Confidence assignment based on game theoretic analysis

Yoav Freund

December 24, 2013

1 Setup

We assume a transductive learning model. The algorithm is first given a labeled training set and a classifier set C. The algorithm uses the training set to identify a subset $F \subseteq C$ of "good" classifiers. We will give two alternative definitions for the meaning of "good". Intuitively, both of them imply that the functions in F have errors smaller than that of random guessing.

In addition to the training set, the algorithm gets a finite test set of unlabeled instances. The goal of the algorithm is to make the min/max optimal predictions on the test set. In other words, the predictions that would guarantee the smallest expected error for the worst case distribution over the labels that conforms with the knowledge that the functions in F are "good".

1. **Function set:** A finite set of classification rules $F = (f_1, \ldots, f_m)$ that map instances $x \in X$ to $y \in \{-1, +1\}$.
2. **Test set:** n unlabeled examples (instances) $x_i \in X$, denoted $T = \{x_1, \ldots, x_n\}$. We assign to each x_i equal weight of $1/n$.
3. **Nature:** Nature chooses a conditional distribution over the label associated with each test instance
$$z_i = E_{(x_i,y_i)}\left(y_i | x_i\right)$$
4. **Good classifiers:** We make one of two types of assumption on the functions in F:
   - **Good average performance:** We are given a normalized weighting over the functions $\{q_j\}_{j=1}^m$ such that
   $$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} q_j f_j(x_i) z_i > l > 0$$
   In words, the average correlation between a randomly chosen function operating on x_i and the label selected by nature is at least l.
   This assumption matches the PAC-Bayesian analysis.
   - **Good individual performance:** We assume that we have upper and lower bounds on the error of *each* of the functions $f \in F$. We denote upper and lower bounds on these correlations by l_j, u_j:
   $$\forall 1 \le j \le m \;\; l_j \le \frac{1}{n}\sum_{i=1}^{n}\left[y_i \cdot f_j(x_i)\right] z_i \le u_i$$
5. The goal of the algorithm is to find a prediction function $g : (1, \ldots, n) \to [-1, +1]$ that maximizes the worst case (over $z_1, \ldots, z_n$) of the correlation between g and the true label: $\sum_{i=1}^{n} g_i z_i$

2 Setup using matrix notation

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 vectors $\mathbf{l}, \mathbf{u}$ to denote the column vectors defining the lower and upper bounds on the correlation of each function in F with the label.

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 upper and lower bounds on the correlations of 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 constraints on $\mathbf{z}$ defined by $(f_1, \ldots, f_m)$, as follows:

- **Average performance**
$$\mathbf{z}^T\mathbf{F}\mathbf{q} \geq nl$$
- **Individual performance**
$$\mathbf{z}^T\mathbf{F} \geq n\mathbf{l} \text{ and } \mathbf{z}^T\mathbf{F} \leq n\mathbf{u}$$

3 The optimization problem

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

The problem is

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

Where $\mathbb{1}^{2n}$ denote a row vector of length $2n$ all of which entries are equal to 1. The vector $\mathbf{d}$ and the matrix $\mathbf{A}$ are defined differently depending on the type of performance bounds that are given.

1. **Average performance bound**

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

We define the average prediction vector:

$$\mathbf{p} \doteq \mathbf{F}\mathbf{q}$$

and $\mathbf{A}$ is a $2n \times 2n+1$ matrix:

$$\mathbf{A} = (\mathbf{p}, -\mathbf{I}) = \begin{pmatrix} \sum_{j=1}^m q_j f_j(x_1) & -1 & 0 & 0 & \cdots & 0 & 0 \\ -\sum_{j=1}^m q_j f_j(x_1) & 0 & -1 & 0 & \cdots & 0 & 0 \\ \sum_{j=1}^m q_j f_j(x_2) & 0 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \sum_{j=1}^m q_j f_j(x_n) & 0 & 0 & 0 & \cdots & -1 & 0 \\ -\sum_{j=1}^m q_j f_j(x_n) & 0 & 0 & 0 & \cdots & 0 & -1 \end{pmatrix} \tag{4}$$

2. **Individual performance bounds**

$$\mathbf{d}^T = (n\mathbf{l}^T, -n\mathbf{u}^T, -\mathbb{1}^{2n}) \tag{5}$$

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

$$\mathbf{A} = (\mathbf{F}, -\mathbf{F}, -\mathbf{I}) = \begin{pmatrix} f_1(x_1) & \cdots & f_m(x_1) & -f_1(x_1) & \cdots & -f_m(x_1) & -1 & 0 & 0 & \cdots & 0 & 0 \\ -f_1(x_1) & \cdots & -f_m(x_1) & f_1(x_1) & \cdots & f_m(x_1) & 0 & -1 & 0 & \cdots & 0 & 0 \\ f_1(x_2) & \cdots & f_m(x_2) & -f_1(x_2) & \cdots & -f_m(x_2) & 0 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ f_1(x_n) & \cdots & f_m(x_n) & -f_1(x_n) & \cdots & -f_m(x_n) & 0 & 0 & 0 & \cdots & -1 & 0 \\ -f_1(x_n) & \cdots & -f_m(x_n) & f_1(x_n) & \cdots & f_m(x_n) & 0 & 0 & 0 & \cdots & 0 & -1 \end{pmatrix} \tag{6}$$

3.1 Transforming the optimization problem into an LP

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

$$\begin{aligned} \text{Find:} \quad & \min_{\mathbf{z}} \mathbf{z}^T\mathbf{g} \\ \text{Such That:} \quad & \mathbf{z}^T\mathbf{A} \geq \mathbf{d}^T \text{ and } \mathbf{z} \geq \mathbf{0} \end{aligned} \tag{7}$$

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

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

Plugging (8) back into (2) we get

$$\begin{aligned} \text{Find:} \quad & \max_{\mathbf{v}} \mathbf{d}^T\mathbf{v} \\ \text{Such That:} \quad & \mathbf{A}\mathbf{v} \leq \mathbf{g} \text{ and } \mathbf{v} \geq \mathbf{0} \\ \text{And:} \quad & -\mathbf{g} \geq -\mathbb{1}^{2n} \text{ and } \mathbf{g} \geq \mathbf{0} \end{aligned} \tag{9}$$

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

$$\begin{aligned} \text{Find:} \quad & \max_{\mathbf{v}} \mathbf{d}^T\mathbf{v} \\ \text{Such That:} \quad & \mathbf{A}\mathbf{v} \leq \mathbb{1}^{2n} \text{ and } \mathbf{v} \geq \mathbf{0} \end{aligned} \tag{10}$$

4 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, suppose that

$$T = \left\{-K + \frac{1}{2}, -K + \frac{3}{2}, \ldots, -\frac{1}{2}, +\frac{1}{2}, \ldots, K - \frac{1}{2}\right\}$$

The set T contains $2K$ points, these partition the threshold functions into $2K + 1$ equivalence sets which we can represent by

$$F = \{f_\theta \,|\theta \in \{-K, -K + 1, \ldots, 0, \ldots, K - 1, K\}\}$$

the size of the function space is thus $m = 2K + 1$.

As constraints on the correlations we assume a lower bound α on the correlations of all of the functions. We assume no upper bounds on the correlations.

We have therefor a special case of the LP (10) with the following settings.

$$\mathbf{d} = (\alpha n \mathbb{1}^m, -n\mathbb{1}^m, -\mathbb{1}^{2n})$$

$$\mathbf{A} = (\mathbf{F}, -\mathbf{F}, -\mathbf{I})$$

$$\mathbf{F} = \begin{pmatrix} +1 & -1 & -1 & \cdots & -1 & -1 \\ -1 & +1 & +1 & \cdots & +1 & +1 \\ +1 & +1 & -1 & \cdots & -1 & -1 \\ -1 & -1 & +1 & \cdots & +1 & +1 \\ +1 & +1 & +1 & \cdots & -1 & -1 \\ -1 & -1 & -1 & \cdots & +1 & +1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ +1 & +1 & +1 & \cdots & -1 & -1 \\ -1 & -1 & -1 & \cdots & +1 & +1 \\ +1 & +1 & +1 & \cdots & +1 & -1 \\ -1 & -1 & -1 & \cdots & -1 & +1 \end{pmatrix}$$

5 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:

$$\begin{aligned} \text{Maximize:} \quad & (n\mathbf{d}^T\mathbf{r} - \|\mathbf{s}\|_1 \\ \text{Such That:} \quad & \mathbf{F}\mathbf{r} - \mathbf{s} \leq \mathbf{g} \\ \text{and} \quad & \mathbf{r} \geq \mathbf{0} \text{ and } \mathbf{s} \geq \mathbf{0} \end{aligned} \tag{11}$$

We can substitute the program (11) into (2) and get the following maximization LP:

$$\text{Maximize:} \quad n\mathbf{d}^T\mathbf{r} - \|\mathbf{s}\|_1 \tag{12}$$

$$\text{Such That:} \quad \mathbf{F}\mathbf{r} - \mathbf{s} \leq \mathbf{g} \tag{13}$$

$$\text{and} \quad \mathbf{r} \geq \mathbf{0} \text{ and } \mathbf{s} \geq \mathbf{0} \tag{14}$$

$$\text{and} \quad -\mathbf{g} \geq -\mathbb{1}^{2n} \text{ and } \mathbf{g} \geq \mathbf{0} \tag{15}$$

To maximize (12) 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 (13), and the only constraints on $\mathbf{g}$ are in (15).

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

$$\begin{aligned} \text{Maximize:} \quad & n\mathbf{d}^T\mathbf{r} - \|\mathbf{s}\|_1 && (16)\\ \text{Such That:} \quad & \mathbf{F}\mathbf{r} - \mathbf{s} \leq \mathbb{1}^{2n} && (17)\\ \text{and} \quad & \mathbf{r} \geq \mathbf{0} \text{ and } \mathbf{s} \geq \mathbf{0} && (18)\\ & && (19) \end{aligned}$$

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

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

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

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