Summary:

Obtaining Well Calibrated Probabilities Using Bayesian Binning

Code: <https://github.com/pakdaman/calibration/>

Paper: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4410090/>

Introduction

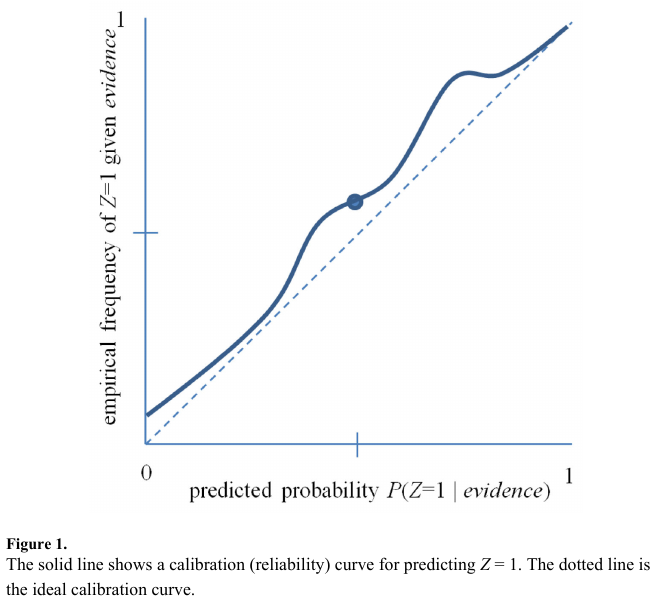
In this paper, the authors propose a novel calibration method for probabilistic predictive models. Bayesian Binning into Qualities (BBQ) is a non-parametric and post-processing calibration method. Their proposed method is a binary classifier calibration method is based on the histogram-binning calibration method ([Zadrozny and Elkan 2001](http://citeseerx.ist.psu.edu/viewdoc/download doi=10.1.1.29.3039&rep=rep1&type=pdf)). It is important to note this method could be extended to multi-class classification tasks, ([Zadrozny and Elkan 2002 (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.13.7457&rep=rep1&type=pdf)).

Problem Statement

In machine learning, classification problems are often solved by deploying 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 (DeGroot and Fienberg 1983; Niculescu-Mizil and Caruana 2005) that is used as an example of a predictive model with poorly estimated probabilities.

Figure 1:



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, [Platt 1999). Non-parametric methods include histogram binning (Zadrozny and Elkan 2001), Platt scaling (Platt 1999), and isotonic regression (Zadrozny and Elkan 2002).

Method

BBQ is an extension of simple histogram binning method ([Zadrozny and Elkan 2001 (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.3039&rep=rep1&type=pdf)), with added capability to consider different binning models (different number of bins) and their combinations under a Bayesian frame work ([Heckerman, Geiger, and Chickering, 1995 (<https://link.springer.com/content/pdf/10.1007/BF00994016.pdf>)). The generated Bayesian score is provides further insight into Bayesian network structure and is used for combining binning models.



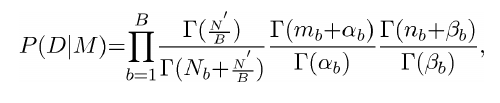
The marginal likelihood, $P(D | M)$, has a closed form solution under the following 3 conditions ([Heckerman, Geiger, and Chickering, 1995 (<https://link.springer.com/content/pdf/10.1007/BF00994016.pdf>)):

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 closed form, ([Heckerman, Geiger, and Chickering, 1995 (https://link.springer.com/content/pdf/10.1007/BF00994016.pdf)):



Where:

+ Gamma(n) = (n-1)!

+ N\_b : total number of training instances in the b'th bin.

+ n\_b : total instances of class \*zero\* among all training instances N\_b.

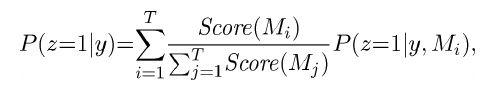
+ m\_b : total instances of class \*one\* among all training instances N\_b.

+ P(M) : 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 Hoeting (Hoeting et al. 1999), model averaging is superior to model selection methods.

BBQ

BBQ framework defines calibrated prediction as:



Where:

+ T: total number of binning models considered.

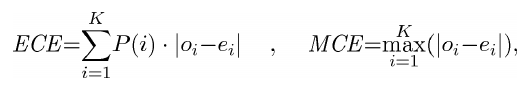
+ P(z=1 | y, M\_i) : probability estimate using model M\_i for uncalibrated classifier output y.

+ B : bins, 

Calibration Measures

+ ECE: Expected Calibration Error is calculated over the bins.

+ MCE: Maximum Calibration Error is calculated among the bins.



Where:

+ o\_i : true fraction of positive instances in the i^th bin

+ e\_i : mean of the post-calibrated probabilities in the i^th bin

+ P(i) : empirical probability (fraction) of all instances in the i^th bin

Empirical Results

+ Acc: accuracy

+ AUC: area under the ROC curve (receiver operator characteristic curve)

+ RMSE

+ ECE

+ MCE

