Learning with Symmetric Label Noise: The Importance of Being Unhinged

Brendan van Rooyen, Aditya Krishna Menon, Bob Williamson

NICTA

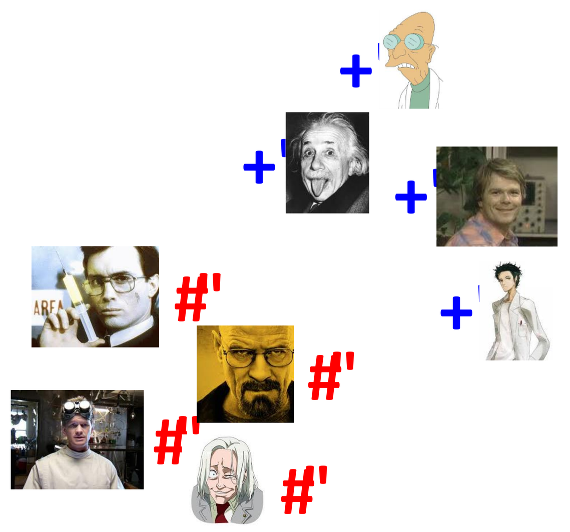
Australian National University

Q: Can we design a loss that is convex but robust to label noise?
A: Yes! Just unhinge the hinge loss from SVMs.

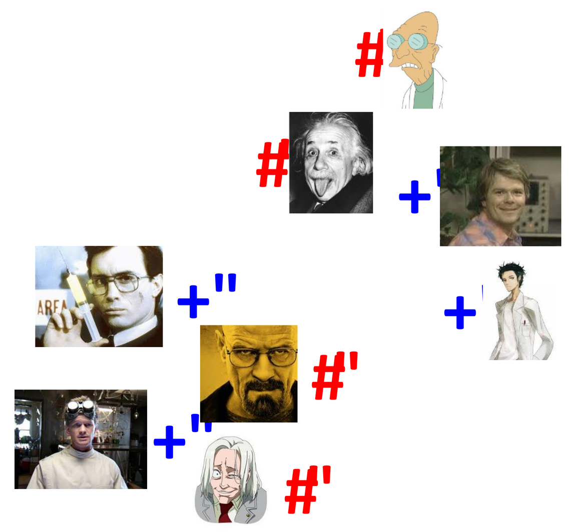
Q: Is there an accurate classification rule that is robust to label noise?
A: Yes! Just use the mean classifier.

The Symmetric Label Noise Problem

Want: Samples from notional "clean" distribution $\boldsymbol{D}$



Get: Samples from "corrupted" distribution $\widetilde{\boldsymbol{D}}$, where labels are flipped with probability σ



Q: Can we still learn a good classifier?

The Usual Approach – Convex Surrogates

The usual approach to learning classifiers is via the minimization of convex potential loss function over a class of linear functions (hinge loss for the SVM, logistic loss for logistic regression, exponential for boosting and so on…). This approach works well if the training samples are clean, but if they are corrupted by noise……..

(Long & Servedio, 2010): with a linear function class, **any** convex potential minimiser resorts to random guessing under nonzero symmetric label noise! This leads to the folk theorem that for robustness to label noise, one needs a non-convex loss.

None of the standard losses are robust to label noise, in fact (Long & Servedio, 2010) says that all convex potential losses share this property.

The devil is in the details: we can circumvent the result if we consider losses that are convex, but not convex potentials!

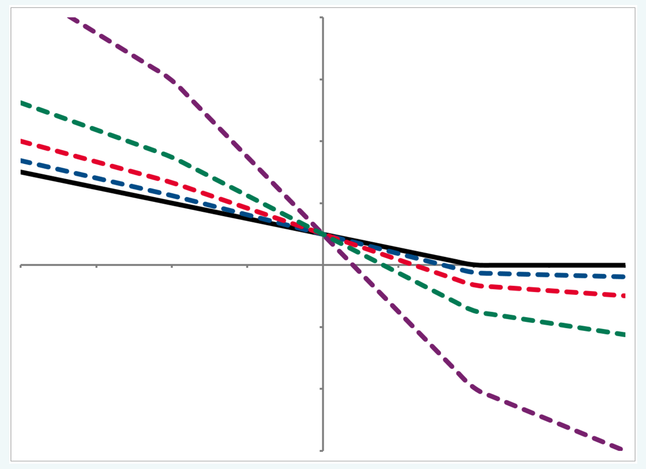
Corruption-Corrected Losses

(Natarajan et al., 2013): introduced a method to correct for symmetric label noise. For any loss ℓ, they associated a corrected loss,

$$\tilde{\ell}(y,v) = \frac{(1-\sigma)\ell(y,v) - \sigma\ell(-y,v)}{1-2\sigma}$$

with the property that for all classifiers f, $R_\ell(f,D) = R_{\tilde{\ell}}(f,\widetilde{D})$. These losses give "bonus points" for correctly classifying a noisy label. Noise corrected losses allow one to learn from corrupted data, **if you know** σ……

Example: for hinge loss, we get a series of negatively unbounded loss functions. Corrected hinge loss is non-convex, other losses remain convex. Ask for details ☺.



Robustness to Label Noise and the Unhinged Loss

To progress, we seek a loss function that is "unaltered" by the above correction, in the sense that,

$$\tilde{\ell}(y,v) = \alpha\ell(y,v) + \beta$$

for some constants α and β. It turns out (see the paper for the details) that for this to occur $\ell(1,v) + \ell(-1,v) = C$, for some constant C. None of the standard losses satisfy this property….however the following **unhinged loss** does!

$$\ell(y,v) = 1 - yv$$

The unhinged loss is classification calibrated. That is, given a rich enough function class, minimizing this loss will yield the optimal classifier for 0-1 loss. Furthermore,

$$regret_{01}(f,D) \le regret_\ell(f,D) = \frac{1}{1-2\sigma} regret_\ell(f,\widetilde{D})$$

so that minimizing the unhinged loss on corrupted samples is consistent means of learning classifiers.

Linear Function Classes and the Mean Classifier

Linear approaches to learning classifiers, such as the SVM, minimize the regularized objective,

$$\min_{\omega\in\boldsymbol{H}} \frac{1}{|S|}\sum_{(x,y)\in S} \ell(y,\langle\omega,\phi(x)\rangle) + \frac{\lambda}{2}\|\omega\|^2$$

where $\phi: X \to \boldsymbol{H}$ is a feature map. For the unhinged loss, performing this minimization is very easy! We have the following **closed form** expression for the optimal weight vector,

$$\omega^* = \frac{1}{\lambda|S|}\sum_{(x,y)\in S} y\phi(x)$$

Note that the regularization parameter only **scales** the weight vector, and therefore makes no difference to the outputted classifier. The final classifier is expressed simply as a **kernel mean.**

$$f(x') = \frac{1}{|S|}\sum_{(x,y)\in S} y\,\boldsymbol{K}(x,x')$$

Extra Goodies that are in the Paper/ Future

All this and more features as a chapter of Brendan's PhD thesis. Ask for details on:

1) Characterizing linear loss' importance when learning under symmetric label noise.
2) Simulating linear loss with high regularization.
3) Robustness properties of linear loss for more general noise processes.
4) Speeding up the evaluation of the mean classifier via kernel herding.
5) Corruption-corrected losses for more general noise, as well as losses that remain convex after being corrected.
6) More general statistical results for learning with corrupted data, featuring both upper and lower bounds.

One Line Summary

"While the truth is rarely pure, it can be simple"