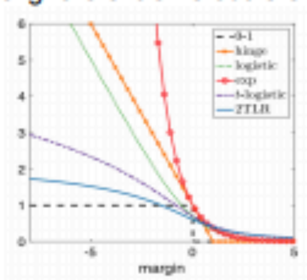
Two-temperature logistic regression based on the Tsallis divergence

Ehsan Amid, Manfred Warmuth, & Sriram Srinivasan
University of California, Santa Cruz
Google Brain, Zürich

Problem and Motivation

- Convex losses are not generally robust to outliers
- Non-convex losses are more desirable, i.e. those that do not grow indefinitely & thus ignore extreme outliers
- We would like to be able to control the level of non-convexity and maintain properties such as Bayes-consistency and properness



Properties

- The normalization constant for probabilities can be determined efficiently using an iterative algorithm
- Loss is Bayes-consistent, even in the non-convex case
- The loss is not proper in general, but the correct distribution can be recovered using a simple transformation
- Quasi-convex loss is significantly more robust to noise

Background

- Generalized log and exponential functions

$$\log_t x = \frac{1}{1-t}(x^{1-t} - 1)$$

$$\exp_t(x) = [1 + (1-t)\,x]_+^{1/(1-t)}$$

- Tsallis divergence

$$D_t(p\|q) = -\int p(\mathbf{x}) \log_t \frac{q(\mathbf{x})}{p(\mathbf{x})}\, d\mathbf{x}$$

Approach

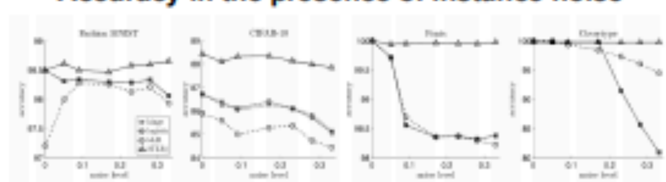
- Two-temperature surrogate loss

$$\frac{1}{N}\sum_n \underbrace{\left[-\log_{t_1} \exp_{t_2}(\mathbf{w}_{c_n}^\top \mathbf{x}_n - G_{t_2}(\mathbf{W}^\top \mathbf{x}_n))\right]}_{\xi_{t_1}^{t_2}(\mathbf{x}_n, c_n \,|\, \mathbf{W})}$$

- Loss is bounded when $0 < t_1 < 1$
- Distribution has a heavier tail when $t_2 > 1$
- Loss is convex when $t_2 = t_1 \geq 1$ and quasi-convex when $t_2 > t_1$. Also, $t_1 = t_2 = 1$ recovers logistic loss

Results

Accuracy in the presence of instance noise



Classification accuracy with 10% label noise

Dataset (#instances, #dim)	Noise Type	Classification Accuracy (%)			
		hinge	logistic	t-LR	2TLR
Fashion MNIST (20K, 784)	random	96.42 ± 0.59	96.42 ± 0.59	94.09 ± 0.48	**99.80 ± 0.12**
	small-margin	98.50 ± 0.26	98.50 ± 0.26	97.35 ± 0.42	**99.13 ± 0.37**
	large-margin	96.42 ± 0.59	96.42 ± 0.59	94.09 ± 0.48	**99.80 ± 0.12**
CIFAR-10 (10.8K, 1024)	random	84.27 ± 1.12	84.39 ± 1.17	82.11 ± 1.01	**87.75 ± 1.40**
	small-margin	84.94 ± 0.97	84.94 ± 0.99	84.22 ± 0.79	**86.28 ± 1.18**
	large-margin	77.79 ± 1.20	77.77 ± 1.20	72.58 ± 1.44	**88.56 ± 1.20**
Fonts (143K, 411)	random	83.78 ± 0.28	83.78 ± 0.28	**84.14 ± 0.27**	**84.14 ± 0.27**
	small-margin	**83.60 ± 0.34**	**83.60 ± 0.34**	83.38 ± 0.36	**83.60 ± 0.34**
	large-margin	72.39 ± 0.32	72.39 ± 0.32	**72.61 ± 0.30**	**72.61 ± 0.30**
Covertype (287K, 54)	random	97.52 ± 0.88	97.52 ± 0.88	**99.26 ± 0.05**	**99.26 ± 0.05**
	small-margin	96.79 ± 0.11	96.79 ± 0.11	**97.25 ± 0.05**	**97.25 ± 0.05**
	large-margin	83.59 ± 0.24	83.59 ± 0.24	84.79 ± 0.20	**94.03 ± 0.13**

Runtime comparison

Dataset (#instances, #dim)	Runtime (s)			
	hinge	logistic	t-LR	2TLR
Fashion MNIST (20K, 784)	4.40 ± 0.28	4.57 ± 0.12	7.02 ± 0.29	7.35 ± 1.21
CIFAR-10 (10.8K, 1024)	31.90 ± 0.22	31.86 ± 0.29	35.08 ± 0.81	28.34±10.56
Fonts (143K, 411)	49.47 ± 4.92	49.75 ± 4.99	82.85 ± 4.98	58.78 ± 6.14
Covertype (287K, 54)	6.45 ± 0.17	6.42 ± 0.18	66.66 ± 1.20	24.53 ± 1.24