Extreme Multi-label Loss Functions for Recommendation, Tagging, Ranking & Other Missing Label Applications

Himanshu Jain
IIT Delhi

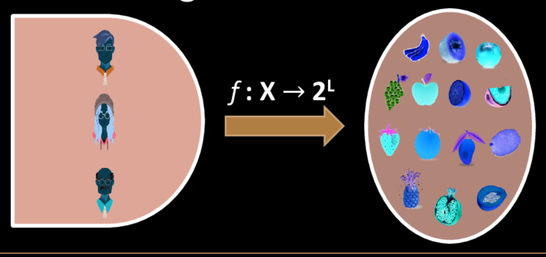
Yashoteja Prabhu
IIT Delhi

Manik Varma
Microsoft Research

Loss Functions for Extreme M-L Learning

Extreme Multi-Label Learning:

- Learning with millions of labels



Loss functions:

- Loss functions influence
  - Training
  - Hyper-parameter tuning
  - Model selection
  - Performance evaluation

Traditional Loss Functions

Examples:
Hamming loss, Precision, Recall, F-score, Coverage

Limitations:

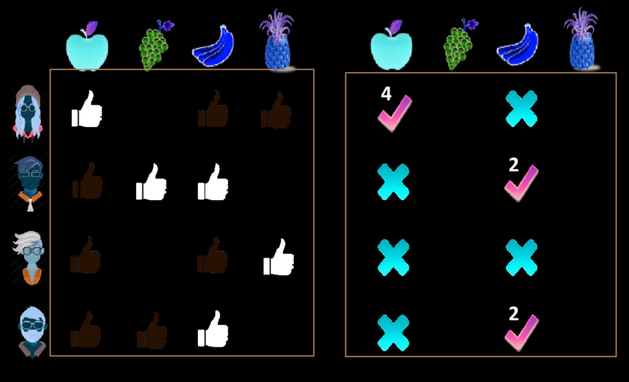
- No priority to predicting few relevant labels over million irrelevant labels e.g. Hamming loss
- Equal priority to all relevant labels
- Biased due to missing labels in ground truth data
- Bad training and test performance

Our Contributions

- We propose novel propensity-weighted loss functions which are unbiased to missing labels in the ground truth
- We propose a new sigmoidal model for label propensities based on empirical evidence
- We develop a novel extreme learning algorithm which achieves best results on these new loss functions
- We train on 9M labels, 70M points, 2M features and achieve significant improvements over state-of-the-art

Propensity Weighted Loss Functions

Intuition:



PSPrecision@2

Formulation:

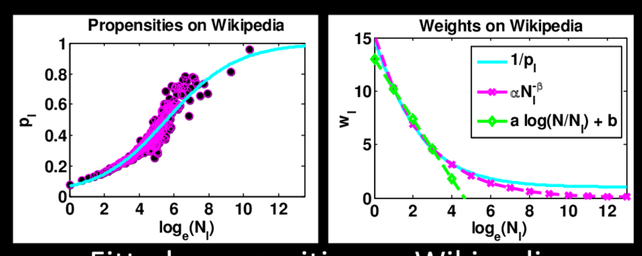
Gain	$-\mathcal{L}(y_l, \hat{y}_l)$
PSPrecision@k	$\frac{1}{k}\sum_l \frac{1}{p_l} y_l \hat{y}_l$
PSnDCG@k	$\frac{\sum_l \frac{y_l \hat{y}_l}{p_l \log(r_l+1)}}{\sum_{l=1}^{k} \frac{1}{\log(1+l)}}$

Propensity Model

$$p_l = \frac{1}{1 + C\, e^{-A \log(N_l + B)}}$$

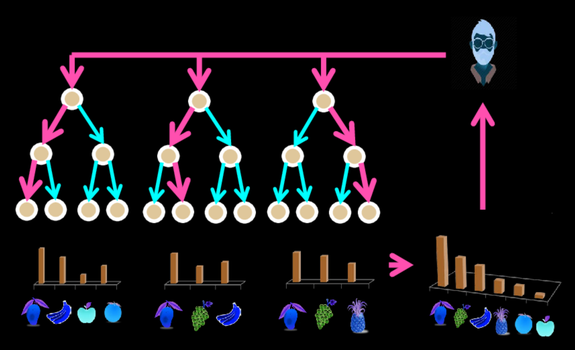
Where,
$C = (\log(N) - 1)(B + 1)^A$

Model



Fitted propensities on Wikipedia

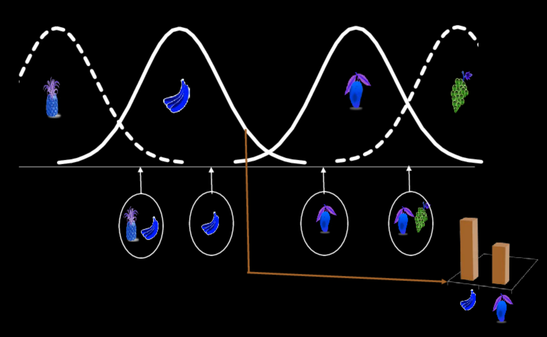
PfastreXML



$$\text{Min}_{\mathbf{w},\delta,\mathbf{r}^{\pm}} \ \|\mathbf{w}\|_1 + \sum_i C_\delta(\delta_i) \log\left(1 + e^{-\delta_i \mathbf{w}^t \mathbf{x}_i}\right) - C_r \frac{(1+\delta_i)}{2} \text{PSnDCG@L}(\mathbf{y}_i, \mathbf{r}^+) - C_r \frac{(1-\delta_i)}{2} \text{PSnDCG@L}(\mathbf{y}_i, \mathbf{r}^-)$$

PfastXML

+



$$P_l(\mathbf{x}_i) = \frac{1}{1 + e^{\frac{\gamma}{2}||\mathbf{x}_i - \boldsymbol{\mu}_l||^2}}$$

Tail Label Classifiers

Results

Dataset Statistics:

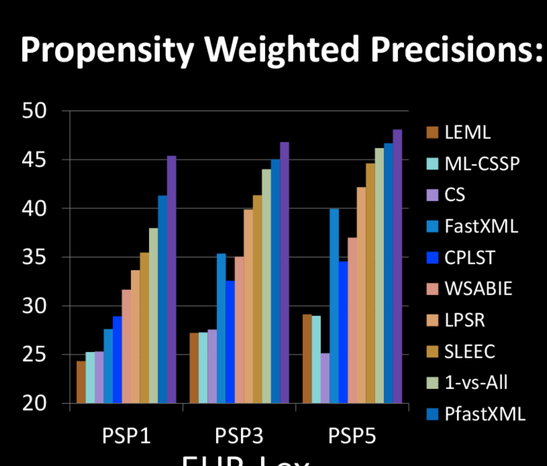
Data Set	# of Training Points	# of Test Points	# of Dimensions	# of Labels
EUR-Lex	15,539	3,809	5000	3993
WikiLSHTC	1,778,351	587,084	1,617,899	325,056
Ads-9M	70,455,530	22,629,136	2,082,698	8,838,461

Qualitative Results:

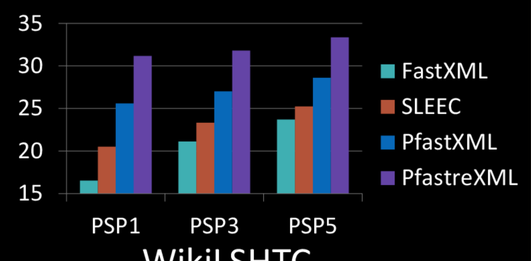
FastXML	PfastreXML
Medieval literature (189)	Works by Dante Alighieri (7)
Philosophical novels (92)	Divine Comedy (20)
1855 births (778)	1321 books (3)
1868 births (976)	1300 in Italy (4)
1977 books (60)	Visionary poems (7)
2003 novels (249)	Epic poems in Italian (7)
2006 books (209)	14th-century Christian texts (10)
21st-century American novels (591)	14th-century books (47)
American poetry collections (77)	Virgil (13)
Electronic music festivals (92)	Dante Alighieri (26)

Predictions for “Divine Comedy”

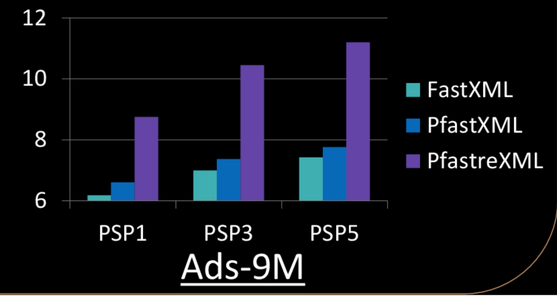
Propensity Weighted Precisions:



EUR-Lex



WikiLSHTC



Ads-9M