Literature Review

Bardia Mojra

February 5, 2021

Robotic Vision Lab

The University of Texas at Arlington

1 Obtaining Well Calibrated Probabilities Using Bayesian Binning

- code: `https://github.com/pakdaman/calibration/`
- paper: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4410090/`
- citation: [1]

1.1 Introduction

In this paper, the authors propose a novel calibration method for probabilistic predictive models. Bayesian Binning into Qualities (BBQ), is a non-parametric and post-processing calibration method. Their proposed method is a binary classifier calibration method is based on the histogram-binning calibration method [2]. It is important to note this method could be extended to multi-class classification tasks [3].

1.2 Problem Statement

In machine learning, classification problems are often solved by deploying a a predictive model trained on some given data set but often underperform and make miscalibrated predictions ("classifier score"). Statistically speaking, for a calibrated prediction of i.e. 40%, there will be an occurrence of 40% for a give test set that is 1) large enough with respect to solution space size, 2) test data set is randomly selected (for more insight read on *Central Limit Theorem*.

Figure 1 is a reliability curve [4] [5] that is used as an example of a predictive model with poorly estimated probabilities.

1.3 Related Work

Mainly, calibration is done two ways 1) *ab initio*, by modifying objective function which increases computational cost. 2) It can be done as a post-processing procedure. Post processing can be categorized into parametric and non-parametric. Platt's method is an example of parametric calibration, [6]. Non-parametric methods include histogram- binning [2], Platt scaling [6], and isotonic regression [3].

1.4 Method

BBQ is an extension of simple histogram binning method [2], with added capability to consider different binning models (different number of bins) and their combinations under a Bayesian framework [7]. The generated Bayesian score provides further insight into network structure and it is also used when combining binning models.

$$Score(M) \ = \ P(M) \ . \ P(D|M)$$

The marginal likelihood, $P(D|M)$, has a closed form solution under the following 3 conditions, [7]:

1. All samples are under i.i.d. assumption and the class distribution $P(Z|B = b)$, which is class distribution for bin b, with a binomial distribution with parameter θ_b.
2. Bin distributions are independent.
3. The prior distribution over binning model parameters θ's are modeled using a Beta distribution.

Marginal likelihood in closed form, [7]:

$$P(D|M) = \prod_{b=1}^{B} \frac{\Gamma(\frac{N'}{B})}{\Gamma(N_b + \frac{N'}{B})} \frac{\Gamma(m_b + \alpha_b)}{\Gamma(\alpha_b)} \frac{\Gamma(n_b + \beta_b)}{\Gamma(\beta_b)}$$

Where:

- $\Gamma(n) = (n-1)!$
- N_b: The total number of training instances in the b'th bin.
- n_b: The total instances of class *zero* among all training instances N_b.
- m_b: The total instances of class *one* among all training instances N_b.
- $P(M)$: The prior distribution of binning model M, uniform distribution for initial condition.

The above equation is used for model averaging by BBQ, they point out that mentioned Bayesian scores could be used for model selection. Per [8], model averaging is superior to model selection methods.

1.5 BBQ

BBQ framework defines calibrated prediction as:

$$P(z=1|y) = \sum_{i=1}^{T} \frac{Score(M_i)}{\sum_{j=1}^{T} Score(M_j)} \ . \ P(z=1|y, M_i)$$

Where:

- T : total number of binning models considered
- $P(z=1|y, M_i)$: probability estimate using model M_i for uncalibrated classifier output y.

1.6 Calibration Measures

- ECE: Expected Calibration Error is calculated over the bins.
- MCE: Maximum Calibration Error is calculated among the bins.

$$ECE = \sum_{i=1}^{K} P(i).|o_i - e_i| \quad , \quad MCE = \max_{i=1}^{K}(|o_i - e_i|),$$

Where:

- o_i: true fraction of positive instances in the ith bin.
- e_i: mean of the post-calibrated probabilities in the ith bin.
- $P(i)$: empirical probability (fraction) of all instances in the ith bin.

1.7 Empirical Results

- Acc: accuracy
- AUC: area under the ROC curve (receiver operator characteristic curve).
- RMSE
- ECE
- MCE

References

[1] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, 2015.

[2] B. Zadrozny and C. Elkan, "Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers," in *Icml*, vol. 1, pp. 609–616, Citeseer, 2001.

[3] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

[4] M. H. DeGroot and S. E. Fienberg, "The comparison and evaluation of forecasters," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 32, no. 1-2, pp. 12–22, 1983.

[5] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632, 2005.

[6] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[7] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," *Machine learning*, vol. 20, no. 3, pp. 197–243, 1995.

[8] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: a tutorial," *Statistical science*, pp. 382–401, 1999.