LEGENDRETRON: Uprising Proper Multiclass Loss Learning

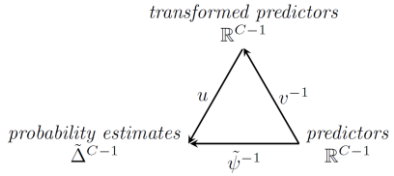
Kevin H. Lam[1], Christian Walder[23], Spiridon Penev[14], Richard Nock[23]
[1]School of Mathematics & Statistics, University of New South Wales, [2]Google Research, [3]ANU College of Engineering, Computing and Cybernetics, The Australian National University, [4]UNSW Data Science Hub, University of New South Wales

Problem Statement

Loss functions underpin supervised learning and are often chosen prior to model development. Existing methods for ***learning proper losses*** illustrate the benefits of transferring learned losses to related domains, but they focus only on the binary setting and ***cannot guarantee properness in the multiclass setting***.

Multiclass probability estimation is conventionally done by reducing the original problem to component 1-vs-rest and 1-vs-1 binary problems. Probability estimates from these component models are ***not guaranteed to be admissible (optimal).*** We desire an approach that can ***jointly learn proper multiclass losses and multiclass probabilities***.

Learning Proper Losses through Canonical Links



- Canonical links bridge predictors, probability estimates, and a strictly proper multiclass loss $\ell = -((\tilde{\psi} \circ \Pi) \cdot J_\Pi)$

$$\tilde{\psi}: \tilde{\Delta}^{C-1} \to \mathbb{R}^{C-1}, \tilde{\psi}(\tilde{p}) = -\nabla \underline{\tilde{L}}(\tilde{p})$$

- Canonical links are invertible and flexibly parameterizable with ***Legendre functions*** and the (u, v)***-geometric structure***

$$\tilde{\psi}^{-1} = u \circ v^{-1}$$

- Solution: set u as invertible squashing function and v^{-1} as invertible net with symmetric & positive definite Hessians!

Contributions

1. Derived necessary and sufficient conditions for a composite function in $\mathbb{R}^{C-1}$ to be monotonic and the gradient of a twice-differentiable convex function
2. Derived sufficient conditions for a composite function in $\mathbb{R}^{C-1}$ to be monotonic and the gradient of a twice-differentiable strictly convex function
3. Proposed LEGENDRETRON as a novel and practical way of learning proper canonical losses and probabilities concurrently in the multiclass problem setting

Theorems

Theorem 4.2. *Let $f : \mathbb{R}^{C-1} \to \mathbb{R}^{C-1}$ and $g : \mathbb{R}^{C-1} \to \mathbb{R}^{C-1}$ be differentiable. Then the following conditions are equivalent:*

1. *$f \circ g = \nabla F$ where F is a twice-differentiable convex function.*
2. *The Jacobian $J_{f \circ g}(\mathbf{x})$ is symmetric for all $\mathbf{x} \in \mathbb{R}^{C-1}$.*
3. *$J_{f \circ g}(\mathbf{x})$ is positive semi-definite for all $\mathbf{x} \in \mathbb{R}^{C-1}$.*
4. *$f \circ g$ is monotone.*

Theorem 4.3. *Let $f : \mathbb{R}^{C-1} \to S$ and $g : \mathbb{R}^{C-1} \to \mathbb{R}^{C-1}$ be differentiable where $S \subseteq \mathbb{R}^{C-1}$, and $J_f(\mathbf{x})$ and $J_g(\mathbf{x})$ are symmetric and positive definite for all $\mathbf{x} \in \mathbb{R}^{C-1}$. Then $f \circ g$ is the gradient of a twice-differentiable Legendre function.*

Experiments

- LEGENDRETRON (LT) shows superior performance over multinomial logistic regression (MLR) and Integrated Squared Gaussian Process (ISGP)

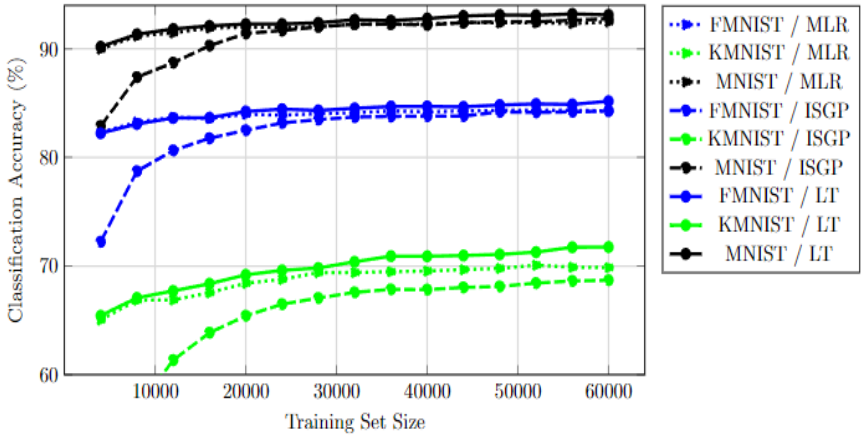


Figure 1. Test performance v.s. training set size for MNIST, KMNIST and FMNIST datasets

Table 3. Test classification accuracies (%) on LIBSVM, UCI and Statlog datasets; at varying levels of label noise (η)

Dataset	# Features	# Classes	$\eta = 0\%$		$\eta = 20\%$		$\eta = 50\%$	
			LT	MLR	LT	MLR	LT	MLR
aloi	128	1,000	**88.11±0.03**	10.34±0.42	**83.03±0.06**	7.07±0.45	**75.23±0.07**	3.53±0.29
sector	55,197	105	**89.71±0.18**	8.77±0.73	**81.00±0.28**	4.12±0.44	**57.38±0.31**	3.17±0.47
letter	16	26	**79.82±0.30**	53.37±0.25	**74.17±0.21**	51.24±0.28	**64.28±0.26**	46.78±0.41
news20	62,061	20	**75.65±0.72**	63.09±0.58	**73.48±0.20**	50.49±1.16	**51.72±0.16**	31.54±1.83
Sensorless	48	11	**88.31±0.19**	34.42±0.46	**82.63±0.99**	32.70±0.50	**52.02±0.78**	29.35±0.84
vowel	10	11	**79.72±1.03**	44.58±1.08	**63.77±1.36**	43.44±1.17	**40.94±1.61**	35.42±1.45
usps	256	10	**95.23±0.16**	93.79±0.17	92.88±0.15	92.95±0.19	90.23±0.26	90.48±0.27
segment	19	7	**95.95±0.24**	87.86±0.40	**92.21±0.40**	87.28±0.40	**86.56±0.47**	82.75±0.46
satimage	36	6	**86.97±0.19**	83.93±0.28	**84.93±0.25**	81.16±0.28	77.44±0.29	77.39±0.29
glass	36	6	58.72±1.94	52.09±1.88	53.72±1.98	50.47±2.11	42.56±1.92	45.47±1.67
vehicle	18	4	**76.91±0.65**	64.94±0.43	**73.59±0.79**	63.06±0.53	**60.94±1.25**	55.18±1.20
dna	180	3	92.79±0.30	**94.43±0.19**	82.61±0.51	**89.55±0.31**	58.23±1.05	**64.18±0.81**
svmguide2	20	3	56.01±1.40	56.01±1.40	56.01±1.40	56.01±1.40	51.65±2.81	52.41±3.04
wine	13	3	96.94±1.14	97.78±0.59	90.97±1.92	96.25±0.99	69.44±2.89	77.36±2.46
iris	4	3	86.67±3.89	83.00±2.08	80.00±3.71	81.50±2.27	63.50±5.13	70.67±3.83