REDUCING MULTICLASS COST SENSITIVE CLASSIFICATION TO BINARY COST SENSITIVE CLASSIFICATION

1. Goals

The goal here is to find a good reduction from multiclass cost sensitive classification to binary cost sensitive classification. These two problems are summarized next, then a method is discussed.

1.1. Multiclass Cost Sensitive Classification.

In multiclass cost sensitive classification, labeled examples have the form $(x, c_1, c_2, ..., c_k)$ where $x \in X$ is an arbitrary feature and $c_l \in [0, \infty)$ is the "cost" of choosing label l. Since the value of a good classifier is independent of a global offset, we'll assume that there is at least one label l which satisfies $c_l = 0$.

To solve this problem, we must find a hypothesis $h : X \to \{1, ..., k\}$ which produces low cost labels given the input feature. With respect to some measure D producing these examples we want to do the optimization:

$$\min_h E_{(x,c_1,c_2,\ldots,c_k)\sim D} c_{h(x)}$$

Here the optimization (or heuristic optimization) is over some unspecified and algorithm dependent set of classifiers.

1.2. Binary Cost Sensitive Classification.

In binary cost sensitive classification, labeled examples have the form (x, y, i) where $x \in X$ is an arbitrary feature, $y \in \{0, 1\}$ is a binary label and $i \in [0, \infty)$ is the "importance" of classifying correctly. To solve this problem, we must find a hypothesis $h : X \to \{0, 1\}$ which is correct on high importance examples. With respect to some measure D producing these examples, we want to solve the optimization:

$$\min_h E_{x,y,i\sim D} iI(h(x) \neq y)$$

Again, this (heuristic) optimization is according to some learning algorithm.

2. A Reduction

Let's attempt a reduction of Multiclass Cost Sensitive Classification to a single instance of Binary cost sensitive classification. In order to do this prediction, we must settle on a method for converting a binary classification into a multiclass classification. The following algorithm seems reasonable:

Predict(classifier h, example x)

(1) $P = \{\}$
(2) for every $l = 1$ to k do
    (a) if $h(x, l) = 1$ then $P \leftarrow P \cup \{l\}$
(3) if $|P| > 0$ then return a uniform draw from P
(4) else return a uniform draw from $\{1, ..., k\}$

This procedure implicitly specifies how errors in the binary classification transform into errors for the multiclass classification. This transformation, in turn, can tell us what the good values for the importance i in the binary classification problem are.

Our reduction maps a multiclass cost sensitive example $(x, c_1, c_2, ..., c_k)$ to a set of binary cost sensitive examples. We start by creating one example for every label. The input feature is (x, l) and the output feature is $y_l = I(\frac{1}{k}\sum_{l'=1}^{k} c_{l'} - c_l \geq 0)$ and the importance is $i_l = |\frac{1}{k}\sum_{l'=1}^{k} c_{l'} - c_l|$[1].

These are not necessarily the "right" values for the importance and label because, for example, our classifier might be capable of distinguishing the best possible label from all other labels. We can't know in advance how powerful our classifier is, so we need an adaptive system to improve our classification. Therefore, we define an iterative process which optimizes according to the ability of our learning algorithm.

Cost_to_binary(binary cost sensitive learning algorithm A, multiclass cost sensitive dataset S)

(1) For every $(x, c_1, ..., c_k) \in S$
   (a) For every $l \in \{1, ..., k\}$ let $x_l \leftarrow (x, l)$, $y_{x,l} \leftarrow I(\frac{1}{k}\sum_{l'=1}^{k} c_{l'} - c_l \geq 0)$, $i_{x,l} \leftarrow \frac{1}{k}\sum_{l'=1}^{k} c_{l'} - c_l|$
(2) Let $S' \leftarrow \{\forall x, l \in \{1, ..., k\}\ (x_l, y_{x,l}, i_{x,l})\}$ (Note that for algorithms which are sensitive to requiring i.i.d. data, we'll only include one sample drawn uniformly from $l = \{1, ..., k\}$.)
(3) Let $h \leftarrow A(S')$
(4) For every $(x, c_1, ..., c_k) \in S$
   (a) For every $l \in \{1, ..., k\}$
      (i) then $i_l \leftarrow E_{l\sim\text{Predict}(h_{\bar{l}},x)} c_l - E_{l\sim\text{Predict}(h_l,x)} c_l$
      (ii) $y_l \leftarrow I(i_l > 0)$
      (iii) $i_l \leftarrow |i_l|$
   (b) let $n_x = |\{l : y_l = 1\}|$
   (c) For every $l \in \{1, ..., k\}$ if $y_l = 1$ then $i_l \leftarrow \frac{\frac{1}{k}\sum_{l'=1}^{k} c_{l'}}{n_x}$
(5) If $E_{(x,c_1,...,c_k)\sim S} E_{l\sim\text{Predict}(h,x)} c_l$ did not improve much then terminate otherwise go to 2

This process has the following properties.

Theorem 2.1. *(optimality) If A has loss rate ϵ and* ***Cost_to_binary*** *converges, then:*

$$E_{x,c_1,c_2,...,c_k\sim D} E_{l\sim Predict(h,x)} c_l = E_{x\sim D} \sum_{y_l, i_l\sim\hat{D}|x} i_l I(h(x) \neq y_l) \leq k E_{x,y_l,i_l\sim\hat{D}} i_l I(h(x) \neq y) = k\epsilon$$

Proof. There are two cases. The first is when for all l, $h(x, l) = 0$. In this case, all errors are false negatives, and we have: $E_{l\sim\text{Predict}(h,x)} c_l = \frac{1}{k}\sum_{l'=1}^{k} c_{l'}$. We also have that :

$$\sum_{y_l, i_l\sim\hat{D}|x} i_l I(h(x \neq y_l) = \frac{1}{k}\sum_{l'=1}^{k} c_{l'}$$

[1]The logic here is that this initial choice is always discriminative (at least one label is 1 and one label is 0) whenever the costs differ. In addition, this seems reasonable for the binary case.

finishing the case. The next case is when there exists l such that $h(x,l)=1$. In this case, $E_{l\sim\text{Predict}(h,x)}c_l = E_{l:h(x,l)=1}c_l$ and $\sum_{y_l,i_l\sim\hat{D}|x} i_l I(h(x)\neq y_l) = \sum_{l|y_l\neq h(x,l)} E_{l\sim\text{Predict}(h,x)}c_l - E_{l\sim\text{Predict}(h_l,x)}c_l$. □

Theorem 2.2. *(convergence) For any fixed error rate, $E_{(x,c_1,\ldots,c_k)\sim D}E_{l\sim Predict(h,x)}c_l$ decreases with each iteration until convergence.*

Proof. I'm not sure how to prove this yet, but something like this should hold. □