Condensed Filter Tree for Cost-Sensitive Multi-Label Classification

A. Proof of Theorem 1

Theorem 1. *Under the proper ordering and K-classifier tricks, for each* $\mathbf{x}$ *and the multi-label classifier* h *formed by chaining* K *binary classifiers* $(h_1, ..., h_K)$ *as in the prediction procedure of Filter Tree, the regret* $rg(h, \mathcal{P})$ *is*

$$rg(h,\mathcal{P}) \leq \sum_{\mathbf{t}\in\langle r,\mathbf{y}^*\rangle} [\![h_k(\mathbf{x},\mathbf{t}) \neq \mathbf{y}[k]]\!] rg\Big(h_k(\mathbf{x},\mathbf{t}), \mathsf{FT}_{\mathbf{t}}(\mathcal{P}, h_{k+1}, ..., h_K)\Big),$$

where k *denotes the layer that* $\mathbf{t}$ *is on, and* $\mathsf{FT}_{\mathbf{t}}(\mathcal{P}, h_{k+1}, ..., h_K)$ *represents the procedure that generates weighted examples* $(\mathbf{x}, b, w)$ *to train the node at index* $\mathbf{t}$ *based on sampling* $\mathbf{y}$ *from* $\mathcal{P}_{|\mathbf{x}}$ *and considering the predictions of classifiers in the lower layers.*

Proof. The proof is similar to the one in (Beygelzimer et al., 2008), which is based on defining the overall-regret of any subtree. The key change in our proof is to define the *path-regret* of any subtree to be the total regret of the nodes on the ideal path of the subtree. The induction step follows similarly from the proof in (Beygelzimer et al., 2008) by considering two cases: one for the ideal prediction to be in the left subtree and one for the ideal prediction to be in the right. Then an induction from layer K to the root proves the theorem.

For each node $\mathbf{t}$ on layer k, h_k makes a weighted binary classification decision of 0 or 1, which directs the prediction procedure to move to either the node $\mathbf{t}_0$ or $\mathbf{t}_1$. Without loss of generality, assume $h_k(\mathbf{x}, \mathbf{t}) = 1$. We denote $\hat{\mathbf{t}}$ as the prediction (leaf) on $\mathbf{x}$ when starting at node $\mathbf{t}$. For each leaf node $\tilde{\mathbf{y}}$, let $\bar{C}(\tilde{\mathbf{y}}) \equiv E_{\mathbf{y}\sim\mathcal{P}|\mathbf{x}} C(\mathbf{y}, \tilde{\mathbf{y}})$. Then, the node regret $rg(\mathbf{t})$ is simply $\bar{C}(\hat{\mathbf{t}}_1) - \min_{i\in\{0,1\}} \bar{C}(\hat{\mathbf{t}}_i)$. Obviously, $rg(\mathbf{t}) \geq \bar{C}(\hat{\mathbf{t}}_1) - \bar{C}(\hat{\mathbf{t}}_0)$ for all node $\mathbf{t}$.

In addition to the regret of nodes, we also define the regret of the subtree $T_{\mathbf{t}}$ rooted at node $\mathbf{t}$. The regret of the subtree $T_{\mathbf{t}}$ is as defined as the regret of the predicted path (vector) $\hat{\mathbf{t}}$ within the subtree $T_{\mathbf{t}}$, that is, $rg(T_{\mathbf{t}}) = \bar{C}(\hat{\mathbf{t}}) - \bar{C}(\mathbf{t}^*)$, where $\mathbf{t}^*$ denotes the optimal prediction (leaf node) in the subtree $T_{\mathbf{t}}$. By this definition, $rg(h, \mathcal{P})$ can be treated as $rg(T_r)$.

We now prove by induction from layer K to the root. The induction hypothesis is that

$$rg(T_{\mathbf{t}}) \leq \sum_{\mathbf{t}'\in\langle\mathbf{t},\mathbf{t}^*\rangle} [\![h_k(\mathbf{x}, \mathbf{t}') \neq \mathbf{y}[k]]\!] rg(\mathbf{t}'),$$

where k is the corresponding layer of each node $\mathbf{t}'$. The hypothesis states that the regret of the subtree is bounded by the sum of the regrets for the wrongly predicted nodes from $\mathbf{t}$ to the ideal prediction $\mathbf{t}^*$. The base case is the reduction tree with one single internal node $\mathbf{t}$ and two leaf nodes, which is a cost-sensitive binary classification with $rg(T_{\mathbf{t}}) = rg(\mathbf{t})$ trivially. If h_1 predicts correctly, then $rg(T_{\mathbf{t}}) = 0$. Otherwise $rg(T_{\mathbf{t}}) = rg(\mathbf{t})$. Then the induction hypothesis is satisfied.

For the inductive step, for node $\mathbf{t}$ on layer k, assume

$$R_0 \equiv rg(T_{\mathbf{t}_0}) \leq \sum_{\mathbf{t}'\in\langle\mathbf{t}_0,\mathbf{t}_0^*\rangle} [\![h_k(\mathbf{x}, \mathbf{t}') \neq \mathbf{y}[k]]\!] rg(\mathbf{t}'),$$

and

$$R_1 \equiv rg(T_{\mathbf{t}_1}) \leq \sum_{\mathbf{t}'\in\langle\mathbf{t}_1,\mathbf{t}_1^*\rangle} [\![h_k(\mathbf{x}, \mathbf{t}') \neq \mathbf{y}[k]]\!] rg(\mathbf{t}').$$

The optimal prediction $\mathbf{t}^*$ is either on the right subtree T_1 or the left subtree T_0. For the first case, it implies $\mathbf{t}^* = \mathbf{t}_1^*$ and $\mathbf{y}[k] = h_k(\mathbf{x}, \mathbf{t}) = 1$, then

$$\begin{aligned} rg(T_{\mathbf{t}}) &= \bar{C}(\hat{\mathbf{t}}_1) - \bar{C}(\mathbf{t}^*) \\ &= \bar{C}(\hat{\mathbf{t}}_1) - \bar{C}(\mathbf{t}_1^*) \\ &= R_1 \leq \sum_{t'\in\langle\mathbf{t}_1,\mathbf{t}_1^*\rangle} [\![h_k(\mathbf{x}, \mathbf{t}') \neq \mathbf{y}[k]]\!] rg(\mathbf{t}') \\ &= \sum_{\mathbf{t}'\in\langle\mathbf{t},\mathbf{t}^*\rangle} [\![h_k(\mathbf{x}, \mathbf{t}') \neq \mathbf{y}[k]]\!] rg(\mathbf{t}'). \end{aligned}$$

For the second case, it implies $\mathbf{t}^* = \mathbf{t}_0^*$ and $\mathbf{y}[k] \neq h_k(\mathbf{x}, \mathbf{t}) = 1$, then

$$\begin{aligned} rg(T_{\mathbf{t}}) &= \bar{C}(\hat{\mathbf{t}}_1) - \bar{C}(\mathbf{t}^*) \\ &= \bar{C}(\hat{\mathbf{t}}_1) - \bar{C}(\mathbf{t}_0^*) \\ &= \bar{C}(\hat{\mathbf{t}}_1) - \bar{C}(\hat{\mathbf{t}}_0) + \bar{C}(\hat{\mathbf{t}}_0) - \bar{C}(\mathbf{t}_0^*) \\ &\leq rg(\mathbf{t}) + R_0 \\ &\leq rg(\mathbf{t}) + \sum_{\mathbf{t}'\in\langle\mathbf{t}_0,\mathbf{t}_0^*\rangle} [\![h_k(\mathbf{x}, \mathbf{t}') \neq \mathbf{y}[k]]\!] rg(\mathbf{t}') \\ &= \sum_{\mathbf{t}'\in\langle\mathbf{t},\mathbf{t}^*\rangle} [\![h_k(\mathbf{x}, \mathbf{t}') \neq \mathbf{y}[k]]\!] rg(\mathbf{t}'). \end{aligned}$$

Then we complete the induction. □

B. Datasets

Here we summarize the basic statistics of the used datasets in Table 1.

Table 1. The properties of each dataset

Dataset	# Instances	# Labels (K)
CAL500	502	174
emotions	593	6
enron	1702	53
imdb	86290	28
medical	662	45
scene	2407	6
slash	3279	22
tmc	28596	22
yeast	2389	144

References

Beygelzimer, A., Langford, J., and Ravikumar, P. Error correcting tournaments, 2008.