Multilabel reductions: what is my loss optimising?

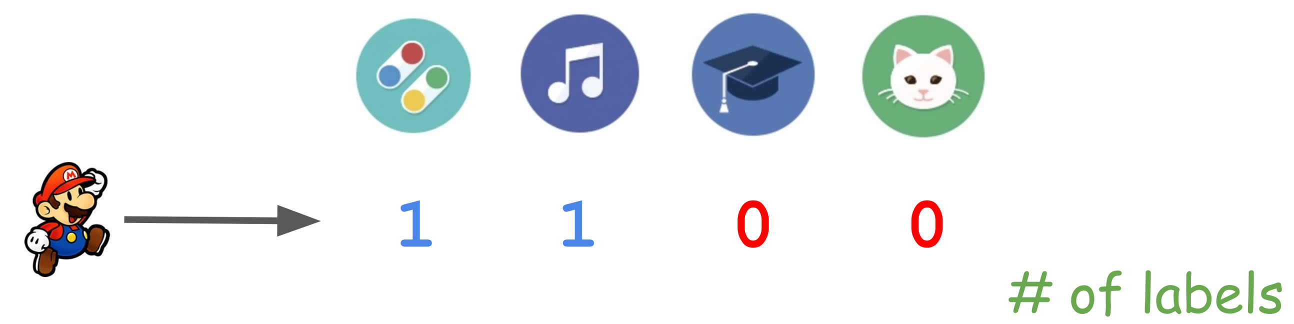
Aditya Krishna Menon, Ankit Singh Rawat, Sashank J. Reddi, Sanjiv Kumar

Google Research NYC

Five common multilabel reductions implicitly optimise for precision <u>or</u> recall

Multilabel classification via reductions

Multilabel classification: predict a binary label vector

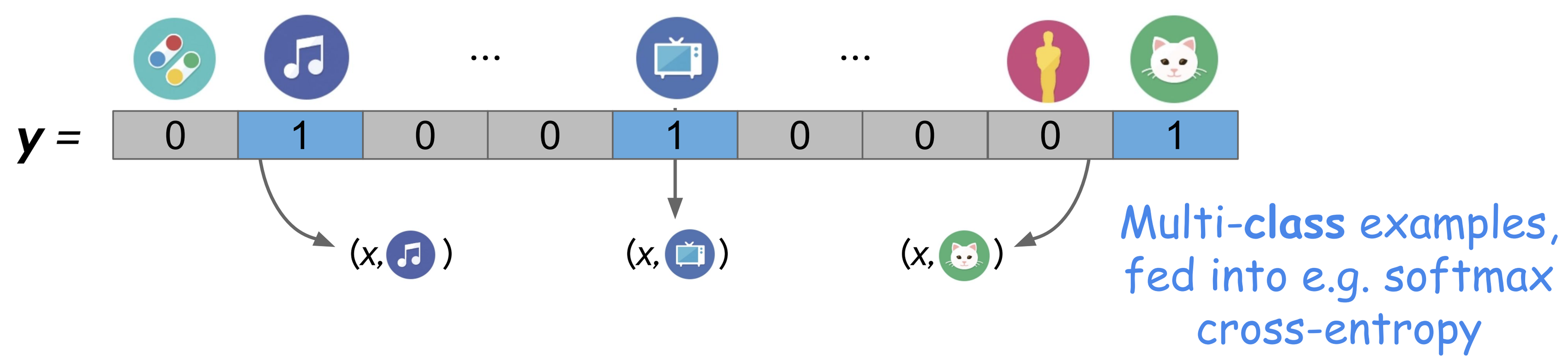


Given $(x, \mathbf{y})$ for $\mathbf{y} \in \{0, 1\}^L$, find $\mathbf{f}(x) \in \mathbb{R}^L$ *# of labels*

Ideally, want $f_i(x)$ high if $y_i = 1$. The challenge is that L may be large; so how do we efficiently find such an $\mathbf{f}$?

Common algorithms reduce the problem to binary or multiclass learning; e.g., we may create

- multiple binary examples, one for each label
- multiple multi-class examples, one for each **positive** label
- one multi-class example for a **random positive** label



Key question of this work

Such reductions can be practically effective, but:

what multilabel metric do they optimise?

Answering this helps inform the choice of reduction, depending on our end goal.

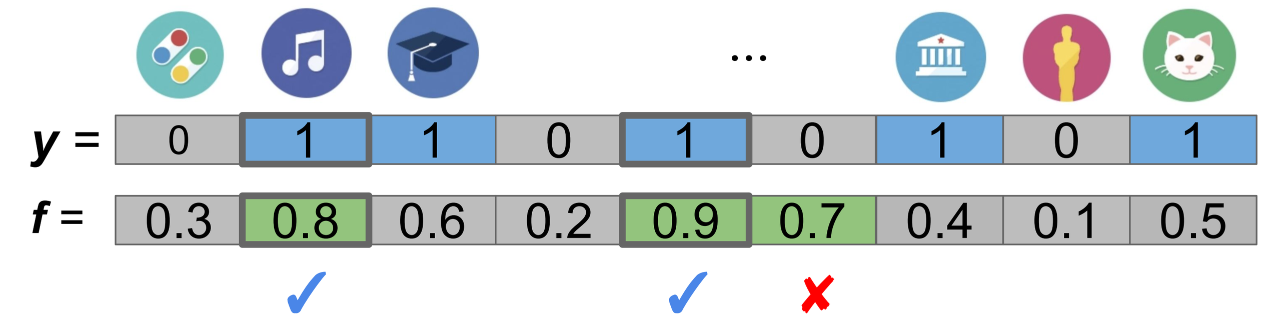
Key contribution of this work

five common reductions *implicitly* optimise for either precision <u>or</u> recall@*k*!

To begin, we study a basic property of these metrics.

Multilabel metrics: precision and recall

Precision and **recall@*k*** are standard metrics for retrieval settings. They measure the # of positives in the top-*k* scoring indices, suitably normalised:



$\mathrm{Prec@}k(\mathbf{f}) = \mathbb{E}_{(x,\mathbf{y})}[\,|\,\mathrm{top}_k(\mathbf{f}(x)) \cap \mathrm{pos}(\mathbf{y})\,| \,/\, k\,]$ *Different normaliser*

$\mathrm{Rec@}k(\mathbf{f}) = \mathbb{E}_{(x,\mathbf{y})}[\,|\,\mathrm{top}_k(\mathbf{f}(x)) \cap \mathrm{pos}(\mathbf{y})\,| \,/\, |\mathrm{pos}(\mathbf{y})|\,]$

Fact: Given a distribution $\mathrm{P}(x, \mathbf{y})$, the optimal $\mathbf{f}^*$ for these metrics preserve the ordering of:

(Prec@k) $\mathrm{P}(y_i = 1 \mid x)$ *Marginal relevance*

(Rec@k) $\mathrm{P}(y'_i = 1 \mid x) = \mathrm{P}(y_i = 1 \mid x) \cdot \mathbb{E}[\,(1 + \Sigma_{j \neq i}\, y_j)^{-1}\,]$ *Non-constant weighting*

Note that owing to the weighting above, these optimal solutions are **incompatible** in general!

Implicit losses for multilabel reductions

To compare different reductions, we explicate their implicit multilabel losses, i.e., $\ell(\mathbf{y}, \mathbf{f}(x))$:

One-versus-all (OVA) $\displaystyle\sum_{i \in [L]} \ell_{\mathrm{BC}}(y_i, f_i)$ *Binary loss*

Pick-all-labels (PAL) $\displaystyle\sum_{i \in [L]} y_i \cdot \ell_{\mathrm{MC}}(i, f)$ *Multiclass loss; has label "competition"*

OVA normalised $\displaystyle\sum_{i \in [L]} \left\{ \frac{y_i}{\sum_{j \in [L]} y_j} \cdot \ell_{\mathrm{BC}}(1, f_i) + \left(1 - \frac{y_i}{\sum_{j \in [L]} y_j}\right) \cdot \ell_{\mathrm{BC}}(0, f_i) \right\}$

PAL normalised / Pick-one-label (POL) $\displaystyle\sum_{i \in [L]} \frac{y_i}{\sum_{j \in [L]} y_j} \cdot \ell_{\mathrm{MC}}(i, f)$

For log-loss, KL(label, prediction)

Optimal scores for multilabel reductions

Equipped with these losses, observe that, e.g.,

$\mathrm{Prec@}k(\mathbf{f}) = \mathbb{E}_{(x,\mathbf{y})}[\,\Sigma_{i \in [L]}\, \ell_{\text{top-k}}(\mathbf{y}, f_i(x)) \,/\, k\,]$,

where $\ell_{\text{top-k}}$ is a "top-*k*" multiclass loss; i.e., **precision is equivalent to PAL** for a specific loss! More generally:

Fact: Given a distribution $\mathrm{P}(x, \mathbf{y})$, the optimal f_i^* is:

(OVA) $\mathrm{P}(y_i = 1 \mid x)$ *Classical*

(PAL) $\mathrm{P}(y_i = 1 \mid x) \,/\, N(x)$ *Expected # of +'ves Doesn't vary with i*

(Rest) $\mathrm{P}(y'_i = 1 \mid x)$ *c.f. recall@k optimal scorer*

Different reductions target either precision or recall! Subtle differences in the loss thus have non-trivial effects; e.g., $\ell_{\text{PAL-Norm}}(\mathbf{y}, \mathbf{f}(x)) = (\Sigma_i\, y_i)^{-1}\, \ell_{\mathrm{PAL}}(\mathbf{y}, \mathbf{f}(x))$, but the optimal solutions for the two are <u>not</u> scalings of each other. Further remarks:

- PAL scores are not calibrated across instances!
- PAL may ≥ OVA since it directly bounds top-*k* loss
- PAL with a multiclass OVA loss ≠ multilabel OVA; PAL implicitly places a greater weight on each "negative" label

Illustration of differences in reductions

Synthetic problem where normalised reduction gives recall gains, at the expense of precision.