Online Local Linear Classification

Joseph Wang, Kirill Trapeznikov, and Venkatesh Saligrama
Department of Electrical & Computer Engineering
Boston University
Boston, MA 02215

***Abstract*—We present a novel convex formulation to learning binary, 2-region local linear classifiers. From this convex formulation, we formulate an online optimization scheme using stochastic gradient descent that allows for efficient training using streaming training data. We demonstrate the fast convergence and accurate classification on the canonical XOR dataset.**

I. Introduction

We present a convex approach to learning local decision boundaries by partitioning the feature space into two regions and learning independent local classifiers in each region, as shown in Fig. 1. Under this structure, the binary decision function g maps each observation to a region where the associated local classifier f_1 or f_2 makes a class prediction. We use an empirical risk minimization approach to jointly train both the decision function, g, and the local classification functions, f_1 and f_2. We present a globally convex formulation, allowing for globally optimal solutions to be efficiently found.

Given that data is often structured such that the optimal decision boundary is locally simple, complex classification functions are generally unnecessary to reduce empirical error [1]. We therefore focus on linear partitioning and classification functions, although the methods proposed are easily extended to kernel-based classifiers. Linear partitioning and classification functions allow for complex decision boundaries to be approximated by piecewise linear functions.

Local linear classifiers of the form shown in Fig. 1 are easily applied to large data sets due to the extremely small test-time run cost. In contrast, non-linear kernel methods or boosting-based methods with high-numbers of weak learners can be computationally costly during test-time, limiting their practicality in high-throughput classification scenarios.

A. Related Work

The architecture of our classifiers is closely related to decision trees [2]. However, decision trees typically are trained in a greedy, top-down fashion in an attempt to minimize some loss or a heuristic, such as region purity or entropy, at each split/partition of the feature space. In contrast, our approach directly minimizes a surrogate on the global empirical risk. Furthermore, the heuristics employed by decision trees are often difficult to optimize, limiting each decision to single feature splits which can be optimized by brute force search. The approach we propose allows for any kernel-based decision boundary for both the partitioning and classification functions.

Approximating decision boundaries with piecewise simple functions has also been proposed in generative learning schemes, such as Mixture Discriminant Analysis (MDA), proposed by Hastie *et al.* [3], where each class is modeled as by mixture of Gaussian distributions. Local Linear Discriminant Analysis (LLDA), proposed by Kim *et al.* [4], clusters data and learns decision boundaries independently within each cluster. Additional piecewise linear techniques have been proposed in the past [5], [6], [7], however these approaches do not learn decision boundaries based on minimizing global empirical risk.

Minimizing the empirical error of local classifiers has been proposed in the mixture of experts framework [8], bilinear separation framework [9], and space partitioning classification framework [1]. The mixture of experts framework hybridizes generative and discriminative approaches by replacing the partitioning classifier, G, with a "latent" probability distribution. Alternating minimization is used, switching between learning the parameters of the "latent" distribution and training local classifiers using standard learning methods. In the bilinear separation problem, an attempt is made to directly minimize empirical risk. The empirical loss is expressed as a product of indicators which are approximated with hinge loss surrogate functions. This introduces a bilinear optimization problem whose globally optimal solution cannot be efficiently found. The space partitioning classifier framework can be seen as a generalization of the bilinear separation problem, where alternating minimization is used to train the partitioning and local classifiers, with each subproblem posed as a standard supervised learning problem. While space partitioning classifiers have been shown to have strong performance, convergence can only be guaranteed to a local minima without resorting to exhaustive search.

Online learning is a well studied problem [10]

978-1-4673-3146-3/13/$31.00 ©2013IEEE

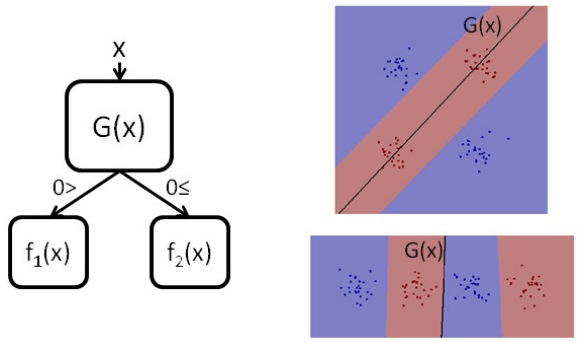

Fig. 1. **Left: A local linear classifier with 2 regions. The binary reject classifiers $g(x)$ partitions the space, and the region classifiers $f_1(x)$ and $f_2(x)$ output labels in each regions. Right: Decision boundaries learned for two synthetic examples.**

in which classifiers are trained over extremely large training data sets. In online learning, the goal is to train classifiers using streaming training data and labels, with the classifier updated after every observation. A well studied approach to optimizing convex loss functions is online gradient descent, which has been shown to be an efficient and effective approach to training classifiers [11], [12].

II. Convex Local Learning

A. Empirical Risk Reformulation

Consider the binary local classifier as shown in Fig. 1. The function given by this structure is:

$$F(x) = \mathbb{1}_{g(x)\leq 0} f_1(x) + \mathbb{1}_{g(x)>0} f_2(x).$$

The function $g(x)$ partitions the data into two regions. Dependent on the output of the function $g(\cdot)$, a binary label y is estimated using either the classifier $f_1(x)$ or $f_2(x)$. Our goal is to jointly learn the partitioning function g and local functions f_1 and f_2.

For a classifier of this form, the event of an error (empirical risk) can be expressed:

$$R(g, f_1, f_2) = \sum_{i=1}^{n} \mathbb{1}_{g(x)\leq 0} \mathbb{1}_{f_1(x)\neq y_i} + \mathbb{1}_{g(x)>0} \mathbb{1}_{f_2(x)\neq y_i}. \tag{1}$$

Replacing indicators with surrogate functions generally yields a non-convex upper bound on the empirical risk. Instead, we reformulate the empirical risk to yield a convex upper-bounding surrogate:

Theorem II.1. *The empirical risk* (1) *can equivalently be expressed:*

$$R(g, f_1, f_2) = \sum_{i=1}^{n} \left[\max\left(\mathbb{1}_{g(x_i)>0} + \mathbb{1}_{f_2(x_i)\neq y_i}, \mathbb{1}_{f_1(x_i)\neq y_i} + \mathbb{1}_{g(x_i)\leq 0} \right) - 1 \right] \tag{2}$$

Proof: Consider the empirical risk with respect to the event of a correct classification:

$$R(g, f_1, f_2) = \sum_{i=1}^{n} \left[1 - \left[\mathbb{1}_{g(x_i)\leq 0} \mathbb{1}_{f_1(x_i)=y_i} + \mathbb{1}_{g(x_i)>0} \mathbb{1}_{f_2(x_i)=y_i} \right] \right]. \tag{3}$$

Equivalently, the product of indicators can be replaced with the minimization:

$$R(g, f_1, f_2) = \sum_{i=1}^{n} \left[1 - \min_{\lambda_i^1\in[0,1], \lambda_i^2\in[0,1]} \left[\lambda_i^1 \mathbb{1}_{g(x_i)\leq 0} + (1-\lambda_i^1)\mathbb{1}_{f_1(x_i)=y_i} + \lambda_i^2 \mathbb{1}_{g(x_i)>0} + (1-\lambda_i^2)\mathbb{1}_{f_2(x_i)=y_i} \right] \right]. \tag{4}$$

The minimization with respect to λ_i^1 and λ_i^2 is not guaranteed to have a unique solution. Due to the fact that $\mathbb{1}_{g(x)\leq 0} = 1 - \mathbb{1}_{g(x)>0}$, one valid optimal solution occurs when $\lambda_i^1 = 1 - \lambda_i^2$. Enforcing this constraint allows for the substitutions $\lambda_i^1 = \lambda_i$ and $\lambda_i^2 = 1 - \lambda_i$, simplifying the empirical risk:

$$R(g, f_1, f_2) = \sum_{i=1}^{n} \max_{\lambda_i\in[0,1]} \left[1 - \left[\lambda_i \mathbb{1}_{g(x_i)\leq 0} + (1-\lambda_i)\mathbb{1}_{f_1(x_i)=y_i} + (1-\lambda_i)\mathbb{1}_{g(x_i)>0} + \lambda_i \mathbb{1}_{f_2(x_i)=y_i} \right] \right]. \tag{5}$$

Changing the sign in the arguments of the indicator functions ($\mathbb{1}_{z<0} = 1 - \mathbb{1}_{z\geq 0}$), the empirical risk can be expressed:

$$R(g, f_1, f_2) = \sum_{i=1}^{n} \max_{\lambda_i\in[0,1]} \left[\lambda_i \mathbb{1}_{g(x_i)>0} + (1-\lambda_i)\mathbb{1}_{f_1(x_i)\neq y_i} + (1-\lambda_i)\mathbb{1}_{g(x_i)\leq 0} + \lambda_i \mathbb{1}_{f_2(x_i)\neq y_i} - 1 \right]. \tag{6}$$

Note that the optimal solution to the variables λ_i is at the boundaries of the constraints, that is $\lambda_i \in \{0, 1\}$. As such, the variable λ_i can be replaced with a maximization, allowing the empirical risk to be expressed:

$$R(g, f_1, f_2) = \sum_{i=1}^{n} \left[\max \left(\mathbb{1}_{g(x_i)>0} + \mathbb{1}_{f_2(x_i)\neq y_i}, \mathbb{1}_{f_1(x_i)\neq y_i} + \mathbb{1}_{g(x_i)\leq 0} \right) - 1 \right].$$

■

Introducing convex surrogate functions for the indicator into the empirical risk as formulated in (2) leads to a convex, upper-bounding loss function. The resulting risk is globally convex and allows for efficient optimization, whereas existing approaches [9], [1] are generally non-convex, preventing efficient global optimization.

B. Convex Relaxations and Uniqueness

Although a globally optimal solution of this convex relaxation can be efficiently found, directly replacing indicators with convex surrogates generally leads to poor classification performance. This is a result of the structure of classifier, which allows for the sign of the partitioning function, $g(x)$, to be flipped and the local classifiers to be exchanged, yielding the same predictions and empirical risk.

Theorem II.2. *For any convex relaxation $\hat{R}(g, f_1, f_2)$, where the indicators are each replaced with upper-bounding surrogate functions of the same form, the set of global optimal solutions includes the partitioning function $g(x) = 0$.*

Proof: Consider set of functions g^*, f_1^*, f_2^* that minimize $\hat{R}(g, f_1, f_2)$. An alternative set of functions $\tilde{g}^* = -g^*, \tilde{f}_1^* = f_2^*, \tilde{f}_2^* = f_1^*$ also minimizes $\hat{R}(g, f_1, f_2)$. Given that the surrogate loss function is convex, any convex combination of these two functions (including the partitioning function $g(x) = 0$) has a value less than or equal to $\hat{R}(g^*, f_1^*, f_2^*)$. ■

The exchangeability of solutions ($\hat{R}(g, f_1, f_2) = \hat{R}(-g, f_2, f_1)$) is a fundamental limitation when constructing convex relaxations of the empirical risk. This fundamental limitation arises in estimating an unobserved exchangeable variable, as previously shown when using convex relaxations to fit latent variable models [13].

To overcome this fundamental limitation, we modify the feasible set of solutions so that it does not include both(g, f_1, f_2) and $(-g, f_2, f_1)$. In particular, we randomly select an observation, x_j, and enforce the constraint $g(x_j) \geq \beta$ for some constant $\beta \geq 1$. Intuitively, this constraint fixes the region that the point x_j is assigned to, removing the exchangeability as $g(x_j) < 0$ is not a feasible solution.

Algorithm 1 Online Update

Input: Observation and label, x_t, y_t, current partitioning classifier, α, and local classifiers β_1, β_2
Output: Updated partitioning classifier, α, updated local classifiers β_1, β_2

1. Find active region

$$r_t = \begin{cases} 1 & \text{if } \log(1 + e^{\alpha^T x_t}) + \log(1 + e^{-y_t \beta_1^T x_t}) > \\ & \quad \log(1 + e^{-\alpha^T x_t}) + \log(1 + e^{-y_t \beta_2^T x_t}) \\ 2 & \text{otherwise} \end{cases}$$

2. Calculate the subgradient for the partitioning classification functions

$$\bigtriangledown \alpha = \begin{cases} \frac{-x_t}{1+e^{-\alpha^T x_t}} & \text{if } r_t = 1 \\ \frac{x_t}{1+e^{\alpha^T x_t}} & \text{if } r_t = 2 \end{cases}$$

$$\bigtriangledown \beta_1 = \begin{cases} \frac{-y_t x_t}{1+e^{y_t \beta_1^T x_t}} & \text{if } r_t = 1 \\ 0 & \text{if } r_t = 2 \end{cases}$$

$$\bigtriangledown \beta_2 = \begin{cases} 0 & \text{if } r_t = 1 \\ \frac{-y_t x_t}{1+e^{y_t \beta_2^T x_t}} & \text{if } r_t = 2 \end{cases}$$

3. Return updated functions

$$\alpha = \alpha - \frac{\bigtriangledown \alpha}{\sqrt{t}}$$
$$\beta_1 = \beta_1 - \frac{\bigtriangledown \beta_1}{\sqrt{t}}$$
$$\beta_2 = \beta_2 - \frac{\bigtriangledown \beta_2}{\sqrt{t}}$$

III. Online Training of Locally Linear Classifiers

In practice, we upper-bound the indicator losses in (2) using logistic loss functions. The logistic loss function is an ideal choice when training local linear classifiers using streaming data, as it is smooth continuously differentiable while asymptotically approximating the tightest convex surrogate functions (hinge losses as shown in [14]). Starting with a random set of functions, we use a stochastic gradient descent algorithm shown in Alg. 1 to find the local linear classifier that minimizes the objective function [11].

Performance of this online algorithm is shown on a

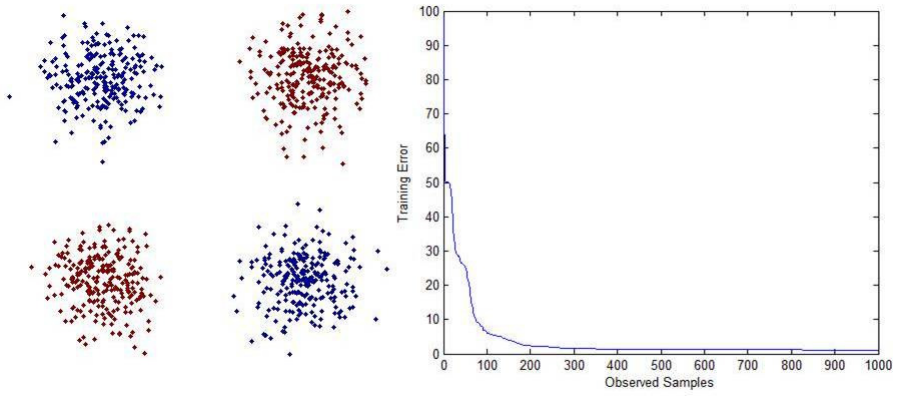

Fig. 2. **Left: Synthetic gaussian XOR data. Right: Average training error over the entire training set vs. observed training observations.**

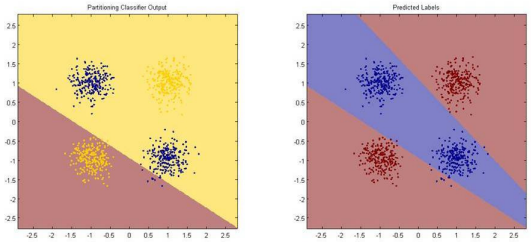

Fig. 3. **Left: Partitioned regions learned via online training. Right: Decision boundaries learned by online training.**

synthetic dataset in Fig. 2. The synthetic dataset, shown in Fig. 2, was generated from a mixture of Gaussians, with a single gaussian distribution centered in each quadrant and labels corresponding to each Gaussian equal to the XOR of the mean coordinates. A randomly initialized local linear classifier is updated by randomly generated training examples. The average training error on the entire training dataset is shown on the right of Fig. 2. On the Gaussian XOR data set, the local linear classifier converges at an extremely fast rate, with convergence approximately after 200 updates.

REFERENCES

[1] Joseph Wang and Venkatesh Saligrama. Local supervised learning through space partitioning. In *Advances in Neural Information Processing Systems 25*. 2012.

[2] R. A. Olshen L. Breiman, J. H. Friedman and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[3] Trevor Hastie and Robert Tibshirani. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society, Series B*, 58:155–176, 1996.

[4] Tae-Kyun Kim and Josef Kittler. Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:318–327, 2005.

[5] Ofer Dekel and Ohad Shamir. There's a hole in my data space: Piecewise predictors for heterogeneous learning problems. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*.

[6] Juan Dai, Shuicheng Yan, Xiaoou Tang, and James T. Kwok. Locally adaptive classification piloted by uncertainty. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 225–232, New York, NY, USA, 2006. ACM.

[7] Marc Toussaint and Sethu Vijayakumar. Learning discontinuities with products-of-sigmoids for switching between local models. In *Proceedings of the 22nd international conference on Machine Learning*, pages 904–911. ACM Press, 2005.

[8] Clodoaldo A.M. Lima, Andre L.V. Coelho, and Fernando J. Von Zuben. Hybridizing mixtures of experts with support vector machines: Investigation into nonlinear dynamic systems identification. *Information Sciences*, 177(10):2049 – 2074, 2007.

[9] Kristin P. Bennett and O. L. Mangasarian. Bilinear separation of two sets in n-space. *Computational Optimization and Applications*, 2, 1993.

[10] Shai Shalev-Shwartz. Online learning and online convex optimization. *Found. Trends Mach. Learn.*, 4(2):107–194, February 2012.

[11] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.

[12] Elad Hazan, Adam Kalai, Satyen Kale, and Amit Agarwal. Logarithmic regret algorithms for online convex optimization. In *In 19th COLT*, pages 499–513, 2006.

[13] Yuhong Guo and Dale Schuurmans. Convex relaxations of latent variable training. *Advances in Neural Information Processing Systems*, 20:601–608, 2008.

[14] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines. *Journal of the American Statistical Association*, 99(465):67–81, 2004.