La

Label Ranking

Label Ranking (LR) corresponds to the problem of learning a hypothesis th
to rankings over a finite set of labels.

Applications: pattern recognition, web advertisement and document cat

**The importance of LR has spurred the development of applied CS appr
task. However, formal theoretical guarantees for this problem are missi**

Label Ranking through Nonparametric Reg

In LR, the learner observes labeled examples $(x, \sigma) \in \mathbb{X} \times \mathfrak{S}_k$ and the goa
$h : \mathbb{X} \to \mathfrak{S}_k$ from instances to rankings of labels with low error on future

According to practical applications, a natural way to represent preferences
alternatives through a real-valued score function.

Distribution-free Nonparametric LR: Let $\mathbb{X} \subseteq \mathbb{R}^d$, $[k]$ be a set of l
functions from $\mathbb{X}$ to $[0,1]^k$ and $\mathcal{D}_x$ be an arbitrary distribution over
distribution $\mathcal{E}$ over $\mathbb{R}^k$. Let m be an unknown target function in $\mathcal{C}$.

An example oracle $\mathrm{Ex}(m, \mathcal{E})$ with complete rankings, works as follows:

- $x \sim \mathcal{D}_x$ and $\xi \sim \mathcal{E}$ independently,
- $\sigma = \mathrm{argsort}(m(x) + \xi)$ and
- it returns a labeled example $(x, \sigma) \in \mathbb{X} \times \mathfrak{S}_k$.

In the noiseless case ($\xi = 0$ almost surely), we simply write $\mathrm{Ex}(m)$.

abel Ranking through Nonpar

Dimitris Fotakis **Alkis Ka**

NT

Main R

hat maps feature vectors

egorization.

oaches for tackling this
ng.

ression

l is to learn a hypothesis
unseen examples.

s is to evaluate individual

abels, $\mathcal{C}$ be a class of
$\mathbb{X}$. Consider a noise

We give the first theoretical guarantees for the
decision trees and random forests in the noisele
for robustness of shallow decision trees and ra

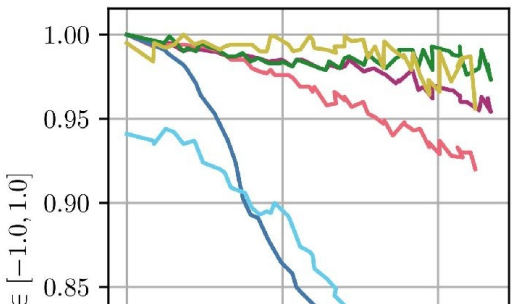
Noiseless Setting

Under sparsity and approximate submodularity
Decision Trees via Level-Splits that draws $\widetilde{O}\left(\log\right.$
$\mathrm{Ex}(m)$ and, in $\mathrm{poly}_{C,r}(d, k, 1/\epsilon)$ time, computes
$\{0,1\}^d \to \mathbb{S}_k$ which, with probability 99%, satisf

We further obtain results for random forests and

Noisy Setting The noise distribution $\mathcal{E}$ satisfies

(i) the α-inconsistency property if $\mathbf{E}_{x\sim\mathcal{D}_x}\left[\mathbf{Pr}_{\xi\sim\mathcal{E}}\right.$
(ii) the β-k_τ gap property if $\mathbf{E}_{x\sim\mathcal{D}_x}\,\mathbf{E}_{\xi\sim\mathcal{E}}[k_\tau(h(x)$



ametric Regression ICML 20

lavasis Eleni Psaroudaki

UA

esults

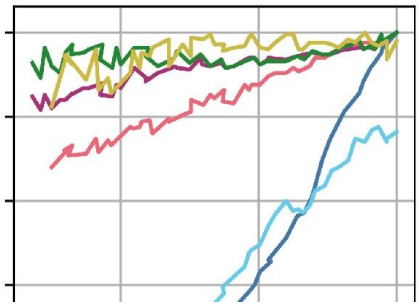
Label Ranking problem for algorithms based on
ss setting and extensive experimental evidence
ndom forests in the noisy case.

conditions, there exists an algorithm based on
$g(d) \cdot \mathrm{poly}_{C,r}(k/\epsilon)\big)$ independent samples from
s a set of splits S_n and an estimate $h^{(n)}(\cdot\,; S_n)$:
ies $\mathbf{E}_{x\sim\mathcal{D}_x}\left[\Delta_{\mathrm{Spearman}}(h(x), h^{(n)}(x; S_n))\right] \leq \epsilon$.

l for the Breiman's criterion.

for some $\alpha \in [0,1]$ and $\beta \in [-1,1]$:

$[h(x) \neq \sigma]] \leq \alpha$, and
$, \sigma))] = \beta$.



Statistical Learning in ra

$h^\star = \mathrm{a}$

The learner is given i.i.c
pothesis $\widehat{h} : \mathbb{X} \to \mathbb{S}_k$ fr
$\mathbf{Pr}_{x\sim\mathcal{D}_x}[\widehat{h}(x) \neq h^\star(x)]$ a

Related Work What is
Clemencon et al. (2018)

C

The model Consider ar
over features. Let $y : [$
$[0,1]^k$. Each sample $(x,$

- Draw $x \in \mathbb{X}$ from $\mathcal{D}_x$
- Draw $q(x)$-biased coi
- If $c_i > 0$, set $y_i = m_i($
- Compute $\sigma = \mathrm{argsort}($

Assumptions We wor
$x \in \mathbb{R}^d$. For any $1 \leq i <$

1. **(Strict Stochastic Tran**
$(p_{i,i}(x) > 1/2 \wedge p_{i,i}(x$

Problem Formulation

nking metric Δ and incomplete rankings Consider the median problem

$$\text{rgmin}_h \mathop{\mathbf{E}}_{(x,\sigma)} [\Delta(h(x), \sigma)] , \quad \text{where} \quad (x, \sigma) \sim \text{Ex}(m, \mathcal{E}) .$$

l. samples from the oracle $\text{Ex}(m, \mathcal{E}, \mathcal{M})$ and its goal is to output a hy- om some hypothesis class $\mathcal{H}$ such that, with high probability, the error gainst the median $h^\star$ is small.

s the optimal solution $h^\star$? When does it exist? See Korba et al. (2017); ; Clemencon and Korba (2018); Clemencon and Vogel (2020).

onditions for Theoretical Guarantees

underlying score hypothesis $m : \mathbb{X} \to [0,1]^k$ and let $\mathcal{D}_x$ be a distribution $] \to [0,1] \cup \{\star\}$ and consider the survival probabilities vector $q : \mathbb{X} \to$ $) \sim \text{Ex}(m, \mathcal{E}, q)$ is generated as follows:

and $\xi \in [-\frac{1}{4}, \frac{1}{4}]^k$ from $\mathcal{E}$.

ns $c \in \{-1, +1\}^k$.

$) + \xi_i$, else $y_i = \star$.

$y)$, ignoring the $\star$ symbol.

k with $\Delta = \Delta_{KT}$. Let $p_{ij}(x) = \mathbf{Pr}_{\xi \sim \mathcal{E}}[m_i(x) + \xi_i > m_j(x) + \xi_j | x]$ for $j \leq k$, we assume that the following hold.

sitivity) For any $x \in \mathbb{R}^d$ and any $u \in [k]$, we have that $p_{ij}(x) \neq 1/2$ and $) > 1/2) \Rightarrow p_{ij}(x) > 1/2$

Problem Formulation

Computational Learning in ranking metric Δ The learner is given i.i.d. $\mathrm{Ex}(m, \mathcal{E})$ and its goal is to *efficiently* output a hypothesis $\widehat{h} : \mathbb{X} \to \mathbb{S}_k$ suc bility, the error $\mathbb{E}_{x\sim\mathcal{D}_x}[\Delta(\widehat{h}(x), h(x))]$ is small.

In practice This problem is mainly solved using decision trees and ra theoretical guarantees were not known for this task.

Conditions for Theoretical Guarantee

Set $\mathbb{X} = \{0,1\}^d$ and the regression function $m : \{0,1\}^d \to [0,1]^k$ with m be the distribution over features. We assume that the following hold for

1. **(Sparsity)** $m_j : \{0,1\}^d \to [0,1]$ is r-sparse, i.e., it depends on r out of d
2. **(Approximate Submodularity)** The mean squared error $\widetilde{L}_j$ of m_j is C-approximate-submodular, i.e., for any $S \subseteq T \subseteq [d]$, $i \in [d]$, it holds t

$$\widetilde{L}_j(T) - \widetilde{L}_j(T \cup \{i\}) \leq C \cdot \Big(\widetilde{L}_j(S) - \widetilde{L}_j(S \cup \{i\})$$

The **Mean Squared Error** (MSE) of a function $f : \{0,1\}^d \to [0,1]$ is equal

$$\widetilde{L}(f, S) = \mathbb{E}_{x\sim\mathcal{D}_x}\Big[\big(f(x) - \mathbb{E}_{w\sim\mathcal{D}_x}[f(w)|w_S = x_S]\big)$$

where x_S is the sub-vector of x, where we observe only the coordinate and $x_S \in \{0,1\}^{|S|}$.

samples from the oracle
:h that, with high proba-

ndom forests; however,

s

$= (m_1, \ldots, m_k)$. Let $\mathcal{D}_x$
any $j \in [k]$.

coordinates.

hat

$\Big)$.

to

$^2\Big]$,

s with indices in $S \subseteq [d]$

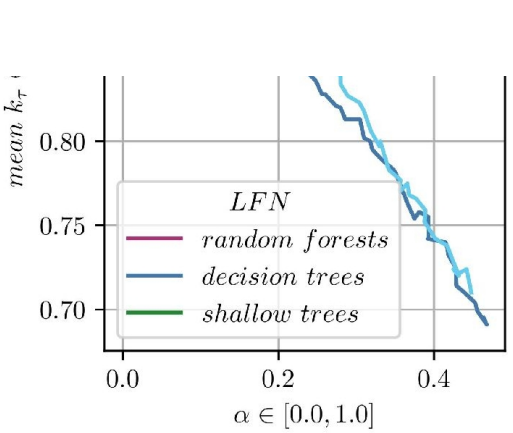


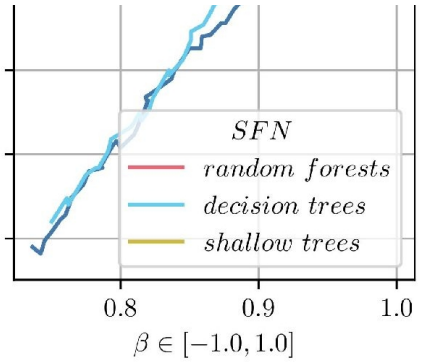
Figure 1. Experimental results in terms of mean

Statistical Label Rank

Distribution-free Nonparametric Incomplete
class of functions from $\mathbb{X}$ to $[0,1]^k$ and $\mathcal{D}_x$ be ar
distribution $\mathcal{E}$ over $\mathbb{R}^k$. Let m be an unknown
mechanism that given a tuple $(x, y) \in \mathbb{X} \times \mathbb{R}^k$ g

An example oracle $\mathrm{Ex}(m, \mathcal{E}, \mathcal{M})$ with incomple

- $x \sim \mathcal{D}_x$ and $\xi \sim \mathcal{E}$ independently,
- $y = m(x) + \xi$,
- $\sigma = \mathcal{M}(x, y)$ and
- it returns a labeled example $(x, \sigma) \in \mathbb{X} \times \mathbb{S}_{\leq}$



KT coefficient for different noise distributions $\mathcal{E}$.

ing with incomplete rankings

LR: Let $\mathbb{X} \subseteq \mathbb{R}^d$, $[k]$ be a set of labels, $\mathcal{C}$ be a
arbitrary distribution over $\mathbb{X}$. Consider a noise
target function in $\mathcal{C}$. Let $\mathcal{M}$ be a randomized
enerates an incomplete ranking $\mathcal{M}(x, y) \in \mathbb{S}_{\leq k}$.

te rankings, works as follows:

k.

2. (Tsybakov's Noise Co
random feature $x \sim \mathcal{I}$
3. (Deletion Tolerance) T
the survival probabilit

Let $\epsilon, \delta \in (0, 1)$. Consi
Under conditions (1)-(3

samples from $\mathrm{Ex}(m, \mathcal{E}$
$h^\star(x)]$ is, with probabil

where $h^\star$ is the optima

$$h^\star(x; i$$

and $L^\star_{i,j}$ is the loss of
constant depending or

$$L_{i,j}(g$$

ndition) There exists $a \in [0,1]$ and $B > 0$ so that the probability that a $_x$ satisfies $\left|p_{ij}(x) - 1/2\right| < 2t$, is at most $B \cdot t^{a/(1-a)}$ for all $t \geq 0$.

There exists $\phi \in (0,1]$ so that $q_{i,j}(x) \geq \phi$ for any $x \in \mathbb{R}^d$, where $q_{i,j}(x)$ is y of the pair $i < j$ in x.

Main Results

der a hypothesis class $\mathcal{G}$ of binary classifiers with finite VC dimension. 3) with parameters a, B, ϕ, there exists an algorithm that draws

$$n = \widetilde{O}\left(\frac{k^{\frac{4(1-a)}{a}}}{\mathrm{poly}_a(\phi \cdot \epsilon)} \cdot \max\left\{\log\left(\frac{k}{\delta}\right), \mathrm{VC}(\mathcal{G})\right\}\right)$$

$, q)$ and computes an estimate $\widehat{h} : \mathbb{R}^d \to \mathbb{S}_k$ so that $\mathbf{Pr}_{x\sim\mathcal{D}_x}[\widehat{h}(x) \neq$ ity $1-\delta$, at most

$$\frac{C_{a,B}}{\phi^2}\left(2\sum_{i<j}\left(\inf_{g\in\mathcal{G}} L_{i,j}(g) - L^\star_{i,j}\right)^a\right) + \epsilon\,,$$

l predictor given by

$$) = 1 + \sum_{j\neq i} 1\{p_{ij}(x) < 1/2\} = 1 + \sum_{j\neq i} 1\{h^\star_{ij}(x) = -1\}$$

the binary Bayes classifiers $h^\star_{i,j}$ for $1 \leq i < j \leq k$, where $C_{a,B}$ is a a, B. In particular,

$$) = \mathop{\mathbf{E}}_{(x,\sigma)\sim\mathrm{Ex}(m,\mathcal{E},q)}[g(x) \neq \mathrm{sgn}(\sigma(i) - \sigma(j))|\sigma \ni i,j]\,.$$