The Importance of Calibration for Estimating Proportions from Annotations: Supplementary Material

Dallas Card
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
dcard@cmu.edu

Noah A. Smith
Paul G. Allen School of CSE
University of Washington
Seattle, WA, 98195, USA
nasmith@cs.washington.edu

1 Experimental Details

For tuning the base classifier, we used grid search to choose the strength of regularization strength, testing 11 values from 0.01 to 1000. On each experiment, the training set was split into five random folds. A classifier was trained for each four-fifths of the data, using the remaining fifth as a validation set in each case. The validation set was used to choose regularization strength (using F_1 or calibration as a performance metric), as well as to estimate secondary models, such as ACC or Platt scaling. The predicted proportions from each of the five models (one for each development fold) were then averaged to produce the final estimate of proportions. Reweighting, Platt scaling, CC, and ACC were all based on the model trained using F_1. For Platt scaling, we do not regularize the secondary model, but instead replace the binary labels with smoothed target values, as suggested in the original paper (Platt, 1999). Because ACC can result in inadmissible values in extreme cases, we threshold its predictions to be in the range $[0, 1]$.

2 Datasets

Media Frames Corpus. For this dataset, we treat each framing dimension for each of three issues (immigration, same-sex marriage, and smoking) as a separate subtask. Because there are fewer labeled instances in this dataset than the others, we only create a single split into a source and target corpus for each subtask, treating the articles published before 2009 as a source corpus, and testing on articles from 2009–2012. Most documents in this dataset were annotated by two annotators, so we weight these inversely proportional to the number of annotators for each instance.

Amazon reviews. For this dataset, we made use of the 5-core subsets for five mid-sized product categories: 1) clothing, shoes and jewelry; 2) home and kitchen; 3) sports and outdoors; 4) toys and games; and 5) tools and home improvement, and treat the proportion of people rating the review as "helpful" as the target. For each category, we create separate subtasks by treating each pair of adjacent years in the range 2010–2014 as a source and target corpus (using the earlier year as the source and the later as the target). As with the MFC, we weight instances with multiple votes inversely proportional to the total number of votes per instance.

Yelp reviews. For this dataset, we used three pairs of cities with approximately the same numbers of reviews: Toronto and Scottsdale; Charlotte and Pittsburgh; and Tempe and Henderson. For each pair, we created multiple subtasks by treating each pair of adjacent years as a source and target corpus, respectively, for the years 2009–2017. We ignore the star rating, the title of the review and information about the author, and only consider the review text and location (as a label).

Twitter sentiment. For this dataset, we only make use of what is designated as the official training set (which is the vast majority of instances). Similar to the other datasets, we create subtasks by creating a source and target corpus from each pair of adjacent days for which both days had at least 4,000 tweets. Note that the tweets from after day 166 appear to be artificially biased (containing only positive or negative tweets), thus we exclude these from the analysis.

3 Simulation Details

To simulate a comparison of PCC and SRS when we are able to randomly sample instances to be labeled from the target corpus, we generate sparse binary data and sparse weights and then fit a model with the same form and hyperparameters to a subset of the data. Specifically, we use the following data generating process, for $i = 1, \ldots, N$ and $j = 1, \ldots, P$:

$$\begin{aligned}
X_{ij} &\sim \text{Bernoulli}(p_x) \\
\beta_j &\sim \text{Laplace}(0, 1) \\
\beta_0 &\sim \mathcal{N}(0, 1) \\
p_i &= \text{Sigmoid}(X_{i,:} \cdot \beta + \beta_0) \\
y_i &\sim \text{Bernoulli}(p_i)
\end{aligned}$$

We then fit this model to a subset of the data using an l_1-regularized logistic regression model with regularization strength equal to 1, and average the predicted probabilities over all instances (PCC), or simply average the observed labels in the subset (SRS). Figure 3 in the paper was made using values of $N = 20000$, $P = 10000$, and $p_x = 0.01$, averaged over 200 repetitions, varying the amount of labeled data available to the models.

4 Variance of Simple Random Sampling

As noted in the paper, if we were able to sample and annotate data from the target corpus *with replacement*, the variance of SRS for binary labels would be $\frac{\bar{p}(1-\bar{p})}{L}$, where $\bar{p} = \frac{1}{N_T}\sum_{i=1}^{N_T} p_i$, and $p_i = p(y_i = 1 \mid \boldsymbol{x}_i)$. In the case where we sample a random set of instances from the target corpus and annotate each one exactly once, the variance of the resulting estimate is somewhat more complicated, as there are two sources of randomness – the set of instances selected for annotation (A) and the labels returned by the annotation function (Y). Using the law of total variance, we have

$$\begin{aligned}
&\mathbb{V}_{A,Y}[\hat{q}^{\text{SRS}}] \\
&\quad = \mathbb{E}_A[\mathbb{V}_Y[\hat{q}^{\text{SRS}} \mid A]] + \mathbb{V}_A[\mathbb{E}_Y[\hat{q}^{\text{SRS}} \mid A]].
\end{aligned} \tag{1}$$

Note that the first component in Equation (1) will be zero if $p_i = 0$ or $p_i = 1, \forall i$, and is maximized if $p_i = 0.5, \forall i$. Conversely, the second component is equal to zero if all p_i have the same value, and is maximized if the p_is are evenly split between $p_i = 0$ and $p_i = 1$. As such, there is a tradeoff between these two components.

We can further simplify the above terms as follows. First,

$$\begin{aligned}
&\mathbb{E}_A[\mathbb{V}_Y[\hat{q}^{\text{SRS}} \mid A]] \\
&\quad = \mathbb{E}_A\left[\mathbb{V}_Y\left[\frac{1}{L}\sum_{i\in A} y_i \,\middle|\, A\right]\right] && (2)\\
&\quad = \mathbb{E}_A\left[\frac{1}{L^2}\sum_{i\in A} p_i(1-p_i)\right] && (3)\\
&\quad = \frac{1}{L}\frac{1}{N_T}\sum_{i=1}^{N_T} p_i(1-p_i) && (4)\\
&\quad = \frac{1}{L}\left(\bar{p} - \frac{1}{N_T}\sum_{i=1}^{N_T} p_i^2\right) && (5)\\
&\quad = \frac{1}{L}\left(\bar{p} - (S^2 + \bar{p}^2)\right) && (6)\\
&\quad = \frac{1}{L}\left(\bar{p}(1-\bar{p}) - S^2\right), && (7)
\end{aligned}$$

where S^2 is the sample variance of the set of p_is in the target corpus. Similarly,

$$\mathbb{V}_A[\mathbb{E}_Y[\hat{q}^{\text{SRS}} \mid A]] = \mathbb{V}_A\left[\mathbb{E}_Y\left[\frac{1}{L}\sum_{i\in A} y_i \,\middle|\, A\right]\right] \tag{8}$$

$$= \mathbb{V}_A\left[\frac{1}{L}\sum_{i\in A} p_i\right]. \tag{9}$$

For a sufficiently large L and N_T, we can approximate this with the central limit theorem for a finite population (Bellhouse, 2001), which gives us

$$\mathbb{V}_A\left[\bar{p}_A\right] \approx S^2\left(\frac{1}{L} - \frac{1}{N_T}\right). \tag{10}$$

When we only have access to a single label per instance, it is not possible to estimate S^2, but we can nevertheless combine the two parts above and use a standard plug-in estimator to approximate an upper bound on the variance of simple random sampling, $\hat{\mathbb{V}}[\hat{q}^{\text{SRS}}] \approx \frac{\bar{y}(1-\bar{y})}{L}$, where $\bar{y} = \frac{1}{L}\sum_{i\in A} y_i$, and empirically this produces a reasonable, if somewhat pessimistic estimate.

References

D. R. Bellhouse. 2001. The central limit theorem under simple random sampling. *The American Statistician* 55(4):352–357. https://doi.org/10.1198/000313001753272330.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74.