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 or recall

Multilabel classification via reductions

Multilabel classification: predict a binary label vector

\[ y = [1 1 0 0] \]

Given \((x, y) \) for \( y \in \{0, 1\}^k \), find \( f(x) \in \mathbb{R}^k \)

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

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 or 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:

\[ \text{Prec@k}(f) = \frac{\mathbb{E}_{(x, y)}[|\text{top}_k(f(x)) \cap \text{pos}(y)|]}{k} \]
\[ \text{Rec@k}(f) = \frac{\mathbb{E}_{(x, y)}[|\text{top}_k(f(x)) \cap \text{pos}(y)|]}{|\text{pos}(y)|} \]

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

\[ (\text{Prec@k}) P(y_i = 1 | x) - \] Marginal relevance

\[ (\text{Rec@k}) P(y_i' = 1 | x) = P(y_i = 1 | x) \cdot \mathbb{E}[ (1 + \Sigma_{j \neq i} y_j)^{-1}] \]

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., \( \ell(y, f(x)) \):

- One-versus-all (OVA)
  \[ \sum_{i \in [L]} \ell_{\text{MC}}(y_i, f_i) \]
- Pick-all-labels (PAL)
  \[ \sum_{i \in [L]} y_i \cdot \ell_{\text{MC}}(i, f) \]
- OVA normalised
  \[ \sum_{i \in [L]} \left\{ \sum_{j \in [L]} y_j f_{\text{OVA}}(0, f_j) + (1 - \sum_{j \in [L]} y_j) f_{\text{OVA}}(0, f_j) \right\} \]
- PAL normalised
  \[ \sum_{i \in [L]} \sum_{j \in [L]} y_j f_{\text{MC}}(i, f) \]

For log-loss, KL(label, prediction)

Optimal scores for multilabel reductions

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

\[ \text{Prec@k}(f) = \mathbb{E}_{(x, y)}[\sum_{i \in [L]} \ell_{\text{top}_k}(y_i, f_i) / k] \]

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 \( P(x, y) \), the optimal \( f^* \) is:

- (OVA) \( P(y_i = 1 | x) \)
- (PAL) \( P(y_i = 1 | x) / N(x) \)
- (Rest) \( P(y_i = 1 | x) \)

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., \( \ell_{\text{PAL-Norm}}(y, f(x)) = (\mathbb{E}[y])^{-1} \ell_{\text{PAL}}(y, f(x)) \), but the optimal solutions for the two are not 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.