Addendum to EMNLP 2011 Paper
“Closing the Loop...”

Burr Settles
May 18, 2012

“It is virtually impossible to get anything exactly right.”
— Carl de Boor

This is a brief addendum to previous work (Settles, 2011), which corrects some errors reported in the results in Table 1. These errors were introduced by third-party code, and the main results of the paper still hold. This note is intended to shed more light on the somewhat surprising result that the proposed multinomial naive Bayes (MNB) variant drastically outperformed maximum entropy (MaxEnt) models trained using generalized expectation (GE) criteria (Druck et al., 2008).

Through conversations with Gregory Druck (the author of the MaxEnt/GE software), we discovered that this was partly due to a bug in the implementation. In particular, my experiments used the MaxEntGETrainer class in the MALLET\(^1\) library 2.0.5 release. This contained some errors in the optimization procedure used for training. Specifically, the \(L^2\) regularization penalty was included in the gradient calculations but not the value computation, which seems to have caused the L-BFGS optimizer to terminate training prematurely. This bug does not seem to adversely effect GE training when large numbers of labeled words (GE constraints) are available. However, with only ten labeled words per class — as with my experimental setup — the code performed poorly. After a bug fix, which is available since the MALLET 2.0.7 release, GE performs significantly better\(^2\).

\(^1\)http://mallet.cs.umass.edu/
\(^2\)This bug is likely related to the poor performance of GE in other evaluations (Small et al., 2011, for example). However, Druck et al. (2008) used different code and was not affected.

![Classification Accuracy](image)

![Training Time (seconds)](image)

Figure 1: New comparison of the MNB variant (Settles, 2011) vs. a corrected GE implementation, and the previous incorrect one. Statistical significance (two-tailed \(t\)-test at 95% after Bonferroni correction) is marked with *.

Figure 1 shows updated accuracies and training times using the exact same experimental setup as reported in the original paper. In the previous evaluation, MNB was always more accurate than GE (significantly so for six of the eight corpora). Now MNB still dominates in six of the eight corpora, and is significantly better on five of them. However, (1) the corrected GE significantly beats MNB on the
WebKB and Movie Reviews corpora, and (2) it is consistently more competitive on all corpora than the previous, incorrect GE implementation.

Reuters and 20 Newsgroups are somewhat “clustery” data sets — that is, the class labels correspond to natural clusterings of documents — which might explain why the EM-based semi-supervised approach employed by the MNB variant still performs better on these tasks. Perhaps WebKB and Movie Reviews are less clustery in this sense, or maybe the naïve independence assumptions of MNB are more violated, giving MaxEnt/GE an advantage. For example, we would expect sports documents to naturally cluster by sport, but movie reviews might naturally cluster by movie genre rather than sentiment, and university webpages might cluster by discipline rather than whether they are course, faculty, or student pages. Nevertheless, the training time of GE still routinely takes much longer than the MNB variant, and for corpora of any significant size (4,000+ documents: Reuters, WebKB, 20 Newsgroups, Science) it requires anywhere from five seconds to several minutes. This is probably too long for a truly interactive annotation (Card et al., 1986; Fails and Olsen, 2003). Thus, the choice of MNB with informative priors and EM for the interactive DUALIST system is still well-justified.

Acknowledgments

Thanks to Gregory Druck for bringing this to my attention and helping to get to the bottom of it.

References


