More sensitive tests in DELVE

David J.C. MacKay
University of Cambridge
Cavendish Laboratory
Madingley Road
Cambridge CB3 0HE
mackay@mrao.cam.ac.uk

Abstract

The current test in DELVE, for sum squared error loss, has the unfortunate feature that if a method does really lousily on a few test cases, the posterior standard deviation of its mean loss is increased such that the test is weakened and may well give a null result.

By the simple hack of modifying the loss function to sum $|\text{error}|^p$ with $p < 2$, for example $p = 0.5$ or $p = 0.1$, it seems plausible that the sensitivity of DELVE might be improved.

DELVE test whether the mean losses of two learning methods are significantly different, under the implicit assumption that the differences in losses have a benign symmetric distribution. When the loss function is the standard sum squared error, this has the consequence that a poor quality method which does particularly badly on a few test cases may be found *not* to be significantly different in performance from a consistently good method, because the large variance of the loss of the poor method causes the standard deviation of its mean loss is increased such that the test is weakened. Another way of putting this is that DELVE's test assumes that 'if method A does really badly on occasion, maybe method B also does really badly sometimes' (RMN).

The best solution would be to make a better statistical test. But perhaps we can squeeze more out of DELVE for little effort. It seems plausible that in some cases where the results are 'not significant', but ought to have been, maybe a modified test would have given a significant result.

To be specific, let us consider using the loss function per test case,

$$\mathcal{E}_p = |\text{error}|^p \tag{1}$$

with $p < 2$, for example $p = 0.5$ or $p = 0.1$.

What we are looking for is a value of p such that a blatantly asymmetric distribution for the traditional loss $\mathcal{E}_2$ is transformed into a more bell–shaped distribution for $\mathcal{E}_p$.

Here I plot the probability distribution for the nasty case where the errors are actually Cauchy distributed,

$$P(x) = \frac{1}{\pi} \frac{1}{(x^2 + 1)}, \tag{2}$$

where x is the error.

For the traditional choices of mean square error ($p = 2$) and mean absolute error ($p = 1$) the probability distribution of the loss is so skew, it does not even have a finite mean. But as we reduce p, the probability distribution of the loss becomes more bell–shaped, and its mean becomes finite. In the case $p = 1/2$, the density of $\mathcal{E}$ is

$$P(\mathcal{E}_{1/2}) = \frac{4}{\pi} \frac{\mathcal{E}_{1/2}}{1 + \mathcal{E}_{1/2}^4} \quad (\mathcal{E}_{1/2} \geq 0), \tag{3}$$

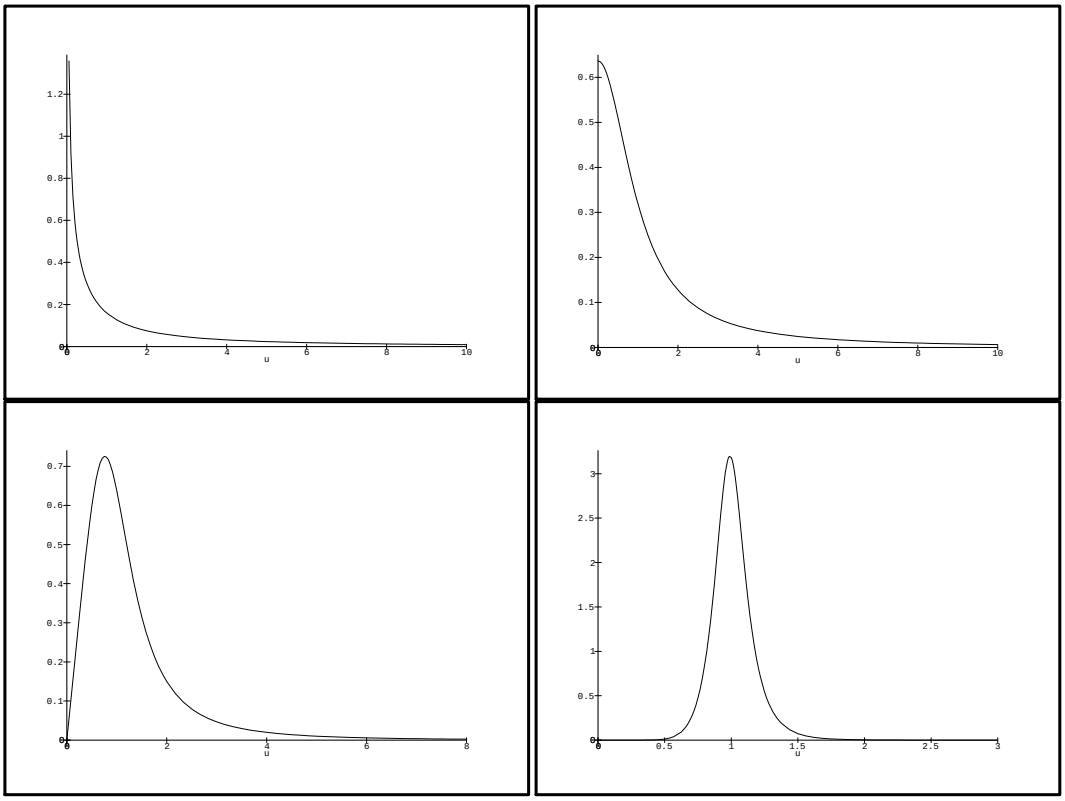

Figure 1: Probability density of $\mathcal{E}_p$ for $p = 2$, $p = 1$, $p = 1/2$, $p = 0.1$, assuming that the errors have a Cauchy distribution.

a distribution whose mean is $\sqrt{2}$. In the case $p = 0.1$, the density of $\mathcal{E}$ is

$$P(\mathcal{E}_{0.1}) = \frac{20}{\pi} \frac{\mathcal{E}_{0.1}^{9}}{1 + \mathcal{E}_{0.1}^{20}} \ (\mathcal{E}_{0.1} \geq 0). \tag{4}$$

These densities are shown in figure 1.

Discussion

Which value of p will give the most significant result will depend on the actual distribution of losses. Perhaps after a little experimentation, a small suite of values of p could be settled upon.

Thanks

Thanks to Radford Neal for discussions. And thanks to any DELVE user who checks out this suggestion!
19th September 1997. DJCM.

Radford's response

> DELVE already supports absolute error loss as well as squared error loss. We haven't really investigated systematically, but the number of significant results using it doesn't seem dramatically different. In my experience, the main reason for non-significant results that one might think should be significant is that variability with respect to training sets is high (as when net-mc doesn't come close to convergence in the allotted time for some training sets, but does for others). This probably can't be cured by fiddling the loss function.