Adversarial Multiclass Classification: A Risk Minimization Perspective

Rizal Fathony, Anqi Liu, Kaiser Asif, and Brian D. Ziebart
{rfatho2, aliu33, kasif2, bziebart}@uic.edu
Department of Computer Science, University of Illinois at Chicago

UIC COMPUTER SCIENCE

Overview

Empirical Risk Minimization (ERM)

- Goal of classification: minimize classification error - zero-one loss
- Minimize zero-one loss over training data: NP-hard in general

Support Vector Machine (SVM)

- Optimize hinge loss (a convex surrogate loss) over training data
- Binary SVM: Fisher consistent and universally consistent
- Generalizing SVM to multiclass case is challenging: loses consistency guarantees or does not perform well in practice

Adversarial Classification

- Optimizes **exact loss** (zero-one) and **approximates training data**
- Promising empirical results for cost-sensitive and multivariate losses (Asif et al. 2015, Wang et al. 2015)

Our Approach:

1. Recast zero-one adversarial classification from ERM perspective by analyzing the Nash equilibrium and define a new multiclass loss
2. Fill the long-standing gap in ERM methods by simultaneously:
   (1) Guaranteeing Fisher consistency and universal consistency
   (2) Enabling computational efficiency via kernel trick and dual sparsity
   (3) Providing competitive performance in practice
3. Significantly improve computational efficiency
   $\rightarrow$ no game solving using linear programming is required

Related Work

Multiclass Support Vector Machine

- Three main formulations:
  1) **WW** by Weston and Watkins (1999):
     $\text{loss}_{\text{WW}}(\mathbf{x}_i, y_i) = \sum_{j \neq y_i} [1 - (f_{y_i}(\mathbf{x}_i) - f_j(\mathbf{x}_i))]_+$
  2) **CS** by Crammer and Singer (2002):
     $\text{loss}_{\text{CS}}(\mathbf{x}_i, y_i) = \max_{j \neq y_i} [1 - (f_{y_i}(\mathbf{x}_i) - f_j(\mathbf{x}_i))]_+$
  3) **LLW** by Lee, Lin and Wahba (2004):
     $\text{loss}_{\text{LLW}}(\mathbf{x}_i, y_i) = \sum_{j \neq y_i} [1 + f_j(\mathbf{x}_i)]_+$, subject to $\sum_j f_j(\mathbf{x}_i) = 0$
- WW and CS: relative margin, LLW: absolute margin
- Only LLW is Fisher consistent and universally consistent. (Tewari and Bartlett 2004, Liu 2007)
- LLW's use of absolute margin $\rightarrow$ often performs poorly in datasets with low dimensional features. (Doğan et al. 2016)

Adversarial Prediction Games (Asif et al. 2015)

- Two player zero-sum games:
  1) Adversarial player: controls conditional label distribution $\check{P}(\check{y}|\mathbf{x})$
     $\rightarrow$ must approximate training data, but otherwise maximize expected loss
  2) Estimator player: controls $\hat{P}(\hat{y}|\mathbf{x})$ and seeks to minimize expected loss
  - Formulation:
$$\min_{\hat{P}} \max_{\check{P}: \mathbb{E}_{P(\mathbf{x})\check{P}(\check{y}|\mathbf{x})}[\phi(\mathbf{X}, \check{Y})] = \tilde{\phi}} \mathbb{E}_{\tilde{P}(\mathbf{x})\hat{P}(\hat{y}|\mathbf{x})\check{P}(\check{y}|\mathbf{x})} \left[\text{loss}(\hat{Y}, \check{Y}) \right]$$
  - Feature moments $\tilde{\phi} = \mathbb{E}_{\tilde{P}(\mathbf{x},y)}[\phi(\mathbf{X}, Y)]$, are measured from training data
- For zero-one loss, it reduces to an optimization convex in θ:
$$\min_{\theta} \sum_i \max_{\check{\mathbf{p}}} \min_{\hat{\mathbf{p}}} \hat{\mathbf{p}}_{\mathbf{x_i}}^T \mathbf{L}'_{\mathbf{x}_i,\theta} \check{\mathbf{p}}_{\mathbf{x_i}}; \quad \mathbf{L}'_{\mathbf{x}_i,\theta} = \begin{bmatrix} \psi_{1,y_i}(\mathbf{x}_i) & \cdots & \psi_{|\mathcal{Y}|,y_i}(\mathbf{x}_i) + 1 \\ \vdots & \ddots & \vdots \\ \psi_{1,y_i}(\mathbf{x}_i) + 1 & \cdots & \psi_{|\mathcal{Y}|,y_i}(\mathbf{x}_i) \end{bmatrix}$$
- Inner zero-sum game can be solved using a linear program

Risk Minimization Perspective

Theorem *The model parameters θ for multiclass zero-one adversarial classification are equivalently obtained from empirical risk minimization under the adversarial zero-one loss function:*

$$AL^{0\text{-}1}_{\mathbf{f}}(\mathbf{x}_i, y_i) = \max_{\mathcal{S} \subseteq \{1,\dots,|\mathcal{Y}|\}, \mathcal{S} \neq \emptyset} \frac{\sum_{j \in \mathcal{S}} \psi_{j,y_i}(\mathbf{x}_i) + |\mathcal{S}| - 1}{|\mathcal{S}|}, \text{ where}$$

$\mathcal{S}$ is any non-empty member of the powerset of classes $\{1, 2, \dots, |\mathcal{Y}|\}$

Plots of AL$^{0\text{-}1}$ in binary and 3-class classification

- AL$^{0\text{-}1}$ is the maximum value over $2^{|\mathcal{Y}|} - 1$ linear hyperplanes
- Binary classification: similar with hinge loss, but with two hinges at -1 and 1 (as shown in figure on the right)
- Three class classification: the loss function has seven facets.
- Comparison with WW and CS surrogates (as shown below)

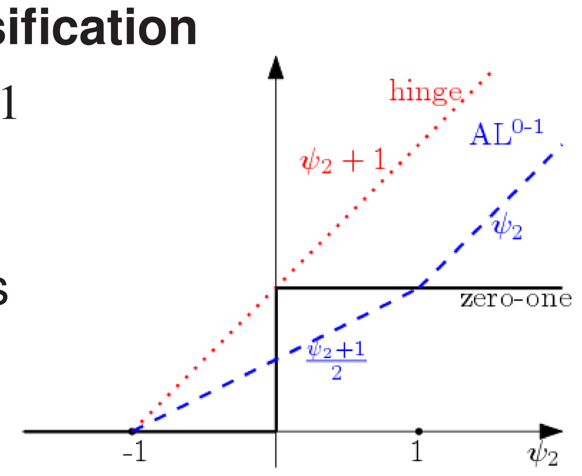

Figure: Binary AL$^{0\text{-}1}$ over the space of potential differences $\psi_{j,y}(\mathbf{x})$ when $y = 1$

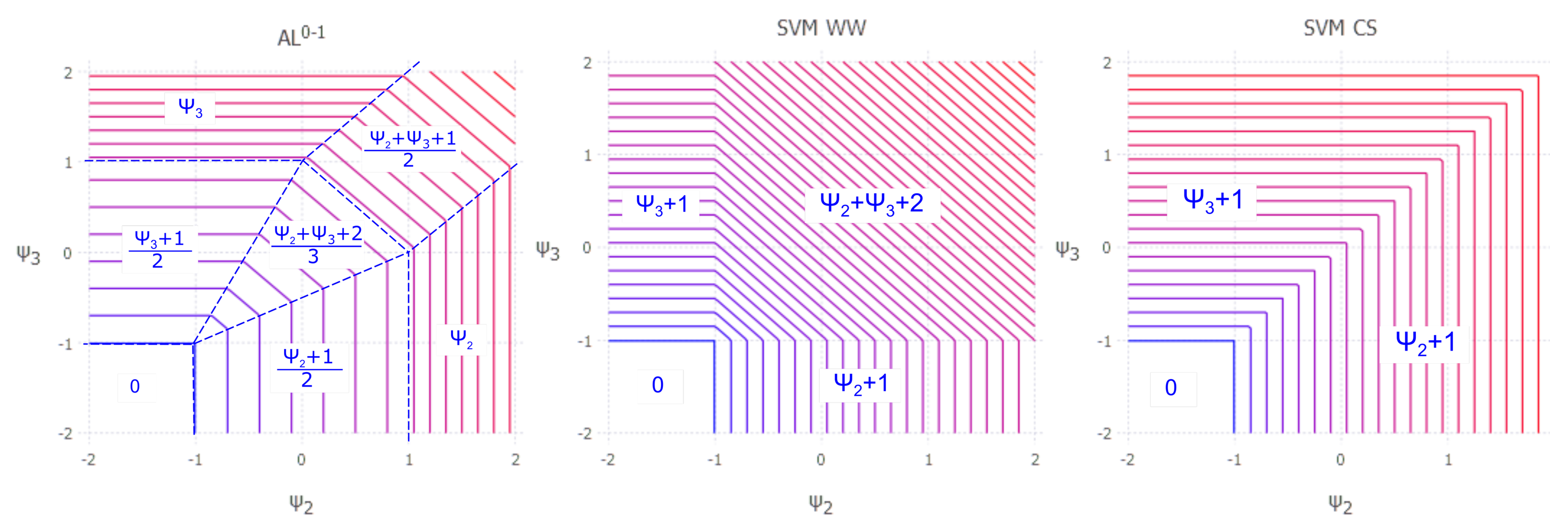

Figure: Loss function contour plots over the space of potential differences for the prediction task with three classes when $y = 1$ under AL$^{0\text{-}1}$, WW and CS.

Statistical Consistency

Fisher Consistency

- Minimizing a Fisher consistent loss will yield the Bayes optimal decision boundary given the true distribution, $P(x, y)$
- Multiclass: it requires $\text{argmax}_j f_j^*(\mathbf{x}) \subseteq \text{argmax}_j P_j(\mathbf{x})$, where $P_j(\mathbf{x}) \triangleq P(Y = j|\mathbf{x})$ and $\mathbf{f}^*(\mathbf{x})$ is the minimizer of $\mathbb{E}\left[\text{loss}_{\mathbf{f}}(\mathbf{X}, Y)|\mathbf{X} = \mathbf{x}\right]$
- The minimizer of $\mathbb{E}\left[\text{AL}^{0\text{-}1}_{\mathbf{f}}(\mathbf{X}, Y)|\mathbf{X} = \mathbf{x}\right]$ resides on the hyperplane defined by the complete set of labels, $\mathcal{S} = \{1, \dots, |\mathcal{Y}|\}$
- It is equivalent with the area where $-\frac{1}{|\mathcal{Y}|} \leq f_j(\mathbf{x}) \leq \frac{|\mathcal{Y}|-1}{|\mathcal{Y}|}, \forall j \in \{1, \dots, |\mathcal{Y}|\}$, s.t. $\sum_j f_j(\mathbf{x}) = 0$
- The minimization reduces to:
$$\max_{\mathbf{f}} \sum_{y=1}^{|\mathcal{Y}|} P_y(\mathbf{x}) f_y(\mathbf{x}) \text{ s.t.: } -\frac{1}{|\mathcal{Y}|} \leq f_j(\mathbf{x}) \leq \frac{|\mathcal{Y}| - 1}{|\mathcal{Y}|} \quad j \in \{1, \dots, |\mathcal{Y}|\}; \quad \sum_{j=1}^{|\mathcal{Y}|} f_j(\mathbf{x}) = 0$$
- The solution satisfies the requirement of Fisher consistency

Universal Consistency

- AL$^{0\text{-}1}$ is a Lipschitz loss with constant 1, optimizing it with universal kernel and reasonablly small regularization on any distribution yields Bayes optimal classifier. (Steinwart et al. 2008)

Optimization

Primal Optimization using Stochastic Sub-gradient Descent

- The sub-gradient of AL$^{0\text{-}1}$ includes the mean of feature differences:
$$\frac{1}{|R|} \sum_{j \in R} [\phi(\mathbf{x}_i, j) - \phi(\mathbf{x}_i, y_i)]$$
- The set R can be computed optimally using a greedy algorithm

Dual Optimization using Quadratic Programming (QP)

- Constrained QP of AL$^{0\text{-}1}$ plus L2 regularization
$$\min_{\theta} \frac{1}{2}\|\theta\|^2 + C \sum_{i=1}^{n} \xi_i \quad \text{s.t.:} \quad \xi_i \geq \Delta_{i,k} \quad \forall i \in \{1, \dots n\}, k \in \{1, \dots, 2^{|\mathcal{Y}|} - 1\}$$
  where $2^{|\mathcal{Y}|} - 1$ possible constraints for example i are denoted as $\Delta_{i,k}$
- Dual QP formulation
$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \sum_{k=1}^{2^{|\mathcal{Y}|}-1} \nu_{i,k} \alpha_{i,k} - \frac{1}{2} \sum_{i,j=1}^{m} \sum_{k,l=1}^{2^{|\mathcal{Y}|}-1} \alpha_{i,k} \alpha_{j,l} [\Lambda_{i,k} \cdot \Lambda_{j,l}]$$
$$\text{s.t.:} \quad \alpha_{i,k} \geq 0, \ \sum_{k=1}^{2^{|\mathcal{Y}|}-1} \alpha_{i,k} = C, \ i \in \{1, \dots, n\}, \ k \in \{1, \dots, 2^{|\mathcal{Y}|} - 1\}$$
  where $\Lambda_{i,k} \triangleq \frac{d\Delta_{i,k}}{d\theta}$, and $\nu_{i,k}$ is the constant part of $\Delta_{i,k}$

Kernel Trick

- Dual QP only depends on dot products
- Enables efficient rich feature expansion

Constraint Generation

- The number of constraints in QP is exponential
- Constraint generation method $\rightarrow$ efficiently solve the problem
- Polynomial time convergence guarantee is provided

Experiments and Results

Table: The mean and standard deviation of the accuracy for each model with linear and Gaussian kernel. Bold numbers indicate that the result is the best or not significantly worse than the best.

Dataset	Linear Kernel				Gaussian Kernel			
	AL$^{0\text{-}1}$	WW	CS	LLW	AL$^{0\text{-}1}$	WW	CS	LLW
iris	**96.3** (3.1)	**96.0** (2.6)	**96.3** (2.4)	79.7 (5.5)	**96.7** (2.4)	**96.4** (2.4)	**96.2** (2.3)	95.4 (2.1)
glass	**62.5** (6.0)	**62.2** (3.6)	**62.5** (3.9)	52.8 (4.6)	**69.5** (4.2)	66.8 (4.3)	**69.4** (4.8)	**69.2** (4.4)
redwine	**58.8** (2.0)	**59.1** (1.9)	56.6 (2.0)	57.7 (1.7)	63.3 (1.8)	64.2 (2.0)	64.2 (1.9)	**64.7** (2.1)
ecoli	**86.2** (2.2)	85.7 (2.5)	**85.8** (2.3)	74.1 (3.3)	**86.0** (2.7)	84.9 (2.4)	**85.6** (2.4)	**86.0** (2.5)
vehicle	**78.8** (2.2)	**78.8** (1.7)	**78.4** (2.3)	69.8 (3.7)	**84.3** (2.5)	**84.4** (2.6)	83.8 (2.3)	**84.4** (2.6)
segment	94.9 (0.7)	94.9 (0.8)	**95.2** (0.8)	75.8 (1.5)	**96.5** (0.6)	**96.6** (0.5)	96.3 (0.6)	96.4 (0.5)
sat	84.9 (0.7)	**85.4** (0.7)	84.7 (0.7)	74.9 (0.9)	**91.9** (0.5)	**92.0** (0.6)	**91.9** (0.5)	**91.9** (0.4)
optdigits	**96.6** (0.6)	96.5 (0.7)	96.3 (0.6)	76.2 (2.2)	98.7 (0.4)	98.8 (0.4)	98.8 (0.3)	**98.9** (0.3)
pageblocks	96.0 (0.5)	96.1 (0.5)	**96.3** (0.5)	92.5 (0.8)	**96.8** (0.5)	96.6 (0.4)	96.7 (0.4)	96.6 (0.4)
libras	**74.1** (3.3)	72.0 (3.8)	71.3 (4.3)	34.0 (6.4)	83.6 (3.8)	83.8 (3.4)	**85.0** (3.9)	83.2 (4.2)
vertebral	**85.5** (2.9)	**85.9** (2.7)	**85.4** (3.3)	79.8 (5.6)	**86.0** (3.1)	**85.3** (2.9)	85.5 (3.3)	84.4 (2.7)
breasttissue	**64.4** (7.1)	59.7 (7.8)	**66.3** (6.9)	58.3 (8.1)	**68.4** (8.6)	**68.1** (6.5)	**66.6** (8.9)	**68.0** (7.2)
avg	81.59	81.02	81.25	68.80	85.14	84.82	85.00	84.93
#bold	9	6	8	0	9	6	6	7

- Experiment using Linear Kernel
  - LLW performs poorly in all datasets
  - AL$^{0\text{-}1}$ has a slight advantage on average accuracy and number of "best"
- Experiment using Gaussian Kernel
  - Provides access to much richer feature spaces
  - Increases performance of all models, especially the LLW
  - AL$^{0\text{-}1}$ maintains a slight advantage

Acknowledgments: This research was supported as part of the Future of Life Institute (futureoflife.org) FLI-RFP-AI1 program, grant#2016-158710 and by NSF grant RI-#1526379.