Confidence assignment based on game theoretic analysis

Yoav Freund

December 22, 2013

- Suppose we have finite set of classification rules $F = (f_1, \ldots, f_m)$ that map instances $x \in X$ to $y \in \{-1, +1\}$.
- Let D be a fixed but unknown (to the algorithm) distribution over (x, y) pairs, where $x \in X$ and $y \in \{-1, +1\}$.
- We work in a transductive framework. In other words, we have n unlabeled examples (instances) $x_i \in X$, denoted $T = \{x_1, \ldots, x_n\}$.
- We assume that some of the functions in F have non-zero correlation with the true distribution over the labels, we denote these correlations (or a lower bound on the correlation) by

$$\forall 1 \leq i \leq m \;\; c_i \leq \frac{1}{n} \sum_{x \in T} E_{y \sim D(x)} \left[y \cdot f_i(x) \right]$$

- The goal of the algorithm is to find a prediction function $g : T \to [-1, +1]$ that maximizes the worst case correlation between g and the true label.

1 Formal Setup

Let $z_i = E_{(x_i, y_i)} (y_i | x_i)$ be the expected value of the label associated with instance x_i. Clearly $-1 \leq z_i \leq 1$. For reasons that have to do with the cannonical representation of linear programs, we partitions each conditional probability into two positive terms: $z_i = z_i^+ - z_i^-$, where $0 \leq z_i^+, z_i^- \leq 1$. We denote by $\mathbf{z}$ the $2n$ dimensional column vector: $\mathbf{z}^T = (z_1^+, z_1^-, \ldots z_n^+, z_n^-)$

Similarly, we use $g_i = g_i^+ - g_i^-$ to denote the predictions made by the algorithm. Again $0 \leq g_i^+, g_i^- \leq 1$ and $\mathbf{g}^T \doteq (g_1^+, g_1^-, \ldots, g_n^+, g_n^-)$.

Finally we use the vector $\mathbf{c}$ to denote the column vector of the function correlations c_i.

The correlation between the prediction vector $\mathbf{g}$ and the conditional probability $\mathbf{z}$ is the inner product $\mathbf{z}^T \mathbf{g}$. The goal of the algorithm is to maximize the correlation and the goal of nature is to minimize it. As we (or the algorithm) are interested in maximizing the worst case performance (over the choices of $\mathbf{z}$). We can formalize the optimization problem faced by the algorithm as

$$\max_{\mathbf{g}} \quad \min_{\mathbf{z}} \mathbf{z}^T \mathbf{g}$$

Where $\mathbf{g}, \mathbf{z} \in [0, 1]^n$ and $\mathbf{z}$ is further constrained by the functions in F.

We denote by $\mathbf{F}$ the matrix the contains the prediction of $(f_1, \ldots, f_m)$ on the instances $(x_1, \ldots, x_n)$. To match the fact that $\mathbf{g}$ and $\mathbf{z}$ have two entries for each x_i we similarly double the number of rows in $\mathbf{F}$, getting the following $2n \times m$ matrix:

$$\mathbf{F} = \begin{pmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_m(x_1) \\ -f_1(x_1) & -f_2(x_1) & \cdots & -f_m(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_m(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(x_n) & f_2(x_n) & \cdots & f_m(x_n) \\ -f_1(x_n) & -f_2(x_n) & \cdots & -f_m(x_n) \end{pmatrix} \tag{1}$$

We can represent the m constraints on $\mathbf{z}$ defined by $(f_1, \ldots, f_m)$, as

$$\mathbf{z}^T\mathbf{F} \geq n\mathbf{c}$$

We represent the constraints $\mathbf{z} \leq \mathbb{1}^{2n}$ as a larger-or-equal constraint

$$\mathbf{z}^T(-\mathbf{I}) \geq -\mathbb{1}^{2n}$$

Where $\mathbb{1}^{2n}$ denote a row vector of length $2n$ all of which entries are equal to 1.

2 The optimization problem

We can rewrite the optimization problem in matrix notation as follows. We use $\mathbf{b}, \mathbf{c}, \ldots$ to denote column vectors. We use the notation $(\mathbf{b}^T, \mathbf{c}^T)$ to denote the row vector which is the concatenation of $\mathbf{b}^T$ and $\mathbf{c}^T$.

The problem is

$$\begin{aligned} \text{Find:} \quad & \max_{\mathbf{g}} \ \min_{\mathbf{z}} \mathbf{z}^T\mathbf{g} \\ \text{Such That:} \quad & \mathbf{z}^T\mathbf{A} \geq \mathbf{d}^T \text{ and } \mathbf{z} \geq \mathbf{0} \\ & -\mathbf{g} \geq -\mathbb{1}^{2n} \text{ and } \mathbf{g} \geq \mathbf{0} \end{aligned} \tag{2}$$

Where

$$\mathbf{d}^T = (n\mathbf{c}^T, -\mathbb{1}^{2n}) \tag{3}$$

and $\mathbf{A}$ is the $2n \times (m + 2n)$ matrix:

$$\mathbf{A} = (\mathbf{F}, -\mathbf{I}) = \begin{pmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_m(x_1) & -1 & 0 & 0 & \cdots & 0 & 0 \\ -f_1(x_1) & -f_2(x_1) & \cdots & -f_m(x_1) & 0 & -1 & 0 & \cdots & 0 & 0 \\ f_1(x_2) & f_2(x_2) & \cdots & f_m(x_2) & 0 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ f_1(x_n) & f_2(x_n) & \cdots & f_m(x_n) & 0 & 0 & 0 & \cdots & -1 & 0 \\ -f_1(x_n) & -f_2(x_n) & \cdots & -f_m(x_n) & 0 & 0 & 0 & \cdots & 0 & -1 \end{pmatrix} \tag{4}$$

2.1 Transforming the optimization problem into an LP

Suppose we fix $\mathbf{g}$, the internal maximization problem is an LP:

$$\begin{aligned} \text{Find:} \quad & \min_{\mathbf{z}} \mathbf{z}^T\mathbf{g} \\ \text{Such That:} \quad & \mathbf{z}^T\mathbf{A} \geq \mathbf{d}^T \text{ and } \mathbf{z} \geq \mathbf{0} \end{aligned} \tag{5}$$

We can write the maximization LP that is the dual to this minimization LP:

$$\begin{aligned} \text{Find:} \quad & \max_{\mathbf{v}} \mathbf{d}^T\mathbf{v} \\ \text{Such That:} \quad & \mathbf{A}\mathbf{v} \leq \mathbf{g} \text{ and } \mathbf{v} \geq \mathbf{0} \end{aligned} \tag{6}$$

Plugging (6) back into (2) we get

$$\begin{aligned} \text{Find:} \quad & \max_{\mathbf{v}} \mathbf{d}^T\mathbf{v} \\ \text{Such That:} \quad & \mathbf{A}\mathbf{v} \leq \mathbf{g} \text{ and } \mathbf{v} \geq \mathbf{0} \\ \text{And:} \quad & -\mathbf{g} \geq -\mathbb{1}^{2n} \text{ and } \mathbf{g} \geq \mathbf{0} \end{aligned} \tag{7}$$

Which is maximized when $\mathbf{g} = \mathbb{1}^{2n}$ so we can substitute for $\mathbf{g}$ and get:

$$\begin{aligned} \text{Find:} \quad & \max_{\mathbf{v}} \mathbf{d}^T\mathbf{v} \\ \text{Such That:} \quad & \mathbf{A}\mathbf{v} \leq \mathbb{1}^{2n} \text{ and } \mathbf{v} \geq \mathbf{0} \end{aligned} \tag{8}$$

3 Solving a simple case numerically

In order to gain some intuition about this problem we start by considering a very simple setup: threshold functions on the line. We define

$$f_\theta(x) = \text{sign}(x - \theta)$$

To start, consider a function set that contains a single function $F = \{f_0\}$, and suppose that the .

4 Towards an interpretation

We partition the vector $\mathbf{v}$ into two parts: $\mathbf{v}^T = (\mathbf{r}^T, \mathbf{s}^T)$, where $\mathbf{r}$ is the m dimensional vector corresponding to the m functions and $\mathbf{s}$ is the $2n$ vector corresponding to the n data points. Using this notation we can rewrite the dual LP as follows:

$$\begin{aligned} \text{Maximize:} \quad & (n\mathbf{c}^T\mathbf{r} - \|\mathbf{s}\|_1 \\ \text{Such That:} \quad & \mathbf{Fr} - \mathbf{s} \leq \mathbf{g} \\ \text{and} \quad & \mathbf{r} \geq \mathbf{0} \text{ and } \mathbf{s} \geq \mathbf{0} \end{aligned} \tag{9}$$

We can substitute the program (9) into (2) and get the following maximization LP:

$$\text{Maximize:} \quad n\mathbf{c}^T\mathbf{r} - \|\mathbf{s}\|_1 \tag{10}$$
$$\text{Such That:} \quad \mathbf{Fr} - \mathbf{s} \leq \mathbf{g} \tag{11}$$
$$\text{and} \quad \mathbf{r} \geq \mathbf{0} \text{ and } \mathbf{s} \geq \mathbf{0} \tag{12}$$
$$\text{and} \quad -\mathbf{g} \geq -\mathbb{1}^{2n} \text{ and } \mathbf{g} \geq \mathbf{0} \tag{13}$$

To maximize (10) when $\mathbf{r}$ is fixed, we want to minimize $\|\mathbf{s}\|_1$. The only constraint on $\mathbf{s}$ (other than $\mathbf{s} \geq \mathbf{0}$) is in (11), and the only constraints on $\mathbf{g}$ are in (13).

We can therefor simplify the linear program to the following form:

$$\text{Maximize:} \quad n\mathbf{c}^T\mathbf{r} - \|\mathbf{s}\|_1 \tag{14}$$
$$\text{Such That:} \quad \mathbf{Fr} - \mathbf{s} \leq \mathbb{1}^{2n} \tag{15}$$
$$\text{and} \quad \mathbf{r} \geq \mathbf{0} \text{ and } \mathbf{s} \geq \mathbf{0} \tag{16}$$
$$\tag{17}$$

For each $1 \leq i \leq 2n$ we have that

$$s_i = \max\left(0, (\mathbf{Fr})_i - 1\right)$$

From the definition of $\mathbf{F}$ we get that fo all $1 \leq i \leq n$

$$\begin{aligned} s_{2i-1} &= \max\left(0, \sum_{j=1}^{m} f_j(x_i) r_j - 1\right) \\ s_{2i} &= \max\left(0, -\sum_{j=1}^{m} f_j(x_i) r_j - 1\right) \end{aligned}$$

which implies

$$s_{2i-1} + s_{2i} = \max\left(0, \left|\sum_{j=1}^{m} f_j(x_i) r_j\right| - 1\right)$$