A Multiclass Classification Approach to Label Ranking

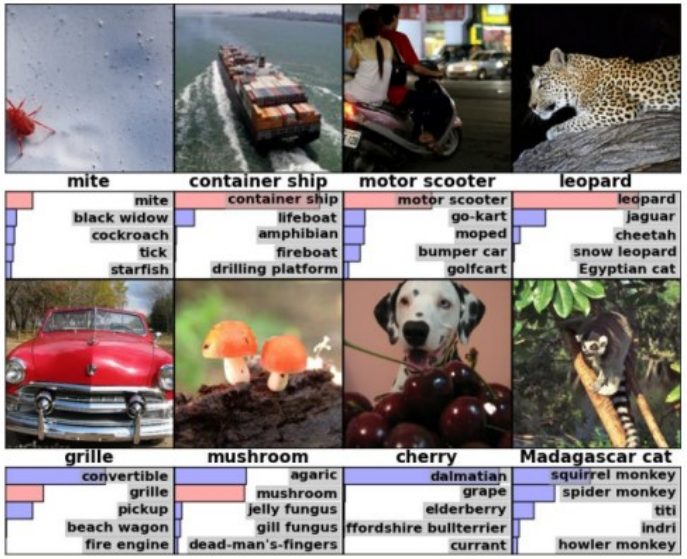
Stephan Clémençon[2] **Robin Vogel**[1,2]

[1] IDEMIA, [2] Télécom Paris

24/03/2020

Outline

Introduction



[Krizhevsky et al., 2012]

Classification:
Introduce a random pair $(X, Y) \sim P$,
and predict a label $Y \in \mathcal{Y} = \{1, \dots, K\}$,
from the features $X \in \mathcal{X} = \mathbb{R}^q$ with $q \geq 1$,
with a classifier $g : \mathcal{X} \to \mathcal{Y}$ from a family $\mathcal{G}$.

For hard problems, one returns a list of the **most likely labels**
for an observation $x \in \mathcal{X}$, which concerns many applications.
e.g. biometrics, search engines, ...

One often uses intermediate values of a classification model,
to derive an ordering on $\mathcal{Y}$, *e.g.* softmax probabilities,
evaluates it with a performance indicator, *e.g.* precision at top-k.

Question: How to **explicitly** learn ordered labels from classif. data ?

From classification to label ranking

Introduce the **posterior probabilities** $\eta(x) := (\eta_1(x), \ldots, \eta_K(x))$, with $\eta_k(x) := \mathbb{P}\{Y = k | X = x\}$, and μ the distribution of X, then (μ, η) characterizes P.

The classification loss writes: $L(g) := \mathbb{P}\{g(X) \neq Y\}$, and its **Bayes classifier** g^* writes: $g^*(x) := \underset{k \in \{1,\ldots,K\}}{\arg\max}\ \eta_k(x)$.

Tasks and targets: by order of difficulty:

· Classification: the maximum of the $\eta_k(x)$'s, in $\{1, \ldots K\}$.

· **Label Ranking** (LR): a decreasing order of $\eta(x)$, in $\mathfrak{S}_K$.

· Conditional density estimation: the vector $\eta(x)$, in $\mathbb{R}^K$.

For $x \in \mathcal{X}$, introduce $\sigma_x^* \in \mathfrak{S}_K$, *s.t.*:

$$\eta_{\sigma_x^{*-1}(1)} > \eta_{\sigma_x^{*-1}(2)} > \cdots > \eta_{\sigma_x^{*-1}(K)},$$

then $g^*(x) = \sigma_x^{*-1}(1)$.

Our contributions

LR is related to **Ranking Median Regression** (RMR). We provide:

- **A description** of the relationship between RMR and LR,
- **A rationale** on the one-vs-one (OVO) approach for LR,
- **Fast convergence proofs** for the OVO approach to LR,
- As a corollary, the **first generalization bound** on an OVO classifier,
- and a **learning bound** on the top-*k* error.

Outline

Label Ranking

Label Ranking: Choose a **ranking rule** $s : \mathcal{X} \to \mathfrak{S}_K$ that minimizes:

$$R(s) := \mathbb{E}[d(s(X), \sigma_X^*)],$$

where $d : \mathfrak{S}_K \times \mathfrak{S}_K \to \mathbb{R}_+$ symmetric and $d(\sigma, \sigma) = 0, \forall \sigma \in \mathfrak{S}_K$.
It is different from classif., unless $d(\sigma, \sigma') = \mathbb{I}\{\sigma^{-1}(1) \neq \sigma'^{-1}(1)\}$.

Many sensible candidates for $d(\sigma, \sigma')$, *e.g.*:

· the error: $\mathbb{I}\{\sigma \neq \sigma'\}$,

· the Hamming distance: $\sum_{k=1}^{K} \mathbb{I}\{\sigma(k) \neq \sigma'(k)\}$,

· the Kendall τ distance d_τ: $\sum_{i<j} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\}$,

Results on $\mathbb{I}\{\sigma \neq \sigma'\}$ can be generalized to most distances, since:

$$d(\sigma, \sigma') \leq \max_{\tau, \tau' \in \mathfrak{S}_K^2} d(\tau, \tau') \times \mathbb{I}\{\sigma \neq \sigma'\}.$$

LR differs from RMR by assumptions on σ_X^* and available data.

Ranking median regression (RMR) (1/2)

Ranking median regression (RMR):

Given a random pair $(X, \Sigma) \sim P$, predict a permutation $\Sigma \in \mathfrak{S}_K$, from $X \in \mathcal{X}$, with a ranking rule $s : \mathcal{X} \to \mathfrak{S}_K$, selected in a family $\mathcal{S}$.

See *e.g.*[Vembu and Gärtner, 2010, Tsoumakas et al., 2009].

In practice, one seeks to find s that nearly minimizes the risk:

$$R(s) := \mathbb{E}[d(s(X), \Sigma)]. \tag{1}$$

Introduce $p_{i,j}(x) = \mathbb{P}\{\Sigma(i) < \Sigma(j) \mid X = x\}$ for $1 \leq i < j \leq K$.

Assumption 1 (Strict Stochastic Transitivity (SST))

For all $x \in \mathbb{R}^q$, *we have:* $\forall (i, k, l) \in \{1, \ldots, K\}^3$, $p_{i,j}(x) \neq 1/2$ *and*

$$p_{i,j}(x) > 1/2 \quad \textit{and} \quad p_{j,k}(x) > 1/2 \quad \Rightarrow \quad p_{i,k}(x) > 1/2.$$

Under Assumption 1 when $d = d_\tau$, the minimizer of Eq. (1) is s^*_X, with:

$$s^*_X(k) := 1 + \sum_{l \neq k} \mathbb{I}\{p_{k,l}(X) < 1/2\} \qquad \text{for any } k \in \{1, \ldots, K\}.$$

Ranking median regression (RMR) (2/2)

Under a complexity assumption on $\mathcal{S}$, with n the size of the sample, *i.e.* the number of i.i.d. observations of P, one can derive generalization bounds in $O(n^{-1/2})$. [Clémençon et al., 2018].

Bounds in $O(n^{-1})$ are derived with an additional noise assumption:

Assumption 2 (Noise condition)

We have:

$$H = ess\inf \min_{i<j} |p_{i,j}(X) - 1/2| > 0.$$

Assumption 2 resembles **Massart's condition** for classification.

Under this condition, guarantees on the risk imply guarantees for the probability of error, since:

$$\mathbb{P}_X \{s(X) \neq s^*_X\} \leq (1/H) \times (R(s) - R(s^*_X)).$$

LR as RMR (1/2)

LR: Label Y and predict ranking σ^*_X,
RMR: Ordering Σ and predict ranking s^*_X.

$\rightarrow$ Need to model the relationship between Y and Σ.

Definition 3 (Conditional Bradley-Terry-Luce-Plackett (BTLP))

The conditional distribution of Σ^{-1} given X is defined by recursion:
Given a **hidden preference vector** $w(X) = (w_1(X), \ldots, w_K(X))$,
set $S_1 := \{1, \ldots, K\}$, $S_k := S_1 \setminus \{\Sigma^{-1}(1), \ldots, \Sigma^{-1}(k-1)\}$ and:

$$\Sigma^{-1}(k) \sim \mathcal{M}\left(1, \left\{ \frac{w_l(X)}{\sum_{m \in S_k} w_m(X)} \;\middle|\; l \in S_k \right\}\right),$$

with $\mathcal{M}(n, p)$ is the multinomial dist. We write $\Sigma \sim \text{BTLP}(w(X))$.

Point of view for LR:
RMR where Σ follows a BTLP model with pref. vector η,
and we observe the partial information $Y = \Sigma^{-1}(1)$.

LR as RMR (2/2)

Lemma 4

Let (X, Y) a random pair on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$.
One may build a r.v. $\Sigma \in \mathfrak{S}_K$, $\Sigma \sim BTLP(\eta(X))$ and $Y = \Sigma^{-1}(1)$ a.s.

Under this model:

$$p_{k,l}(X) = \mathbb{P}\{\Sigma(k) < \Sigma(l) \mid X\} = \frac{\eta_k(X)}{\eta_k(X) + \eta_l(X)} =: \eta_{k,l}(X),$$

hence, SST is satisfied as soon as the $\eta_k(X)$'s are distinct.

Our knowledge of RMR implies:

$$\sigma^*_X(k) = 1 + \sum_{l \neq k} \mathbb{I}\{\eta_{k,l}(X) < 1/2\},$$

which rewrites with Bayes classifiers $g^*_{k,l}(x) := 2\mathbb{I}\{\eta_{k,l}(x) \geq 1/2\} - 1$ for the problem of separating class k from l.

Other approaches to RMR with partial information:
[Korba et al., 2018, Brinker and Hüllermeier, 2019].

Outline

One-Versus-One for classification

Many algorithms are tailored for binary classification, *i.e.* SVMs.

The One-vs-One (OVO) approach extends them to multiclass classif., by running them $K(K-1)/2$ times to separate l from k for any $k < l$, and output the one which won the most duels,
see *e.g.* [Allwein et al., 2000, Wu et al., 2004].

A justification for OVO is that $g^*(x) = \underset{k \in \{1,\dots,K\}}{\arg\max}\ N_k^*(x)$, with:

$$N_k^*(x) = \sum_{l<k} \mathbb{I}\left\{g_{l,k}^*(x) = +1\right\} + \sum_{k<l} \mathbb{I}\left\{g_{k,l}^*(x) = -1\right\}.$$

Our knowledge of the relationship between RMR and LR gives:

$$\sigma_X^*(k) = 1 + \sum_{l \neq k} \mathbb{I}\{g_{k,l}^*(X) = -1\}.$$

which is related to the Copeland score, see [Copeland, 1951].

One-Versus-One for Label Ranking (1/2)

Let $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P$ and $Y_{k,l,i} = \mathbb{I}\{Y_i = l\} - \mathbb{I}\{Y_i = k\}$, for any $k < l$ compute a minimizer $\widehat{g}_{k,l}$ of the empirical risk over $\mathcal{G}$:

$$\widehat{L}_{k,l}(g) := \frac{1}{n_k + n_l} \sum_{i:Y_i \in \{k,l\}} \mathbb{I}\left\{g(X_i) \neq Y_{k,l,i}\right\},$$

approximates $L_{k,l}(g) := \mathbb{P}\{g(X) \neq Y_{k,l} \mid Y \in \{k,l\}\}$, $L^*_{k,l} := L_{k,l}(g^*_{k,l})$.

Proposition 5

Under the assumptions:

· *$\mathcal{G}$ has finite VC-dimension V,*

· *there exists $\epsilon > 0$ s.t. for all $k \neq l$ and $x \in \mathcal{X}$, $\eta_k(x) + \eta_l(x) > \epsilon$,*

· *there exists $\alpha \in [0,1]$ and $B > 0$, s.t.: for all $k \neq l$ and $t \geq 0$:*

$$\mathbb{P}\left\{|2\eta_{k,l}(X) - 1| < t\right\} \leq Bt^{\frac{\alpha}{1-\alpha}}.$$

W.p. $\geq 1 - \delta$,

$$L_{k,l}(\widehat{g}_{k,l}) - L^*_{k,l} \leq 2\left(\inf_{g \in \mathcal{G}} L_{k,l}(g) - L^*_{k,l}\right) + r_n(\delta),$$

with $r_n(\delta) = O(n^{-\frac{1}{2-\alpha}})$.

One-Versus-One for Label Ranking (2/2)

Introduce the scoring rule $\widehat{s}(X)(k) = 1 + \sum_{k \neq l} \mathbb{I}\{\widehat{g}_{k,l}(X) = -1\}$, with $\widehat{s}(X) = \widehat{\sigma}_X$, the union bound implies that:

$$\mathbb{P}\{\widehat{\sigma}_X \neq \sigma_X^*\} \leq \sum_{k<l} \mathbb{P}\left\{\widehat{g}_{k,l}(X) \neq g_{k,l}^*(X)\right\}. \tag{2}$$

Eq. (2) and Proposition 5 imply guarantees for LR.

Theorem 6

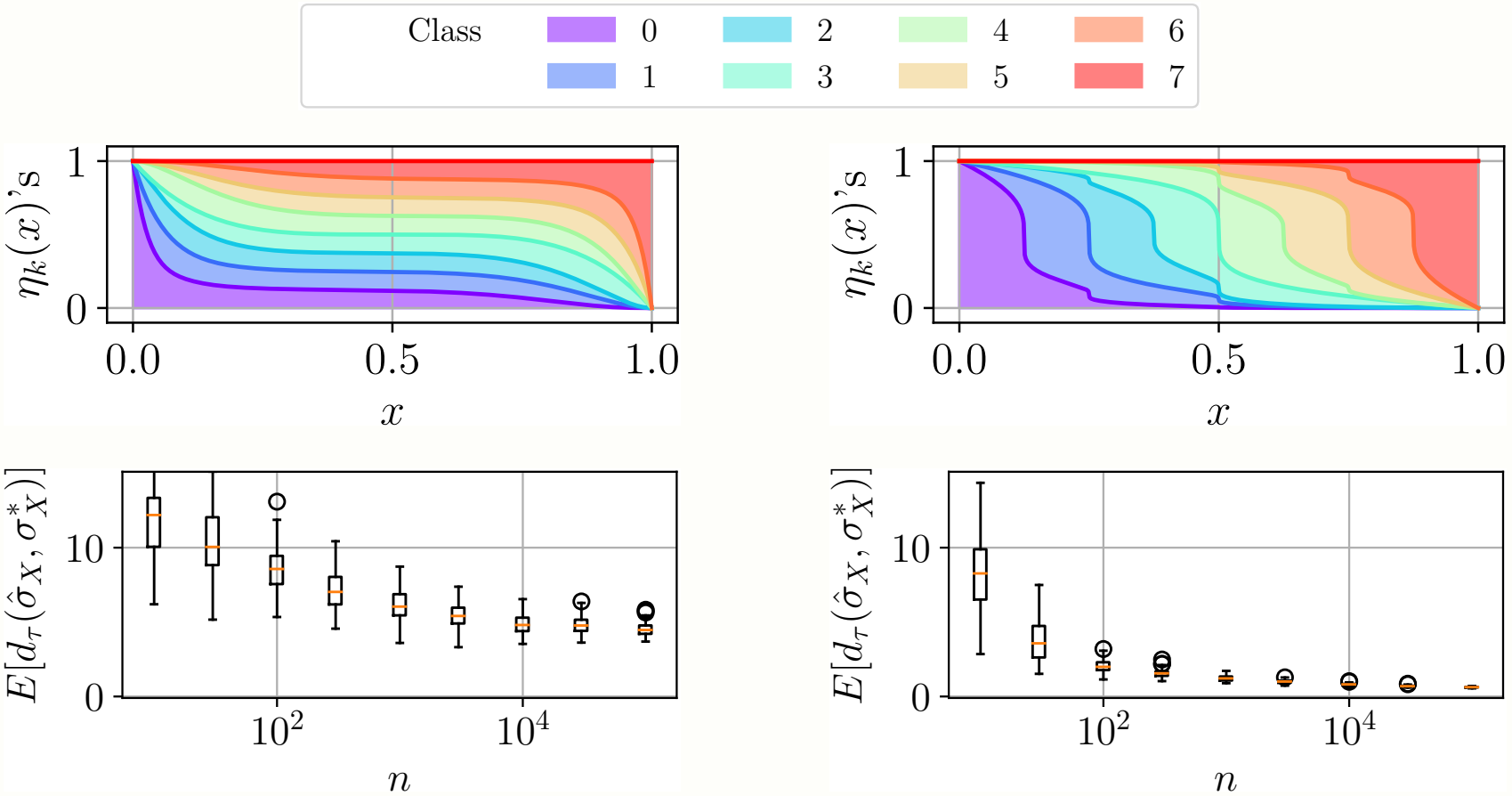
Under the same assumptions as those of Proposition 5, for all $\delta \in (0,1)$ and $n \geq n_0(\delta, \alpha, \epsilon, B, V)$, w.p. $\geq 1 - \delta$,

$$\mathbb{P}\{\hat{\sigma}_X \neq \sigma_X^*\} \leq \frac{\beta}{\epsilon}\left(\sum_{k<l} 2\left(\inf_{g \in \mathcal{G}} L_{k,l}(g) - L_{k,l}^*\right)^{\alpha} + \binom{K}{2} r_n^{\alpha}\left(\frac{\delta}{\binom{K}{2}}\right)\right).$$

Theorem 6 gives a generalization bound in $O(n^{-\frac{\alpha}{2-\alpha}})$,
e.g. $\alpha = 1/2$ gives $O(n^{-\frac{1}{3}})$ versus $O(n^{-\frac{2}{3}})$ for the usual $O(n^{-\frac{1}{2-\alpha}})$.

Toy example

Set $\mathcal{X} = [0, 1]$, in two settings: noisy and separated (see η's below), we plot the expected d_τ between $\widehat{\sigma}_X$ and σ^*_X over 100 simulations, as a function of n.



Noisy setting: $\alpha = 0.2$

Separated setting: $\alpha = 0.8$

Guarantees with OVO for top-k and classification

The top-k loss writes $\ell_k(y, \sigma) = \mathbb{I}\{y \notin \{\sigma^{-1}(1), \ldots, \sigma^{-1}(k)\}\}$, hence the top-$k$ risk of the ranking rule $s : \mathbb{R}^q \to \mathfrak{S}_K$ is:

$$W_k(s) = \mathbb{E}\left[\ell_k(Y, s(X))\right].$$

Introduce $W_k^* = \min_{s \in \mathcal{S}} W_k(s)$, then we have the following result:

Proposition 7

Let $k \in \{1, \ldots, K\}$, then $W_k^* = W_k(\sigma_X^*)$.
Under the same assumptions as those of Proposition 5,
for all $\delta \in (0, 1)$ and $n \geq n_0(\delta, \alpha, \epsilon, B, V)$:

$$W_k(\hat{\sigma}_X) - W_k^* \leq \frac{\beta}{\epsilon} C_{K,k} \left(2 \max_{m \neq l} \left(\inf_{g \in \mathcal{G}} L_{l,m}(g) - L_{l,m}^* \right)^{\alpha} + r_n^{\alpha} \left(\frac{\delta}{\binom{K}{2}} \right) \right).$$

The proof relies on: $W_k(s) - W_k^* \leq \mathbb{P}_X \{\mathrm{Top}_k(\hat{\sigma}_X) \neq \mathrm{Top}_k(\sigma_X^*)\}$.

It implies, w/ $k = 1$, guarantees for the OVO classif $\bar{g}(X) := \widehat{\sigma}_X{}^{-1}(1)$.

Outline

Conclusion

Main contributions:

· We derived a theory for solving label ranking with classif. data,

· and incidentally prove learning bounds for the OVO classifier.

Future work:

One could study the optimization of other accuracy measures, such as ones that include the notion of scoring function.

Thank you !

References I


Allwein, E., Schapire, R., and Singer, Y. (2000).
Reducing multiclass to binary: a unifying approach for margin classifiers.
Journal of Machine Learning Research, 1:113–141.


Brinker, K. and Hüllermeier, E. (2019).
A reduction of label ranking to multiclass classification.
In *ECML PKDD*.


Clémençon, S., Korba, A., and Sibony, E. (2018).
Ranking median regression: Learning to order through local consensus.
In *Proceedings of the conference Algorithmic Learning Theory*.


Copeland, A. H. (1951).
A reasonable social welfare function.
In *Seminar on applications of mathematics to social sciences, University of Michigan*.


Korba, A., Garcia, A., and d'Alché Buc, F. (2018).
A structured prediction approach for label ranking.
In *NeurIPS*.


Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012).
Imagenet classification with deep convolutional neural networks.
In *NIPS*.

References II

Tsoumakas, G., Katakis, I., and Vlahavas, I. (2009).
Mining multi-label data.
In *Data mining and knowledge discovery handbook*, pages 667–685. Springer.

Vembu, S. and Gärtner, T. (2010).
Label ranking algorithms: A survey.
In *Preference learning*, pages 45–64. Springer.

Wu, T., Lin, C., and Weng, R. (2004).
Probability estimates for multi-class classification by pairwise coupling.
Journal of Machine Learning Research, 5:975–1005.