Kendall τ distance

- Based on a sample of i.i.d. copies of X, simply replace the $\mu(\mathcal{C}_k^*)$'s by their empirical counterparts $\Rightarrow \widehat{d}_{\tau_X}(\preccurlyeq_{s_1}, \preccurlyeq_{s_2})$
- Alternately, $\widehat{d}_{\tau_X}(\preccurlyeq_{s_1}, \preccurlyeq_{s_2})$ may be represented by a U-statistic with kernel

$$
\begin{aligned}
K(x, x') = \mathbb{I}\{(s_1(x) - s_1(x')) \cdot (s_2(x) - s_2(x')) < 0\} \\
+ \frac{1}{2}\mathbb{I}\{s_1(x) = s_1(x'),\ s_2(x) \neq s_2(x')\} \\
+ \frac{1}{2}\mathbb{I}\{s_1(x) \neq s_1(x'),\ s_2(x) = s_2(x')\}.
\end{aligned}
$$

- Required results for U-processes are available, see Clémençon, Lugosi Vayatis (2008)

Some theoretical background for ranking aggregation

- randomized scoring function based on a training dataset $\mathcal{D}_n$

$$S_{\mathcal{D}_n}(x, Z),$$

  where the r.v. Z is drawn conditionally to $\mathcal{D}_n$, describes the randomization mechanism.

- Build a profile of scoring functions by drawing m i.i.d. copies of Z

$$S_{\mathcal{D}_n}(x, Z_j),\ j = 1,\ \ldots,\ m$$

- Let $\mathcal{S}_0$ be a set of scoring functions. Consider a (supposedly existing) median scoring function $\bar{S}_m$ w.r.t. d_{τ_X}

$$\sum_{j=1}^{m} d_{\tau_X}(\preccurlyeq_{\bar{S}_m}, \preccurlyeq_{\mathbf{s}_{\mathcal{D}_n}(.,Z_j)}) = \inf_{s\in\mathcal{S}_0} \sum_{j=1}^{m} d_{\tau_X}(\preccurlyeq_s, \preccurlyeq_{\mathbf{s}_{\mathcal{D}_n}(.,Z_j)})$$

Some theoretical background for ranking aggregation

Aggregation preserves AUC consistency and the learning rate

Theorem (Clémençon, 2010)

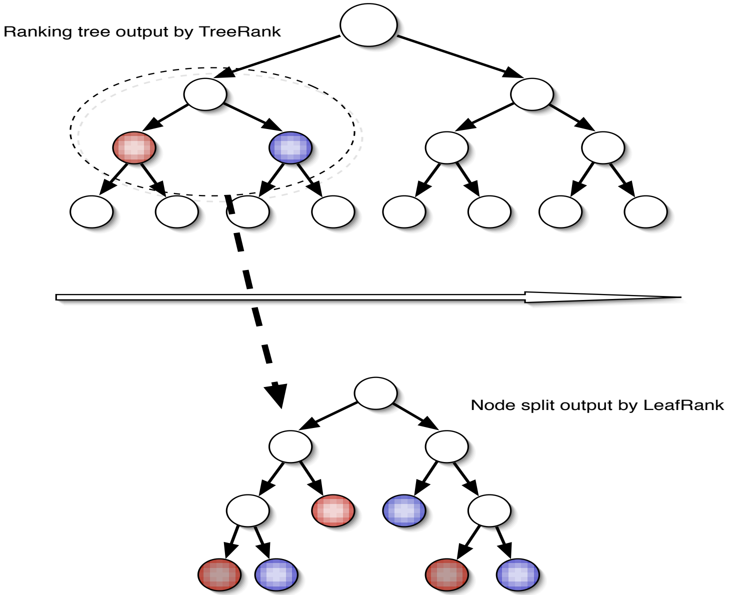
If $S_{\mathcal{D}_n}(x, Z)$ is (strongly) AUC-consistent, so is the median $\bar{S}_m$.
The result still holds true when median computation is performed using $\widehat{d}_{\tau_X}$ provided that $\mathcal{S}_0$ is of finite VC dimension.
If v_n is the rate of $S_{\mathcal{D}_n}(x, Z)$, the rate of the aggregated rule is $O_{\mathbb{P}}(\max\{n^{-1/2},\ v_n\})$.

Feature randomization in TreeRank

$\mathcal{FR}_1$: *At each node* (d, k) *of the master ranking tree* $\mathcal{T}_D$*, draw at random a set of* $q_0 \leq q$ *indexes* $\{i_1, \ldots, i_{q_0}\} \subset \{1, \ldots, q\}$*. Implement the* LeafRank *splitting procedure based on the descriptor* $(X_{i_1}, \ldots, X_{i_{q_0}})$ *to split the cell* $C_{d,k}$*.*

$\mathcal{FR}_2$: *For each node* (d, k) *of the master ranking tree* $\mathcal{T}_D$*, at each node of the cost-sensitive classification tree describing the split of the cell* $\mathcal{C}_{d,k}$ *into two children, draw at random a set of* $q_1 \leq q$ *indexes* $\{j_1, \ldots, j_{q_1}\} \subset \{1, \ldots, q\}$ *and perform an axis-parallel cut using the descriptor* $(X_{j_1}, \ldots, X_{j_{q_1}})$*.*

Feature randomization in TreeRank



Ranking Forest - the Algorithm

1. **Parameters.** B number of bootstrap replicates, n^* bootstrap sample size, TreeRank tuning parameters (depth D and presence/absence of pruning) $\mathcal{FR}$ feature randomization strategy, d pseudo-metric.

2. **Bootstrap profile makeup.**
   1. (Resampling step.) Build B independent bootstrap samples $\mathcal{D}_1^*, \ldots, \mathcal{D}_B^*$, by drawing with replacement $n^* \times B$ pairs among the original training sample $\mathcal{D}$.
   2. (Randomized TreeRank.) For $b = 1, \ldots, B$, run TreeRank combined with the feature randomization method $\mathcal{FR}$ based on the sample $\mathcal{D}_b^*$, yielding the ranking tree $\mathcal{T}_b^*$, related to the partition $\mathcal{P}_b^*$ of the space $\mathcal{X}$.

3. **Aggregation.** Compute the largest subpartition partition $\mathcal{P}^* = \bigcap_{b=1}^{B} \mathcal{P}_b^*$. Let $\preceq_b^*$ be the ranking of $\mathcal{P}^*$'s cells induced by $\mathcal{T}_b^*$, $b = 1, \ldots, B$. Compute a median ranking $\preceq^*$ related to the bootstrap profile $\Pi^* = \{\preceq_b^* : 1 \leq b \leq B\}$ with respect to the metric d.

Ranking stability

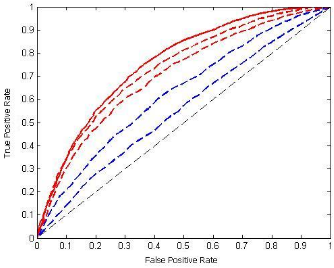
- Ranking algorithm $\mathbf{S} : \mathcal{D}_n \mapsto S_{\mathcal{D}_n}$
- A natural way of measuring (in)stability

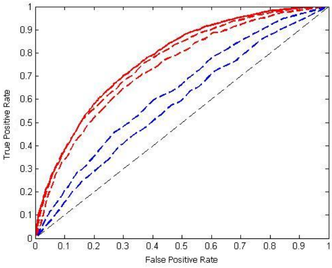
$$\mathbf{Stab}_n(\mathbf{S}) = \mathbb{E}\left[d_{\tau_X}\left(\preccurlyeq_{\mathbf{s}_{\mathcal{D}}}, \preccurlyeq_{\mathbf{s}_{\mathcal{D}'}}\right)\right],$$

- A bootstrap estimate

$$\widehat{\mathbf{Stab}}_n(\mathbf{S}) = \frac{2}{B(B-1)} \sum_{1 \leq b < b' \leq B} \widehat{d}_{\tau_X}\left(\preccurlyeq_{\mathbf{s}_{\mathcal{D}^*_b}}, \preccurlyeq_{\mathbf{s}_{\mathcal{D}^*_{b'}}}\right).$$

Numerical experiments





Conclusion

- Empirically, aggregation combined with randomization enhances ROC accuracy and increases stability both at the same time
- No theoretical grounds for supporting this fact, see Friedman & Hall (2007) in the context of regression
- In progress:
  - Convexification of the median issue
  - boosting ranking trees through a weighted consensus

Elements of Bibliography

- Tree-based ranking methods. S. Clémençon & N. Vayatis (2008). IEEE Information Theory
- Ranking the Best Instances. S. Clémençon & N. Vayatis (2007). JMLR.
- Ranking and Empirical Minimization of U-statistics. S. Clémençon, G. Lugosi & N. Vayatis (2008). Annals of Statistics.
- Overlaying classifiers: a practical approach for optimal ranking. S. Clémençon & N. Vayatis (2010). Constructive Approximation.
- On Partitioning Rules for Bipartite Ranking. S. Clémençon & N. Vayatis (2009). In JMLR&PW
- Empirical maximization performance based on linear rank statistics. S. Clémençon & N. Vayatis (2008). NIPS'08.
- On AUC maximization and the two-sample problem. S. Clémençon, M. Depecker & N. Vayatis (2009). NIPS'09.
- Adaptive partitioning schemes for bipartite ranking - How to grow and prune a ranking tree. S. Clémençon, M. Depecker & N. Vayatis (2010). Machine-Learning
- Empirical maximization performance based on linear rank statistics. S. Clémençon (2010). IEEE PAMI.
- Kantorovich distances between rankings with applications to rank aggregation. S. Clémençon & J. Jakubowicz (2010). ECML'10.