Class Proportion Estimation with Application to Multiclass Anomaly Rejection

Tyler Sanderson [‡]
University of Michigan in Ann Arbor

Clayton Scott
University of Michigan in Ann Arbor

A PROOF OF PROPOSITION 4

Observe

$$
\begin{aligned}
|R(f) - \widehat{R}(f)| &= |\underline{R}_M(f) - \widehat{\underline{R}}_M(f)| \\
&\quad + \left| \sum_{i=1}^{M-1} \pi_i (R_i(f) - \widehat{R}_i(f)) + \sum_{i=1}^{M-1} (\pi_i - \widehat{\pi}_i) \widehat{R}_i(f) \right| \\
&\leq |\underline{R}_M(f) - \widehat{\underline{R}}_M(f)| \\
&\quad + \sum_{i=1}^{M} |R_i(f) - \widehat{R}_i(f)| + \sum_{i=1}^{M-1} |\pi_i - \widehat{\pi}_i|. \qquad \text{(S.1)}
\end{aligned}
$$

From (S.1) and by consistency of the $\widehat{\pi}_i$, it suffices to show that

$$\sup_{f \in \mathcal{F}_{k(\boldsymbol{n})}} |\underline{R}_M(f) - \widehat{\underline{R}}_M(f)| \to 0 \qquad \text{(S.2)}$$

and that for each i, $1 \leq i < M$,

$$\sup_{f \in \mathcal{F}_{k(\boldsymbol{n})}} |R_i(f) - \widehat{R}_i(f)| \to 0 \qquad \text{(S.3)}$$

in probability as $\boldsymbol{n} \to \infty$. For $i < M$, (S.3) follows from the standard (two-class) VC theorem (Devroye et al., 1996), by (7), and because the standard VC dimension of $\{x : f(x) \neq i\}_{f \in \mathcal{F}}$ is upper bounded by the multiclass VC dimension.

To establish (S.2), recall Eqns. (5) and (6). For brevity we omit the dependence of $R_{i\ell}$ and $\widehat{R}_{i\ell}$ on f at times. For any f

$$
\begin{aligned}
&|\underline{R}_M(f) - \widehat{\underline{R}}_M(f)| \\
&\leq \left[|R_{0M} - \widehat{R}_{0M}| + \sum_{i=1}^{M-1} |\pi_i R_{iM} - \widehat{\pi}_i \widehat{R}_{iM}| \right] \\
&= \Bigg[|R_{0M} - \widehat{R}_{0M}| \\
&\qquad + \sum_{i=1}^{M-1} |\pi_i (R_{jM} - \widehat{R}_{iM}) + (\pi_i - \widehat{\pi}_i) \widehat{R}_{iM}| \Bigg] \\
&\leq \Bigg[|R_{0M}(f) - \widehat{R}_{0M}(f)| \\
&\qquad + \sum_{i=1}^{M-1} \Big(|R_{iM}(f) - \widehat{R}_{iM}(f)| + |\pi_i - \widehat{\pi}_i| \Big) \Bigg].
\end{aligned}
$$

Standard VC theory (Devroye et al., 1996) implies that for any $\epsilon > 0$ and for $0 \leq i \leq M-1$, $\sup_{f \in \mathcal{F}_k} |R_{iM}(f) - \widehat{R}_{iM}(f)| \to 0$ with probability one, by (7), and because the standard VC dimension of $\{x : f(x) \neq M\}_{f \in \mathcal{F}}$ is upper bounded by the multiclass VC dimension. The other terms tend to zero in probability by consistency of the $\widehat{\pi}_i$. The result now follows.

B PROOF OF THEOREM 1

Consider the decomposition into estimation and approximation errors,

$$R(\widehat{f}) - R^* = R(\widehat{f}) - \inf_{f \in \mathcal{F}_{k(\boldsymbol{n})}} R(f) + \inf_{f \in \mathcal{F}_{k(\boldsymbol{n})}} R(f) - R^*.$$

The approximation error converges to zero by the stated approximation property and because $k(\boldsymbol{n}) \to \infty$.

To establish convergence in probability of the estimation error, let $\epsilon > 0$. For each positive integer k, let $f_k^* \in \mathcal{F}_k$ such that $R(f_k^*) \leq \inf_{f \in \mathcal{F}_k} R(f) + \frac{\epsilon}{4}$. Then

$$
\begin{aligned}
&R(\widehat{f}) - \inf_{f \in \mathcal{F}_{k(\boldsymbol{n})}} R(f) \leq R(\widehat{f}) - R(f^*_{k(\boldsymbol{n})}) + \frac{\epsilon}{4} \\
&\leq \widehat{R}(\widehat{f}) - \widehat{R}(f^*_{k(\boldsymbol{n})}) + \frac{\epsilon}{2} \\
&\quad \text{(with prob. tending to 1, by previous result)} \\
&\leq \tau_{k(\boldsymbol{n})} + \frac{\epsilon}{2} \\
&\leq \epsilon,
\end{aligned}
$$

where the last step holds for $\boldsymbol{n}$ sufficiently large. The result now follows.

[‡] Current affiliation: Google Inc

C ADDITIONAL DETAILS OF EXPERIMENTS

For each permutation of each dataset, hyperparameters for Kernel Logistic Regression were selected via grid-search maximizing classification accuracy using 3-fold cross validation. For the subsequent binary classification step between each training class and the test sample, the bandwidth parameter from the previous step is used (to save computation) but the regularization parameter is again selected, this time to maximize area under the ROC curve.

Before fitting our ROC regression models, we employed a Bayesian bootstrap method to reduce noise and provide better fits (Gu et al., 2008). The Bayesian bootstrap method also provided confidence intervals on the ROC. By fitting the model from Eqn. (9) to the lower confidence interval of the ROC, we were able to estimate an upper confidence interval on $\widehat{\pi}$. We estimate a corresponding lower confidence interval as one minus the sum of the remaining class upper confidence intervals. Table 1 shows the percentage of true class proportions which fall between the upper and lower estimated 95th-percentile confidence intervals. As expected for the two sided interval, we see it is valid in greater than 90% of cases. We also find that the bounds are tighter when more examples are available.

Note we truncated the sizes of some multiclass datasets in order to process them in a timely manner. Namely, the Opportunity dataset (Roggen et al., 2010), and the SensIT dataset (Duarte and Hu, 2004).

References

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

Jiezhun Gu, Subhashis Ghosal, and Anindya Roy. Bayesian bootstrap estimation of roc curve. *Statistics in medicine*, 27(26):5407–5420, 2008.

D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, G. Trster, P. Lukowicz, G. Pirkl, D. Bannach, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, and J. Mill n. Collecting complex activity data sets in highly rich networked sensor environments. In *Proc. 7th Int. Conf. on Networked Sensing Systems*, 2010.

M. Duarte and Y. H. Hu. Vehicle classification in distributed sensor networks. *Journal of Parallel and Distributed Computing*, 64(7):826–838, 2004.

Table 1: Percentage of true class proportions that fall in the estimated $\widehat{\pi}$ 95th percentile confidence intervals, and the standard deviation of the upper confidence interval from the true class proportion.

Dataset (# Classes)	% in range	Train Counts	Test Counts	Upper-Interval Std. Dev.
All Binary	0.947			0.26
All Multiclass	0.972			0.10
Australian (2)	0.955	350	153	0.17
Banana (2)	0.991	2677	1188	0.06
Breast-cancer (2)	0.900	140	41	0.54
Diabetis (2)	0.991	389	134	0.29
German (2)	0.982	506	150	0.34
Image (2)	0.945	1167	495	0.10
Ionosphere (2)	0.918	178	63	0.23
Ringnorm (2)	0.982	3738	1832	0.03
Saheart (2)	0.891	234	80	0.41
Splice (2)	0.964	1605	763	0.11
Thyroid (2)	0.818	109	33	0.28
Twonorm (2)	0.991	3738	1849	0.03
Waveform (2)	0.982	2526	824	0.08
SensIT (3)	0.991	1011	492	0.17
DNA (3)	0.985	1011	474	0.09
Opportunity (4)	0.975	1150	300	0.12
SatImage (6)	0.982	2241	536	0.06
Segment (7)	0.949	1167	165	0.09