Active learning using muffler

Akshay Balsubramani, Yoav Freund, Shay Moran

November 2016

1 Introduction

As a first step towards using Muffler for active learning, we describe a setup in which Muffler converges to the Bayes optimal rule.

We operate in a restricted context which emulates the kNN convergence rate analysis of Chaudhuri and Dasgupta.

2 Preliminaries

The main tools we use in this paper are linear programming and uniform convergence. We therefore use a combination of matrix notation and the probabilistic notation given in the introduction. The algorithm is first described in a deterministic context where some inequalities are assumed to hold; probabilistic arguments are used to show that these assumptions are correct with high probability.

The ensemble's predictions on the unlabeled data are denoted by $\mathbf{F}$:

$$\mathbf{F} = \begin{pmatrix} h_1(x_1) & h_1(x_2) & \cdots & h_1(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ h_p(x_1) & h_p(x_2) & \cdots & h_p(x_n) \end{pmatrix} \in [-1,1]^{p \times n} \tag{1}$$

The **true labels** on the test data U are represented by $\mathbf{z} = (z_1; \dots; z_n) \in [-1,1]^n$.

Note that we allow $\mathbf{F}$ and $\mathbf{z}$ to take any value in the range $[-1,1]$ rather than just the two endpoints. This relaxation does not change the analysis, because intermediate values can be interpreted as the expected value of randomized predictions. For example, a value of $\frac{1}{2}$ indicates $\{+1 \text{ w.p. } \frac{3}{4}, -1 \text{ w.p. } \frac{1}{4}\}$. This interpretation extends to our definition of the correlation on the test set, $\widehat{\text{corr}}_U(h_i) = \frac{1}{n}\sum_{j=1}^n h_i(x_j) z_j$. [1]

The labels $\mathbf{z}$ are hidden from the predictor, but we assume the predictor has knowledge of a **correlation vector** $\mathbf{b} \geq \mathbf{0}^n$ such that $\widehat{\text{corr}}_U(h_i) \geq b_i$ for all $i \in [p]$, i.e. $\frac{1}{n}\mathbf{F}\mathbf{z} \geq \mathbf{b}$. From our development so far, the correlation vector's components b_i each correspond to a constraint on the corresponding classifier's test error $\frac{1}{2}(1 - b_i)$.

The following notation is used throughout the paper: $[a]_+ = \max(0, a)$ and $[a]_- = [-a]_+$, $[n] = \{1, 2, \dots, n\}$, $\mathbf{1}^n = (1; 1; \dots; 1) \in \mathbb{R}^n$, and $\mathbf{0}^n$ similarly. Also, write I_n as the $n \times n$ identity matrix. All vector inequalities are componentwise. The probability simplex in d dimensions is denoted by $\Delta^d = \{\sigma \geq \mathbf{0}^d : \sum_{i=1}^d \sigma_i = 1\}$. Finally, we use vector notation for the rows and columns of $\mathbf{F}$: $\mathbf{h}_i = (h_i(x_1), h_i(x_2), \cdots, h_i(x_n))^\top$ and $\mathbf{x}_j = (h_1(x_j), h_2(x_j), \cdots, h_p(x_j))^\top$.

[1] We are slightly abusing the term "correlation" here. Strictly speaking this is just the expected value of the product, without standardizing by mean-centering and rescaling for unit variance. We prefer this to inventing a new term.

3 The Transductive Binary Classification Game

We now describe our prediction problem, and formulate it as a zero-sum game between two players: a predictor and an adversary.

In this game, the predictor is the first player, who plays $\mathbf{g} = (g_1; g_2; \ldots; g_n)$, a randomized label $g_i \in [-1, 1]$ for each example $\{\mathbf{x}_i\}_{i=1}^n$. The adversary then plays, setting the labels $\mathbf{z} \in [-1, 1]^n$ under ensemble test error constraints defined by $\mathbf{b}$. The predictor's goal is to minimize (and the adversary's to maximize) the *worst-case expected classification error on the test data* (w.r.t. the randomized labelings $\mathbf{z}$ and $\mathbf{g}$): $\frac{1}{2}\left(1 - \frac{1}{n}\mathbf{z}^\top \mathbf{g}\right)$. This is equivalently viewed as maximizing worst-case correlation $\frac{1}{n}\mathbf{z}^\top \mathbf{g}$.

To summarize concretely, we study the following game:

$$V := \max_{\mathbf{g} \in [-1,1]^n} \min_{\substack{\mathbf{z} \in [-1,1]^n, \\ \frac{1}{n}\mathbf{F}\mathbf{z} \in [\mathbf{b}_l, \mathbf{b}_u]}} \frac{1}{n}\mathbf{z}^\top \mathbf{g} \tag{2}$$

It is important to note that we are only modeling "test-time" prediction, and represent the information gleaned from the labeled data by the parameter $\mathbf{b}$. Inferring the vector $\mathbf{b}$ from training data is a standard application of Occam's Razor [**?**], which we provide in Section **??**.

The minimax theorem (e.g. [**?**], Theorem 7.1) applies to the game (2), since the constraint sets are convex and compact and the payoff linear. Therefore, it has a minimax equilibrium and associated optimal strategies $\mathbf{g}^*, \mathbf{z}^*$ for the two sides of the game, i.e. $\min_{\mathbf{z}} \frac{1}{n}\mathbf{z}^\top \mathbf{g}^* = V = \max_{\mathbf{g}} \frac{1}{n}{\mathbf{z}^*}^\top \mathbf{g}$.

As we will show, both optimal strategies are simple functions of a particular *weighting* over the p hypotheses – a nonnegative p-vector. Define this weighting as follows.

Definition 1 (**Slack Function and Optimal Weighting**). *Let $\sigma \geq 0^p$ be a weight vector over $\mathcal{H}$ (not necessarily a distribution). The vector of **ensemble predictions** is $\mathbf{F}^\top \sigma = (\mathbf{x}_1^\top \sigma, \ldots, \mathbf{x}_n^\top \sigma)$, whose elements' magnitudes are the **margins**. The **prediction slack function** is*

$$\gamma(\sigma, \mathbf{b}) = \gamma(\sigma) := \frac{1}{n}\sum_{j=1}^n \left[\left|\mathbf{x}_j^\top \sigma\right| - 1\right]_+ - \mathbf{b}^\top \sigma \tag{3}$$

*An **optimal weight vector** σ^* is any minimizer of the slack function: $\sigma^* \in \arg\min_{\sigma \geq 0^p} [\gamma(\sigma)]$.*

Our main result uses these to describe the solution of the game (2).

Theorem 2 (Minimax Equilibrium of the Game)**.** *The minimax value of the game* (2) *is $V = -\gamma(\sigma^*)$. The minimax optimal strategies are defined as follows: for all $i \in [n]$,*

$$g_i^* \doteq g_i(\sigma^*) = \begin{cases} \mathbf{x}_i^\top \sigma^* & \left|\mathbf{x}_i^\top \sigma^*\right| < 1 \\ \operatorname{sgn}(\mathbf{x}_i^\top \sigma^*) & \textit{otherwise} \end{cases} \qquad \textit{and} \qquad z_i^* = \begin{cases} 0 & \left|\mathbf{x}_i^\top \sigma^*\right| < 1 \\ \operatorname{sgn}(\mathbf{x}_i^\top \sigma^*) & \left|\mathbf{x}_i^\top \sigma^*\right| > 1 \end{cases} \tag{4}$$

The proof of this theorem is a standard application of Lagrange duality and the minimax theorem. The minimax value of the game and the optimal strategy for the predictor $\mathbf{g}^*$ (Lemma **??**) are our main objects of study and are completely characterized, and the theorem's partial description of $\mathbf{z}^*$ (proved in Lemma **??**) will suffice for our purposes. [2]

Theorem 2 illuminates the importance of the optimal weighting σ^* over hypotheses. This weighting $\sigma^* \in \arg\min_{\sigma \geq 0^p} \gamma(\sigma)$ is the solution to a convex optimization problem (Lemma **??**), and therefore we can

[2] For completeness, Corollary **??** in the appendices specifies z_i^* when $\left|\mathbf{x}_i^\top \sigma^*\right| = 1$.

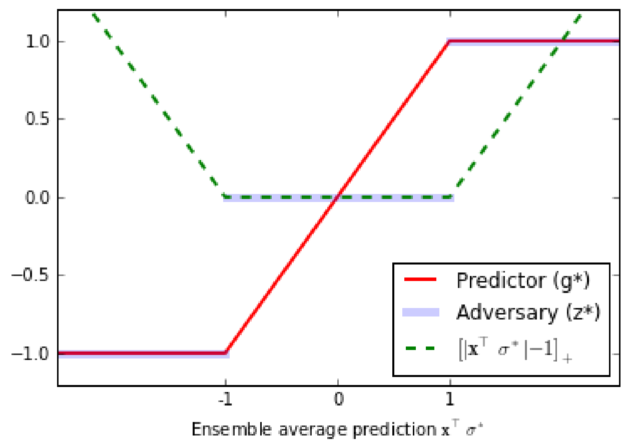

Figure 1: The optimal strategies and slack function as a function of the ensemble prediction $\mathbf{x}^\top \sigma^*$.

efficiently compute it and $\mathbf{g}^*$ to any desired accuracy. The ensemble prediction (w.r.t. this weighting) on the test set is $\mathbf{F}^\top \sigma^*$, which is the only dependence of the solution on $\mathbf{F}$.

More specifically, the minimax optimal prediction and label (4) on any test set example $\mathbf{x}_j$ can be expressed as functions of the ensemble prediction $\mathbf{x}_j^\top \sigma^*$ on that test point alone, without considering the others. The $\mathbf{F}$-dependent part of the slack function also depends separately on each test point's ensemble prediction. Figure 1 depicts these three functions.

4 Ball Specialists

We restrict our attention to a special case which corresponds, roughly, to nearest neighbor methods.

1. The input space $\mathcal{X}$ is a finite set in R^d. We assume a uniform distribution over $\mathcal{X}$. [3]

2. The rules that we use are "specialists" that are balls. The set $\mathcal{B}$ contains all rules of the form

$$B_{r,\vec{c},s}(\vec{x}) = \begin{cases} s & \text{if } \|\vec{c} - \vec{x}\| \le r \\ 0 & \text{otherwise} \end{cases}$$

   Where $r \ge 0$ is the radius of the ball, $\vec{c} \in R^d$ is the center of the ball and $s \in \{-1, +1\}$ is the polarity of the ball. We will drop the subscripts of B when clear from context.

3. We use $\vec{x} \in B$ to indicate that $B(\vec{x}) \neq 0$.

4. We denote the *probability* of a ball B by $p(B) \doteq \frac{|B|}{|\mathcal{X}|}$

5. We use the term *bias* of a ball to refer to the conditional expectation of the label for a ball by

$$\text{bias}(B) \doteq E\left(y | \vec{x} \in B\right)$$

[3] We use an arrow notation $\vec{x}$ to differentiate between $\vec{x} \in R^d$ and $\mathbf{x}$ which denotes a row of the matrix $\mathbf{F}$.

5 Degrees of Safety

We say that a point $\vec{x} \in \mathcal{X}$ is *safe* if we can confidently identify the label of $\vec{x}$. We distinguish three levels of safety of increasing strength: version-space (VS) safety, pairwise (PW) safety and asymptotic (A) safety. We define each one in turn.

First, we need some notation. We denote the set of all balls by $\mathcal{S}$. For any $\epsilon, \gamma > 0$ and $s \in \{-1, +1\}$ we define the set of (ϵ, γ, s)- good balls $\mathcal{S}^s_{\epsilon,\gamma} \subset \mathcal{S}$ to be:

$$\mathcal{S}^s_{\epsilon,\gamma} \doteq \{B \in \mathcal{S} \,|p(B) \geq \epsilon, s \text{ bias}(B) > \gamma\}$$

We define $\mathcal{S}^{\pm}_{\epsilon,\gamma} \doteq \mathcal{S}^{+}_{\epsilon,\gamma} \cup \mathcal{S}^{-}_{\epsilon,\gamma}$

Denote by $V(\mathcal{S}^{\pm}_{\epsilon,\gamma})$ the set of all point-wise biases $\mathbf{z}$ that satisfy the constraints defined by $\mathcal{S}^{\pm}_{\epsilon,\gamma}$

- **Version space (VS) safety**
  We are (ϵ, γ, s)-**VS safe** on $\vec{x}$ if $s \cdot \text{sign}(\mathbf{z}^*(\vec{x})) \geq 0$ for $\mathbf{z}^*$ that satisfy $\frac{1}{n}\mathbf{F}\mathbf{z}^* \geq \mathbf{b}$ and are min/max optimal.

- **Pair-Wise (PW) safety**
  We define the set of (ϵ, γ, s)-**PW safe** instances to be

$$\mathcal{X}^s_{\epsilon,\gamma} \doteq \left\{ \vec{x} \in R^d \;\middle|\; \begin{array}{l} \exists B \in \mathcal{S}^s_{\epsilon,\gamma} \text{ s.t. } \vec{x} \in B \text{ and} \\ \forall B \in \mathcal{S}^{-s}_{\epsilon,\gamma} \text{ s.t. } \vec{x} \in B, \;\; \exists B' \in \mathcal{S}^s_{\epsilon,\gamma} \text{ s.t. } \vec{x} \in B' \text{ and } B' \subset B \end{array} \right\}$$

- **Asymptotic (A) Safety**
  We say that $\vec{x}$ is (ϵ, γ, s)- **A-safe** if it is (ϵ, γ', s)-**PW safe** for all $0 < \epsilon' \leq \epsilon$ and $0 < \gamma' \leq \gamma$ and for the same polarity s.

6 Pairwise safety implies version space safety

Fix a point $\vec{x}$ and the parameters $(\epsilon, edge, s)$. Let $\mathcal{A}(\vec{x}, \epsilon, \gamma)$ be the sets of all balls B that contain $\vec{x}$, have weight $\epsilon > 0$ and edge γ with respect to *some* polarity $s \in \{-1, +1\}$. In other words:

$$\mathcal{A}(\vec{x}, \epsilon, \gamma) \doteq \left\{ B \;\middle|\; \frac{|B|}{|\mathcal{X}|} \geq \epsilon \quad \text{and} \quad \exists s \in \{-1, +1\} \quad \text{such that} \quad \frac{s}{|B|} \sum_{\vec{x} \in B} \mathbf{z}(\vec{x}) \geq \gamma \right\}$$

Consider the partial order of containment defined over the balls in $\mathcal{A}(\vec{x}, \epsilon, \gamma)$. Let the "set of minima" $\mathcal{M}(\vec{x}, \epsilon, \gamma) \subseteq \mathcal{A}(\vec{x}, \epsilon, \gamma)$ be the set of balls that are minimial with respect to this partial order. An alternative specification of pairwise safety is that that all balls in $\mathcal{M}(\vec{x}, \epsilon, \gamma)$ set have the same polarity s. More formally, $\vec{x}$ is (ϵ, γ, s)-pairwise safe if and only if

$$\forall B \in \mathcal{M}(\vec{x}, \epsilon, \gamma), \quad \frac{s}{|B|} \sum_{\vec{x} \in B} \mathbf{z}(\vec{x}) \geq \gamma$$

Before proving that Pairwise Safety implies Version Space safety, we need the following technical lemma:

Lemma 6.1. *Let $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$ be a finite collection of non-empty sets over a finite domain. Then there exist a set of at most n points $x_1, \ldots, x_m$ such that each set in $\mathcal{A}$, contains exactly one point.*

Proof. Denote by $\mathcal{C}$ a collection of sets. We use the notation $\cap\mathcal{C}$ to denote the intersection of the sets in $\mathcal{C}$.

The proof is constructive and recursive. We start with $\mathcal{C} = \mathcal{A}$ and continue until $\mathcal{C}$ is empty.

At each stage of the recursion we distinguish two cases:

1. $\cap\mathcal{C} \neq \emptyset$. In this case We choose $\vec{x}$ from the intersection of all sets and we are done.

2. $\cap\mathcal{C} = \emptyset$. In this case we partition $\mathcal{C}$ into two disjoint collections $\mathcal{D}$ and $\mathcal{F}$, such that $\cap\mathcal{D} \neq \emptyset$ and for all $A \in \mathcal{F}$, $\cap\mathcal{D} \cap A = \emptyset$. Note that $x \in A$ for all $A \in \mathcal{D}$ and $x \notin A$ for $A \in \mathcal{F}$. We choose an arbitrary x element from $\cap\mathcal{D}$ and we can remove $\mathcal{D}$ from consideration and continue recursively with $\mathcal{C} = \mathcal{F}$.

   The construction of the collection $\mathcal{D}$ is greedy. We start with an arbitrary set A_1 in $\mathcal{C}$, by assumption, this set is not empty. We then repeatedly add sets to $\mathcal{D}$ as long as their addition does not result in $\cap\mathcal{D} \neq \emptyset$. As $\cap\mathcal{C} = \emptyset$ we know that this process must at some point before $\mathcal{D} = \mathcal{C}$. We define $\mathcal{D}$ to be a collection of sets whose intersection is not empty such that the addition of any set will make the intersection empty.

   In other words, any point x chosen from $\cap\mathcal{D}$ is a member of $A \in \mathcal{D}$ and is not a member of $\in \mathcal{F}$

$\square$

For a long time I (YF) thought that the following is true, but it is not.

Claim 6.1. *If* $\mathbf{x}$ *is pairwise safe for some* ϵ, γ, s *then it is version space safe.*

Counter Example

Consider a space with 4 points, and associate with the points the adversarial values: z_1, z_2, z_3, z_4 consider the following three constraints:

$$P_1 : \quad \frac{1}{2}(z_1 + z_2) \geq 0.1 \tag{5}$$

$$P_2 : \quad \frac{1}{2}(z_2 + z_3) \geq 0.1 \tag{6}$$

$$N_1 : \quad \frac{1}{4}(z_1 + z_2 + z_3 + z_4) \leq -0.25 \tag{7}$$

Solving just for the two positive constraints: P_1, P_2 we find that the min/max solution is $z_1 = z_3 = 0$ and $z_2 = 0.2$.

All three points: $1, 2, 3$ are pairwise safe. The negative constraint is a super-set of the positive constaints. However, z_1, z_3 are not version-space safe. To see that consider the negative constraint N_1. As $z_4 \geq -1$ we get that:

$$z_1 + z_2 + z_3 + z_4 \leq -1$$
$$z_1 + z_2 + z_3 \leq 0$$

But for the solution for P_1, P_2 yields $z_1 + z_2 + z_3 = 0.2 > 0$, so that solution is not valid when N_1 is added. A solution that holds is $z_1 = z_3 = -0.2$, $z_2 = 0.2$.

But in that solution the signs of z_2, z_3 have been flipped.