ON MISSING LABELS, LONG-TAILS AND PROPENSITIES IN EXTREME MULTI-LABEL CLASSIFICATION

Erik Schultheis[1] Marek Wydmuch[2] Rohit Babbar[1] Krzysztof Dembczyński[2,3]
[1]Aalto University, Helsinki, Finland [2]Poznan University of Technology, Poland [3]Yahoo! Research, New York, USA

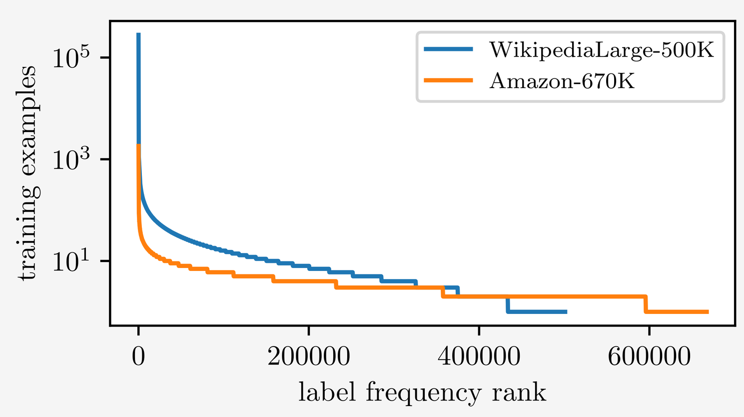
Extreme Multi-Label Classification (XMLC)

Problem setting

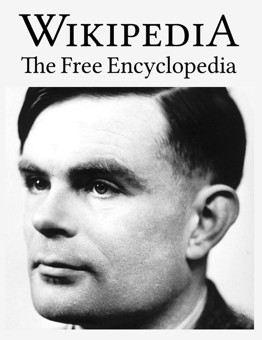
$$x \in \mathcal{X} \xrightarrow{h(x)} \boldsymbol{y} \in \mathcal{Y} := \{0,1\}^m$$

– Number of **labels** m is **large** ($\geq 10^5$).
– **Each** example has only **few** relevant labels, $\|\boldsymbol{y}\|_1 \ll m$.
– **Most** labels are relevant only to **few** instances $\Rightarrow$ **tail** labels.
– Obtaining labels is **challenging** $\Rightarrow$ **missing** labels.
– Applications: tagging, recommendation, ranking.

Long-tailed label distribution



Missing labels



Alan Turing
(over 8000 revisions)

Categories (55 assigned): Alan Turing, 1912 births, 1954 deaths, 1954 suicides, 20th-century mathematicians, 20th-century atheists, 20th-century British scientists, 20th-century English philosophers, Academics of the University of Manchester, [...], Computer designers, English atheists, English computer scientists, English inventors, English logicians, English male long-distance runners, English mathematicians, English people of Irish descent, English people of Scottish descent, [...], Recipients of British royal pardons, Suicides by cyanide poisoning, Suicides in England, Theoretical computer scientists, Deaths by poisoning
Missing (e.g.): 20th-century English scientists, Enigma machine, Suicides by poisoning

Missing labels and propensity models

Missing labels

– $\boldsymbol{Y}$ — ground-truth labels,
– $\tilde{\boldsymbol{Y}}$ — observed labels.

e.g.:

	y_1	y_2	$\dots$	y_m
$\boldsymbol{y}=$	1	1	$\dots$	0
$\tilde{\boldsymbol{y}}=$	1	0	$\dots$	0

– In general, we have:

$$\mathbb{P}\big[\tilde{\boldsymbol{Y}} \preceq \boldsymbol{Y} \,|\, X\big] = 1, \qquad \mathbb{P}\big[\tilde{\boldsymbol{Y}} \npreceq \boldsymbol{Y} \,|\, X\big] = 0,$$

where:
– $\tilde{\boldsymbol{Y}} \preceq \boldsymbol{Y}$ means $\tilde{Y}_j \leq Y_j$ for all $j \in [m]$,
– $\tilde{\boldsymbol{Y}} \npreceq \boldsymbol{Y}$ means that there is at least one label for which $\tilde{Y}_j > Y_j$.

General propensity model

– Propensities defined over entire label vectors:

$$p_{\tilde{\boldsymbol{y}}}(\boldsymbol{y}, x) := \mathbb{P}\big[\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{y}} \,|\, \boldsymbol{Y} = \boldsymbol{y}, X = x\big]$$

– Reconstruction of the ground truth distribution from the observed one requires an **exponential number of parameters**.

Label-wise propensity model

– Assumes the propensities to be defined label-wise:

$$p_j(X) := \mathbb{P}\big[\tilde{Y}_j = 1 \,|\, Y_j = 1, X\big] \,.$$

– Define observed and ground-truth conditional probabilities:

$$\tilde{\eta}_j(x) := \mathbb{P}\big[\tilde{Y}_j = 1 \,|\, X = x\big] \,, \qquad \eta_j(x) := \mathbb{P}[Y_j = 1 \,|\, X = x] \,.$$

– The relation between them is given by:

$$\tilde{\eta}_j(x) = p_j(x)\eta_j(x) \,, \qquad \eta_j(x) = \tilde{\eta}_j(x)/p_j(x) \,.$$

Unbiased losses

– **Task risk** of classifier $h : \mathcal{X} \longrightarrow \mathbb{R}^m$ ($x \mapsto h(x) =: \hat{\boldsymbol{y}}$):

$$\mathrm{Risk}_{\ell_{\mathrm{task}}}[h; X, \boldsymbol{Y}] := \mathbb{E}[\ell_{\mathrm{task}}(\boldsymbol{Y}, h(X))] \,,$$

where $\ell_{\mathrm{task}} : \mathcal{Y} \times \mathbb{R}^m \longrightarrow \mathbb{R}_{\geq 0}$ is the (**task**) **loss**.
– If propensities are known, then they can be used to construct an **unbiased loss** $\tilde{\ell}$ s.t. $\forall h :\ \mathrm{Risk}_{\ell}[h; X, \boldsymbol{Y}] = \mathrm{Risk}_{\tilde{\ell}}[h; X, \tilde{\boldsymbol{Y}}]$.

Recipes to follow

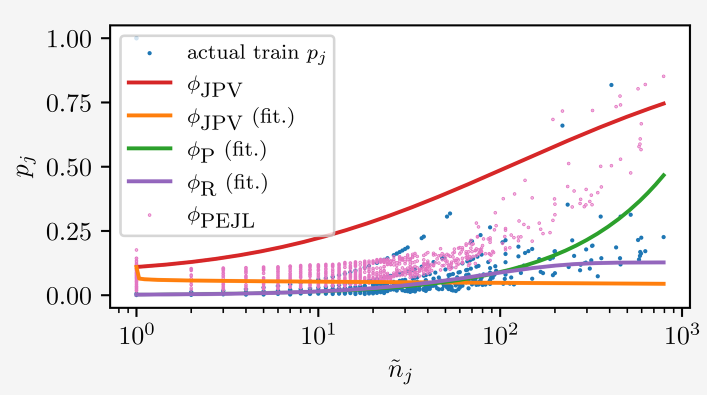
Bias-controlled test sets, alternative propensity models, and different estimation approaches

Estimates based on a bias-controlled test set for Yahoo R3 data with different propensity models fitted to the data:

$$\phi_{\mathrm{P}}(\tilde{\pi}_j; \beta, \gamma) := (\beta\tilde{\pi}_j)^{\gamma}$$

$$\phi_{\mathrm{R}}(\tilde{\pi}_j; c, \dots, h) := c + \frac{d-c}{(e + f\exp(-g\tilde{\pi}_j))^{1/h}}$$

ϕ_{PEJL} - Propensity Estimation via Joint Learning [1, 2] estimates p_js jointly with training a classifier on a biased train set.



Bias-controlled test set allows to evaluate propensity models:

$p_j =$	ϕ_{JPV}	ϕ_{JPV} (fit.)	ϕ_{P} (fit.)	ϕ_{R} (fit.)	ϕ_{PEJL}	p_j
P@1 (%)	66.03	48.58	63.53	71.23	68.09	73.72

Alternative task losses for long-tails

$$\mathrm{F}^{\mathrm{macro}}_{\beta}(\{\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i\}_1^n) = \frac{1}{m}\sum_j \frac{(1+\beta^2)\sum_i y_{ij}\hat{y}_{ij}}{\beta^2\sum_i y_{ij} + \sum_i \hat{y}_{ij}}$$

$$\mathrm{Abandonment@}k(\boldsymbol{y}, \hat{\boldsymbol{y}})^{[3]} = \mathbb{1}[\forall j \in \mathrm{top}_k(\hat{\boldsymbol{y}}) : y_j \neq 1]$$

$$\mathrm{Coverage}(\{\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i\}_1^n) = m^{-1}|\{j \in [m] : \exists i \in [n] \text{ s.t. } y_{ij} = \hat{y}_{ij} = 1\}|$$

Current state-of-the-art and its shortcomings

– A seminal paper by Jain et al. [3] has introduced propensities into XMLC to deal with missing and long-tail labels.
– Results from this paper have been followed in many other papers.

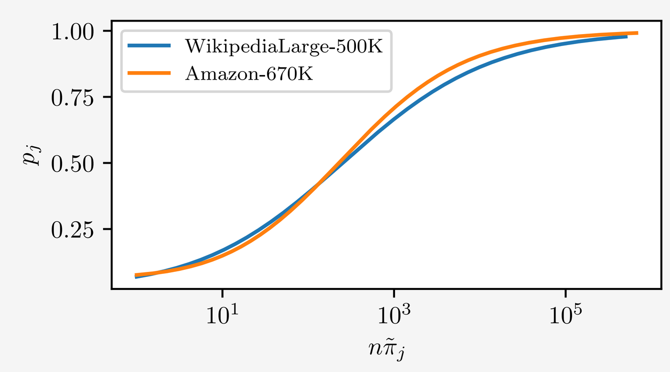
Propensity-scored losses (measures)

Measure definition		Unbiased estimate	
$\mathrm{P@}k(\boldsymbol{y}, \hat{\boldsymbol{y}})$	$= k^{-1}\sum_{j \in \mathrm{top}_k(\hat{\boldsymbol{y}})} y_j$	$\mathrm{PSP@}k(\boldsymbol{y}, \hat{\boldsymbol{y}})$	$= k^{-1}\sum_{j \in \mathrm{top}_k(\hat{\boldsymbol{y}})} \tilde{y}_j/p_j$
$\mathrm{nDCG@}k(\boldsymbol{y}, \hat{\boldsymbol{y}})$	$= \frac{\sum_{j \in \mathrm{top}_k(\hat{\boldsymbol{y}})} \frac{y_j}{\log(\mathrm{r}_j(\hat{\boldsymbol{y}})+1)}}{\sum_{j=1}^k \frac{1}{\log(j+1)}}$	$\mathrm{PSnDCG@}k(\boldsymbol{y}, \hat{\boldsymbol{y}})$	$= \frac{\sum_{j \in \mathrm{top}_k(\hat{\boldsymbol{y}})} \frac{\tilde{y}_j}{p_j \log(\mathrm{r}_j(\hat{\boldsymbol{y}})+1)}}{\sum_{j=1}^k \frac{1}{\log(j+1)}}$

(top_k maps a vector to the indices of its top-k components; $\mathrm{r}_j(\hat{\boldsymbol{y}})$ gives the rank of the j-th element in the vector.)

The JPV propensity model

$$p_j = \phi_{\mathrm{JPV}}(\tilde{\pi}_j; n, a, b) := \frac{1}{1 + (\log n - 1)(b+1)^a e^{-a\log(n\tilde{\pi}_j + b)}} \,,$$



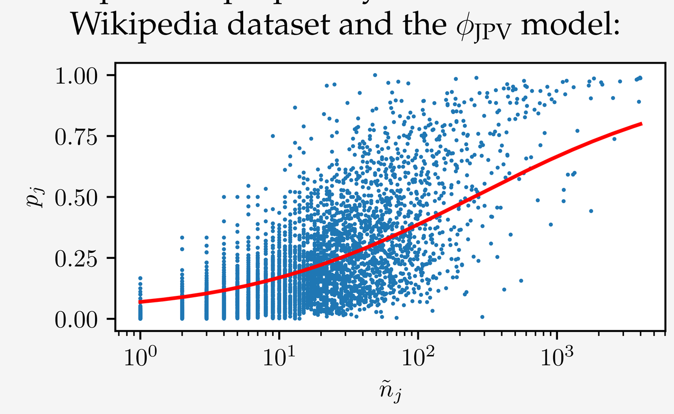
where $\tilde{\pi}_j := \mathbb{P}\big[\tilde{Y}_j = 1\big]$, n is the number of training instances, and a and b are **dataset-dependent** parameters. It is assumed that p_js are **constant** values without dependence on x.

Shortcomings and pitfalls

– **Unclear missing-labels assumptions**
Jain et al. [3] prove $\mathbb{E}[\ell(\boldsymbol{Y}, \hat{\boldsymbol{y}})] = \mathbb{E}\Big[\tilde{\ell}(\tilde{\boldsymbol{Y}}, \hat{\boldsymbol{y}})\Big]$ for any **fixed** prediction $\hat{\boldsymbol{y}}$ without a clear **dependence** on X, which further implies the assumptions behind propensities to be unclear.

– **Estimation of parameters, reproducibility, and propensities as a function of frequency**
In order to determine values for a and b, Jain et al. 2016 [3] investigated two datasets (Wikipedia and Amazon) in which auxiliary information could be used to identify some missing labels.

Replicated propensity estimates for the Wikipedia dataset and the ϕ_{JPV} model:



– **Sensitivity to the number of instances**
Propensities obtained by the JPV model converge to 1 in the limit:

$$\lim_{n\to\infty} \phi_{\mathrm{JPV}}(\tilde{\pi}_j, n) = \frac{1}{1 + (b+1)^a \lim_{n\to\infty}(\log n - 1)e^{-a\log(\tilde{\pi}_j n + b)}} = 1.$$

– **Implausible results and hidden normalization**
PSP@k usually reported as:

$$\mathrm{Norm\ PSP@}k = \frac{\sum_{i=1}^n \mathrm{PSP@}k(\tilde{\boldsymbol{y}}_i, \hat{\boldsymbol{y}}_i)}{\sum_{i=1}^n \max_{\mathbf{z}} \mathrm{PSP@}k(\tilde{\boldsymbol{y}}_i, \mathbf{z})} \,.$$

Effect on the results of PfastreXML [3] on Amazon-670K:

PSP(%)	@1	@3	@5
Unnormalized	326.47	282.28	250.57
Normalized	29.93	31.26	32.80

– **Mismatched usage for missing and tail labels**
Missing labels are an orthogonal problem to tail labels.

References

[1] Paweł Teisseyre, Jan Mielniczuk, and Małgorzata Łazęcka. Different strategies of fitting logistic regression for positive and unlabelled data. In *ICCS*, 2020
[2] Ziwei Zhu, Yun He, Yin Zhang, and James Caverlee. Unbiased implicit recommendation and propensity estimation via combinational joint learning. In *RecSys*, 2020
[3] Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. Learning diverse rankings with multi-armed bandits. In *ICML*, 2008
[4] Himanshu Jain, Yashoteja Prabhu, and Manik Varma. Extreme multi-label loss functions for recommendation, tagging, ranking and other missing label applications. KDD, 2016

The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), 2022, Washington, DC, USA.