ON-LINE LEARNING IN PATTERN CLASSIFICATION USING ACTIVE SAMPLING

Jong-Min Park and Yu-Hen Hu

Department of Electrical and Computer Engineering
University of Wisconsin - Madison
1415 Engineering Drive
Madison, WI 53706-1691 USA
jong-min@ece.wisc.edu, hu@ece.wisc.edu

ABSTRACT

An adaptive on-line learning method is presented to facilitate pattern classification using active sampling to identify optimal decision boundary for a stochastic oracle with minimum number of training samples. The strategy of sampling at the *current estimate* of the decision boundary is shown to be optimal in the sense that the probability of convergence toward the true decision boundary at each step is maximized, offering theoretical justification on the popular strategy of category boundary sampling used by many query learning algorithms. Analysis of convergence in distribution is formulated using the Markov chain model.

1. INTRODUCTION

Pattern recognition via active sampling can trace its roots to statistical experiment design where performing an experiment (acquiring one training sample) may incur significant cost.

A number of active learning strategies, based on the concept of optimal experiment design, as well as importance sampling have been reported ([1, 2, 3, 4, 5, 6]). References [1] and [2] focused on active learning for pattern classification applications, with a common heuristic to sample at or near the present estimate of the category boundary using a justification that the function approximation of the posterior probability is most uncertain near the category boundary.

In this paper, we examine the validity of this argument using a two-class pattern classification problem as an example. We show that the variance of the approximation error reaches its maximum at the **true** category boundary.

Based on a stochastic oracle model, we show that the strategy of sampling at the **present estimate** of category boundary is optimal by using a perceptron-like learning algorithm. This result offers a direct theoretical justification of the "sample-at-current-boundary" strategy.

Convergence toward the true decision boundary is analyzed using the Markov chain model to prove the convergence in distribution.

2. PROBLEM FORMULATION

In a two-class pattern recognition problem, the feature vector $x \in \Re$ and the class label $C \in \{0,1\}$ are random variables with conditional probability density function $f_{x|C}(x \mid C = i) = f_i(x)$, and prior probability $P(C = i) = \pi_i$, where $i = \{0,1\}$. We also denote the posterior probability that $C = i$ given x is

$$q_i(x) = P\{C = i \mid x\} = \frac{\pi_i f_i(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}, \quad i = 0, 1.$$

Since $q_0(x) = 1 - q_1(x)$, for simplicity, we shall denote $q_1(x)$ by $q(x)$ in the rest of this paper.

The set of points $B = \{x \mid q_0(x) = q_1(x) = 1/2\}$ is called the decision boundary. In general, B may contain more than one point. In this work, we are mainly interested in applications where B contains exactly one point. This will be the case if, say, we are doing fine-tuning of a local decision boundary.

In an active learning (also known as *query learning*, [7]) problem formulation, the set of training samples are not given. Instead, a "learner" (the classification algorithm) will sample a feature vector x and present to an oracle (by performing an experiment or running a simulation) to learn the corresponding class label of x. In a two-class pattern recognition problem, this is equivalent to the evaluation of a function $y(x)$ at a specific value of x. The oracle will return $y(x) = 1$ or $y(x) = 0$ as the class label associated with x according to the posterior probability $P\{C = 1 \mid x\} = q(x)$ and $P\{C = 0 \mid x\} = 1 - q(x)$.

For $x \gg w^*$ (w^* is unknown to the learner) $q(x) \to 1$ and the oracle will most likely return $y(x) = 1$, while for $x \ll w^*$, it will most likely return $y(x) = 0$. For $x \approx w^*$, it is equally likely for the oracle to return $y(x) = 0$ or $y(x) = 1$.

3. MINIMUM ERROR ACTIVE LEARNING

To devise a learning rule that learns the optimal decision boundary w^* using active learning, let us define a 0–1 loss function $L(y(x), f(w,x)) = [y(x) - f(w,x)]^2 = [y(x) - u(x - w)]^2$, where $u(x) = 1$ if $x > 0$ and $u(x) = 0$ if $x < 0$. That is, if $f(w,x)$ and $y(x)$ have the same value for a given x, then the loss is 0. Otherwise, the loss is 1. Then, a cost

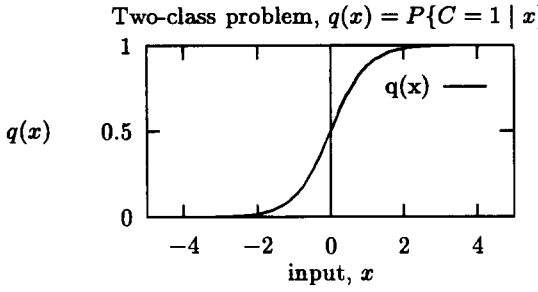

Figure 1: Two-class problem, $q(x) = P\{C = 1 \mid x\}$

function as the conditional risk given x can be defined as:

$$\begin{aligned} Cost(x) &= E[[y(x) - f(w,x)]^2 \mid x] \\ &= P\{y(x) = 1 \mid x\} \cdot [1 - u(x-w)]^2 \\ &\quad + P\{y(x) = 0 \mid x\} \cdot [0 - u(x-w)]^2 \\ &= q(x) \cdot [1 - u(x-w)] + (1 - q(x)) \cdot u(x-w) \\ &= \begin{cases} 1 - q(x) & x > w \\ q(x) & x < w \end{cases} \end{aligned} \quad (1)$$

Since $Cost(x)$ is not differentiable with respect to the decision boundary w, we use an adaptive formula similar to that of the classical perceptron learning algorithm:

$$w_{n+1} = w_n - [\epsilon(y(x_n) - 0.5)] \quad (2)$$

where ϵ is the learning rate, a.k.a. step size, and the new estimate of the boundary moves to the left or right by $\epsilon/2$ depending on the sample output $y(x_n)$ at the next sampling of x_n. Note that $E[y(w^*)] = 0.5$.

From (2),

$$|w_{n+1} - w^*| = |w_n - w^*| \pm \frac{\epsilon}{2}. \quad (3)$$

The algorithm will move toward convergence in the present step if $|w_{n+1} - w^*| = |w_n - w^*| - \frac{\epsilon}{2}$.

Let

$$P_e = P\{|w_{n+1} - w^*| = |w_n - w^*| + \frac{\epsilon}{2} \mid w_n\} \quad (4)$$

be the probability of error of moving away from the true boundary, in the one-step move described in (2).

Theorem 1 *The new sample x_n which minimizes the maximum possible value of P_e over each possible w_n is $x_n = w_n$. Moreover, if $x_n = w_n$, then*

$$P_e < 0.5 \quad (5)$$

Proof 1 *From (4),*

$$\begin{aligned} P_e &= P\{|w_{n+1} - w^*| = |w_n - w^*| + \frac{\epsilon}{2} \mid w_n\} \\ &= P\{y(x_n) = 0 \mid w_n > w^*\} \cdot P\{w_n > w^*\} \\ &\quad + P\{y(x_n) = 1 \mid w_n < w^*\} \cdot P\{w_n < w^*\} \end{aligned} \quad (6)$$

The event $y(x_n) = 0$ given x_n and the event $w_n > w^*$ are independent, hence

$$\begin{aligned} P\{y(x_n) = 0 \mid w_n > w^*\} &= P\{y(x_n) = 0\} \\ &= 1 - q(x_n) \end{aligned} \quad (7)$$

$$\begin{aligned} P\{y(x_n) = 1 \mid w_n < w^*\} &= P\{y(x_n) = 1\} \\ &= q(x_n) \end{aligned} \quad (8)$$

Thus (6) becomes

$$P_e = (1 - q(x_n)) \cdot P\{w_n > w^*\} + q(x_n) P\{w_n < w^*\} \quad (9)$$

If $w_n > w^*$, to minimize P_e, one would want to minimize the term $1 - q(x_n)$ by choosing x_n where $q(x_n)$ is as large as possible, i.e., where x_n is as large as possible. Conversely, if $w_n < w^*$, one would choose x_n such that x_n is as small as possible.

Since we have no knowledge on $P\{w_n > w^*\}$ or $P\{w_n < w^*\}$, we opt to use the min-max criterion to minimize the maximum probability value of P_e regardless of whether $w_n > w^*$ or $w_n < w^*$.

In particular, we note that when $w_n < w^*$, choosing $x_n < w_n$ will run into the risk of $x_n < w^*$, which implies $P_e = 1 - q(x_n) > 0.5$. Only for $x_n \geq w_n$ is it guaranteed that $P_e < 0.5$. Similarly, when $w_n < w^*$, only for $x_n \leq w_n$ does it guarantee that $P_e < 0.5$.

Taking the intersection of the two sets, $\{x_n \geq w_n\}$ and $\{x_n \leq w_n\}$, one concludes that $x_n = w_n$ is the only solution which guarantees that $P_e < 0.5$. Thus (5) is proved.

This theorem establishes that, with a min-max criteria, the optimal active learning strategy for the two-class pattern classification problem is to sample at the **current estimate** of the category boundary w_n. Thus (2) becomes

$$w_{n+1} = w_n - [\epsilon(y(w_n) - 0.5)] \quad (10)$$

4. CONVERGENCE ANALYSIS

In this section we show that the learning algorithm (10) converges in distribution toward the true boundary w^* using a Markov chain model.

Given an initial condition w_0, if ϵ is constant, then the set of random variables $\{w_n\}$ in (10) constitute a Markov chain,

$$w(k) = w_0 + k\frac{\epsilon}{2} \quad (11)$$

where k is any integer, and $w(k)$ denotes the boundary estimate w_n which falls in the state k of the Markov chain. We also define the state k^* to be the state closest to w^*.

Given $w_n = w(k)$, the output of the sampled value $y(w_n)$ dictates the state transition probability from state k to the next state k', $w_{n+1} = w(k')$. In particular,

$$P\{w_{n+1} = w(k') \mid w_n = w(k)\} = \begin{cases} P\{y(w_n) = 1 \mid w_n = w(k)\} = q(w(k)) \\ \qquad \text{if} \quad k' = k - 1, \\ P\{y(w_n) = 0 \mid w_n = w(k)\} = 1 - q(w(k)) \\ \qquad \text{if} \quad k' = k + 1, \\ 0 \\ \qquad \text{if} \quad |k' - k| \neq 1. \end{cases} \quad (12)$$

Since we are fine-tuning local boundary within a region as noted in section 2, the state space have bounds, thus it is considered to have finite number of states, $k_L \le k \le k_U$, where k_L and k_U are the lower and upper bounds, respectively.

Let $T(i \mid j)$ be the notation for the transition probability from state j to i, i.e.,

$$T(i \mid j) = P\{w_{n+1} = w(j) \mid w_n = w(i)\}.$$

Also define $T^{(n)}(i \mid j)$ as the transition probability of moving from state j to i in n steps. By induction, it can be shown that

$$T^{(m+n)}(i \mid j) = \sum_k T^{(m)}(k \mid j) T^{(n)}(i \mid k) \qquad (13)$$

Lemma 1 *Given a state i, the transition probability for the next state $j = i \pm 1$ satisfies*

$$T(j \mid i) > 0.5 \qquad |w(j) - w^*| < |w(i) - w^*| \qquad (14)$$

$$T(j \mid i) < 0.5 \qquad |w(j) - w^*| > |w(i) - w^*| \qquad (15)$$

Proof 2 *From (11), we note that*

$$w(i+1) > w(i) > w(i-1).$$

When $w(i) > w^*$, then $w(i+1) - w^* > w(i) - w^* > w(i-1) - w^* \ge 0$ and $|w(i+1) - w^*| > |w(i) - w^*| > |w(i-1) - w^*|$. From (12), and since $q(w(i)) > q(w^*) = 0.5$,

$$T(i+1 \mid i) = 1 - q(w(i)) < 1 - q(w^*) = 0.5,$$

then

$$T(i-1 \mid i) = 1 - T(i+1 \mid i) > 0.5$$

When $w(i) < w^*$, then $w(i-1) - w^* < w(i) - w^* < w(i+1) - w^* \le 0$, so $|w(i-1) - w^*| > |w(i) - w^*| > |w(i+1) - w^*|$. from (12), and since $q(w(i)) < q(w^*) = 0.5$,

$$T(i-1 \mid i) = q(w(i)) < q(w^*) = 0.5,$$

then

$$T(i+1 \mid i) = 1 - T(i-1 \mid i) > 0.5$$

This proves that the transition probability toward the true boundary is always greater than 0.5, and the probability away from the boundary is always less than 0.5.

Definition 1 *A set of states A is closed if*

$$T(A \mid k) = 1 \quad \forall k \in A$$

Definition 2 *A chain is indecomposable, if there is no two or more disjoint subset of states that are closed.*

Lemma 2 *For an indecomposable finite-state Markov chain with transition probabilities such that there is non-zero probability of reaching any state, then for any set of states A there is one solution $T(A)$ for all starting states k_0 that*

$$\lim_{n \to \infty} T^{(n)}(k \mid k_0) \to T(k) \quad \forall k \in A$$

This Markov chain is called regular or stable. Full description of Markov chain and its convergence proof may be referred in [8, 9].

Theorem 2 *The learning method (10) with constant ϵ and bounded region converges toward an asymptotic probability.*

Proof 3 *We only have to prove that (11) constitutes a regular Markov chain.*

A transition moving toward w^* is possible for all states. This can be shown by noting that one possible path from a state k to a state k^* defined to be closest to the true boundary w^* is to always move toward state k^* without moving away, and the probability is

$$\begin{cases} \prod_{i=k}^{k^*+1} T(i-1 \mid i) & \text{when } k > k^* \\ \prod_{i=k}^{k^*-1} T(i+1 \mid i) & \text{when } k < k^* \end{cases} \qquad (16)$$

which is > 0, from (14).

This chain is indecomposable. This is proven by noting that any state can reach w^*, which are then part of the subset of states which includes w^*. If there were to exist a state that is not an element of that subset, it can never reach w^*, contradicting the above statement.

Clearly the learning method above satisfies the criteria for a regular Markov chain, which proves its convergence in distribution toward the asymptotic probability.

Lemma 3 *If a Markov chain is regular, for any set of states A the proportion of time the system spends in A goes to the asymptotic probability $T(A)$.*

Let N_n be the number of times the system spends in state k up to time n, then using the central limit theorem, as $n \to \infty$, $P(|\frac{N_n}{n} - T(k)| < \epsilon) = 1$ for every arbitrary $\epsilon > 0$ for all k, which is called the weak law of large numbers ([8, 9]).

This shows an important corollary:

Corollary 1 *In n moves, as n becomes large, the state i is reached $nT(i)$ times, and the transition from i to j occurs $nT(j \mid i)T(i)$ times.*

Theorem 3 *Let $w_\infty = \lim_{n \to \infty} w_n$, then*

$$P\{w_\infty = w^*\} > P\{w_\infty = w'\} \quad \forall w' \ne w^* \qquad (17)$$

Proof 4 *We use the Markov chain model (11) and its transition probabilities $T(i \mid j)$.*

First, given two states i and $i+1$, in n steps, if there are m number of transitions from i to $i+1$, then the number of transitions from $i+1$ to i must differ from m by at most 1. This can be proved by looking at the transitions that cross between i and $i+1$. A second transition in the same direction can only occur if a matching transition in the other direction has already occurred.

From Corollary 1, the number of times spent in the transition between i and $i+1$ approaches

$$nT(i+1 \mid i)T(i) = nT(i \mid i+1)T(i+1) + \alpha,$$

where $\alpha \in \{-1, 0, 1\}$.

As $n \to \infty$, *term* α *drops out, and cancelling out* n *and rearranging, we get*

$$T(i) = \frac{T(i \mid i+1)}{T(i+1 \mid i)} T(i+1) \tag{18}$$

Since from (14) and (15), when $i > k^*$, $T(i \mid i+1) > 0.5$ *and* $T(i+1 \mid i) < 0.5$,

$$\frac{T(i \mid i+1)}{T(i+1 \mid i)} > 1,$$

thus

$$T(i) > T(i+1) \quad \forall i > k^*$$

and since $i > k^*$,

$$T(k^*) > T(i) \quad \forall i > k^*.$$

When $i < k^*$, *again from (14) and (15),* $T(i \mid i+1) < 0.5$ *and* $T(i+1 \mid i) > 0.5$,

$$\frac{T(i \mid i+1)}{T(i+1 \mid i)} < 1,$$

thus

$$T(i) < T(i+1) \quad \forall i < k^*$$

and since $i < k^*$,

$$T(i) < T(k^*) \quad \forall i < k^*.$$

Combining both, we have

$$T(k^*) > T(i) \quad \forall i \neq k^*.$$

Since $T(k^*)$ *is equivalent to* $P\{w_\infty = w^*\}$, *this proves (17).*

The above formulation thus proves the convergence in distribution of the "Sample-at-current-Boundary" learning algorithms toward the true boundary point w^*.

5. CONCLUSION

Active learning in a stochastic environment reflects the method of estimating the learning model given existing samples, then querying new samples that may optimize the estimation process, and iterate this process.

It is theoretically shown that sampling near the boundary is the optimal way for active learning in a stochastic environment.

Convergence analysis is done using the Markov chain model to prove that the method converges toward the true decision boundary in distribution.

6. REFERENCES

[1] D. A. Cohn, "Neural network exploration using optimal experiment design," in *Advances in Neural Information Processing Systems*, vol. 6, 1994.

[2] J.-N. Hwang, J. J. Choi, S. Oh, and R. J. M. II, "Query-based learning applied to partially trained multilayer perceptrons," *IEEE Transactions on Neural Networks*, vol. 2, pp. 131–136, Jan. 1991.

[3] P. Sollich, "Learning from minimum entropy queries in a large committee machine," *Physical Review E:LH 7P56 R329 Physics Library*, vol. 53, pp. R2060–R2063, 1996.

[4] Y. Kabashima and S. Shinomoto, "Incremental learning with and without queries in binary choice problems," in *Proc. International Joint Conference on Neural Networks*, vol. 2, pp. 1637–1640, 1993.

[5] D. MacKay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, no. 4, pp. 590–604, 1992.

[6] M. Plutowski and H. White, "Active selection of training examples for network learning in noiseless environments," Tech. Rep. CS91-180, Univ. of California-San Diego, Feb. 1990.

[7] D. Angluin, "Queries and concept learning," *Machine Learning*, vol. 2, pp. 319–342, 1988.

[8] W. Feller, *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons, Inc., New York, 1957.

[9] G. Grimmett and D. Stirzaker, *Probability and Random Processes*. Clarendon Press, 1982.