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 $\mathbf{g} = (g_1; g_2; \dots; g_n)$, 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 $\mathbf{F}^\top \sigma = (\mathbf{x}_1^\top \sigma, \dots, \mathbf{x}_n^\top \sigma)$, 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: $\sigma^* \in \underset{\sigma \geq 0^p}{\arg\min}\,[\gamma(\sigma)]$.*

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 $i \in [n]$,*

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

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 $\left|\mathbf{x}_i^\top \sigma^*\right| = 1$.

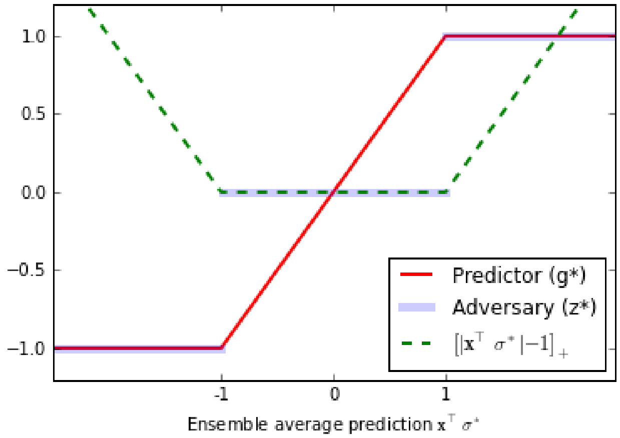

Figure 1: The optimal strategies and slack function as a function of the ensemble prediction $\mathbf{x}^\top \sigma^*$.

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

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

   Where $r \geq 0$ 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.

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

3. We denote by $\mathcal{D}$ the distribution of examples $(\mathbf{x}, y) \sim R^d \times \{-1, +1\}$.

4. We denote the *probability* of a ball B by $p(B) \doteq P_{\mathbf{x} \sim \mathcal{D}}(\mathbf{x} \in B)$

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

$$\text{bias}(B) \doteq E_{(\mathbf{x},y) \sim \mathcal{D}}\left(y | \mathbf{x} \in B\right)$$

6. We define the *Bayes optimal* rule for the distribution $\mathcal{D}$ to be $O(\mathbf{x}) = \text{sign}(\text{bias}(\mathbf{x}))$ and the Bayes optimal error to be $\text{err}^* = \text{err}(O(\mathbf{x}))$

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 the set of (ϵ, γ, s)-good instances to be:

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

For a given ϵ, γ we define $\mathcal{S}_{\epsilon,\gamma} = \mathcal{S}^{+1}_{\epsilon,\gamma} \cup \mathcal{S}^{-1}_{\epsilon,\gamma}$

We denote the solution of the the game 2 (now defined using integrals rather than finite summation) by $\mathbf{SOL}(\mathcal{S}_{\epsilon,\gamma})$, and the error of this solution by $\text{err}_{\epsilon,\gamma} \doteq \text{err}(\mathbf{SOL}(\mathcal{S}_{\epsilon,\gamma}))$.

assumption 4.1. *The error $err_{\epsilon,\gamma}$ is well defined for all $\epsilon, \gamma > 0$ and*

$$\lim_{\epsilon \to 0, \gamma \to 0} err_{\epsilon,\gamma} = err^*$$

5 Good Examples are Clipped

Finally, we relate these definitions back to the slack function optimization problem. A definition that is useful is:

Definition 3. *For $s = \pm 1$, define*

$$T^s_{\epsilon,\gamma}(\vec{x}) := \left\{ B \in \mathcal{S}^s_{\epsilon,\gamma} \text{ s.t. } \vec{x} \in B \right\}$$

5.1 Two Change-of-Basis Lemmas

Lemma 5.1. *Suppose the slack function over a set of specialists $\mathcal{S}^s_{\epsilon,\gamma}$ is minimized by giving each example x a score of $S^*(x)$. Take any $B \in \mathcal{S}^s_{\epsilon,\gamma}$, and suppose that $\exists B' \in \mathcal{S}^s_{\epsilon,\gamma}$ s.t. $B' \subset B$ and $B'(x) \neq B(x) \;\; \forall x \in B'$.*

Define the specialist C predicting only on $B \setminus B'$ according to B. Then the slack function over the specialist class

$$\mathcal{S}^s_{\epsilon,\gamma} \setminus \{B\} \cup \{C\}$$

has the same minimizer, with each example's score being $S^*(x)$.

Proof. Note that B' and B influence the slack function only through the corresponding weights put on them by the algorithm $\sigma_{B'}, \sigma_B \geq 0$. Suppose the predictions of B', B are $\mathbf{h}_{B'}, \mathbf{h}_B \in \mathbb{R}^n$. Recall that $h_{B',i} = \frac{1}{p(B')} B'(x_i)$ where $B'(x_i) \in \{-1, 0, 1\}$. Then B, B' are represented together in the optimization by $\sigma_{B'} \mathbf{h}_{B'} + \sigma_B \mathbf{h}_B$. This vector can be rewritten in terms of the two hypotheses B' and C, where C predicts according to B on its domain $B \setminus B'$.

$$\begin{aligned} \sigma_{B'} h_{B',i} + \sigma_B h_{B,i} &= \frac{\sigma_{B'}}{p(B')} B'(x_i) + \frac{\sigma_B}{p(B)} B(x_i) \\ &= \left(\frac{\sigma_{B'}}{p(B')} - \frac{\sigma_B}{p(B)} \right) B'(x_i) \mathbf{1}(x_i \in B') + \frac{\sigma_B}{p(B)} B(x_i) \mathbf{1}(x_i \in B \setminus B') \\ &= \left(\sigma_{B'} - \sigma_B \frac{p(B')}{p(B)} \right) h_{B',i} + \sigma_B \frac{p(C)}{p(B)} h_{C,i} \end{aligned}$$

Thus optimizing with $\mathbf{h}_{B'}$ and $\mathbf{h}_B$ is equivalent to optimizing with $\mathbf{h}_{B'}$ and $\mathbf{h}_C$, and the conclusion follows. □

Lemma 5.2. *For $s = \pm 1$, suppose the slack function over a set of specialists $\mathcal{S}^s_{\epsilon,\gamma}$ is minimized by giving each example x a score of $S^*(x)$. Take any $B, B' \in \mathcal{S}^s_{\epsilon,\gamma}$ with the same polarity, and define the specialists C, C', C'' predicting with this polarity on $(B \setminus B'), (B' \setminus B), (B \cap B')$ respectively. Then the slack function over the specialist class*

$$\mathcal{S}^s_{\epsilon,\gamma} \setminus \{B, B'\} \cup \{C, C', C''\}$$

has the same minimizer, with each example's score being $S^*(x)$.

Proof. Similar to above. □

5.2 Main Result

Lemma 5.3. *For any $\epsilon > 0, \gamma > 0, s = \pm 1$, take any $\vec{x} \in \mathcal{X}^s_{\epsilon,\gamma}$. If $\sigma^* \geq \mathbf{0}^{|\mathcal{S}^s_{\epsilon,\gamma}|}$ is the minimizer of the slack function over good specialists $\mathcal{S}^s_{\epsilon,\gamma}$, and $\mathbf{x} \in \mathbb{R}^{|\mathcal{S}^s_{\epsilon,\gamma}|}$ is the vector of these specialists' predictions on $\vec{x}$,*

$$s\mathbf{x}^\top \sigma^* \geq 1 \tag{5}$$

Proof. For the given $\mathbf{x}$, by definition of $\mathcal{X}^s_{\epsilon,\gamma}$, any $B \in T^{-s}_{\epsilon,\gamma}(\vec{x})$ has a corresponding $B' \subset B$ such that $B' \in T^s_{\epsilon,\gamma}(\vec{x})$. Lemma 5.1 shows that the slack function problem is equivalent to one where B is replaced in the ensemble by a specialist corresponding to $C := B \setminus B'$, which does not contain $\vec{x}$. Replacing each $B \in T^{-s}_{\epsilon,\gamma}(\vec{x})$ in this way removes it from $T^{-s}_{\epsilon,\gamma}(\vec{x})$ but conserves $T^s_{\epsilon,\gamma}(\vec{x})$, to catalyze replacement of other elements of $T^{-s}_{\epsilon,\gamma}(\vec{x})$. (Note that C may not be in $\mathcal{S}^s_{\epsilon,\gamma}$ because we could have $p(C) \leq \epsilon$.) When all $B \in T^{-s}_{\epsilon,\gamma}(\vec{x})$ are replaced, $T^{-s}_{\epsilon,\gamma}(\vec{x})$ is empty, but $T^s_{\epsilon,\gamma}(\vec{x})$ is identical, and contains all specialists in the new basis that predict on $\vec{x}$.

Lemma 5.2 shows that any intersecting pair of specialists in $T^s_{\epsilon,\gamma}(\vec{x})$ can be replaced by three disjoint specialists of the same polarity without changing the minimization solution. Performing this replacement repeatedly, we eventually arrive at a basis in which exactly one specialist predicts on x, and potentially on other examples as well, for which it must be the only predictor. Call this specialist h. Putting weight on h will always lower the slack function if the examples in its support are hedged, proving the result. □

Akshay: Not sure about this because h may have bias < 0 even though its parent specialists do not.

6 Intuition

Here is some rough intuition. One can think of $\mathcal{S}^{+1}_{\epsilon,\gamma}$ and $\mathcal{S}^{-1}_{\epsilon,\gamma}$ as competing teams of specialists, those in the first team predict $+1$ (when they don't abstain) while those in the other predict -1. Consider a particular point $\mathbf{x}$, and consider subset of the specialists in $\mathcal{S}_{\epsilon,\gamma}$ that predict (i.e. do not abstain) on $\mathbf{x}$.

It is intuitively clear that if all of the predicting specialists are members one team, that the solution $\mathbf{SOL}$ will predict with the same label as that team.

A less intuitive fact is that there are points for which there are predictors from both teams but are still guaranteed to predict according to one of the teams. These are the elements of $\mathcal{X}^s_{\epsilon,\gamma}$. Suppose $\sigma = +1$ and $\mathbf{x}\mathcal{X}^{+1}_{\epsilon,\gamma}$. There might be specialists $B \in \mathcal{S}^{-1}_{\epsilon,\gamma}$such that $\mathbf{x} \in B$, but for each such specialist there is another specialist $B' \in \mathcal{S}^{+1}_{\epsilon,\gamma}$ such that $\mathbf{x} \in B'$ *and* $B' \subset B$. Intuitively, the opinion of the more specialized specialist masks the opinion of the more general one.

In terms of the game, the points in $\mathcal{X}_{\epsilon,\gamma}$ are clipped points. I am not sure whether points outside of $\mathcal{X}_{\epsilon,\gamma}$ can also be clipped.

Predictions on $\mathbf{x} \in \mathcal{X}_{\epsilon,\gamma}$ can change only if there is some $\epsilon' \leq \epsilon$ and $\gamma' \leq \gamma$ such that for some $B \in \mathcal{S}_{\epsilon',\gamma'}$ B belongs to the opposite team. Thus if we have reached the probability radius ρ such that for all B where $\mathbf{x} \in B$ and $p(B) < \rho$ the sign of $B(vx)$ is constant, we have reached the Bayes rule.

We can therefor call $\mathbf{x} \in \mathcal{X}_{\epsilon,\gamma}$ the "unknown unknown", while the complement of $\mathbf{x}$ are the "known unknown".

7 From true dist to empirical

Suppose now that we take a sample of n points, and suppose that, at each point of time, we have a sample of labels associated with these points.

We can define the sets above relative to the empirical distribution over R^d but using the *true* distribution over the conditional $p(y|\mathbf{x})$.

Fixing ϵ and γ we get, using uniform convergence, that the empirical probability of all $B \in \mathcal{S}_{\epsilon,\gamma}$ is bounded below by $\epsilon/2$ and the bias is bounded below by $\gamma/2$

etc etc