Obtaining Well Calibrated Probabilities Using Bayesian Binning

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht

2015

Bardia Mojra
February 3, 2021
Robotic Vision Lab
University of Texas at Arlington

Introduction

- 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).
- It is important to note this method could be extended to multi-class classification tasks, (Zadrozny and Elkan 2002).

Introduction

- Bayes' rule:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

- P(A|B): Posterior - probability of A occurring given that B has occurred
- P(B|A): Likelihood – likelihood of B occurring given A has occurred
- P(A): Prior – general probability or initial degree of belief that A has occurred
- P(B): Marginalization – general probability that B has occurred

Introduction

- Bayes' rule:
  - Remember your priors (Bayes' rate neglect)
  - Imagine your theory is wrong. Would the world look different?
  - Update incrementally (snow flakes of evidence)

Problem Statement

- Models are often underperform and make miscalibrated predictions ("classifier score")
- For a prediction of i.e. 40% to be considered calibrated, there has to be an occurrence of 40% for a give large test set
  - Dataset:
    - Must be large enough with respect to solution space size
    - Test dataset must be randomly selected (for more insight read on Central Limit Theorem – super important)

Problem Statement

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.

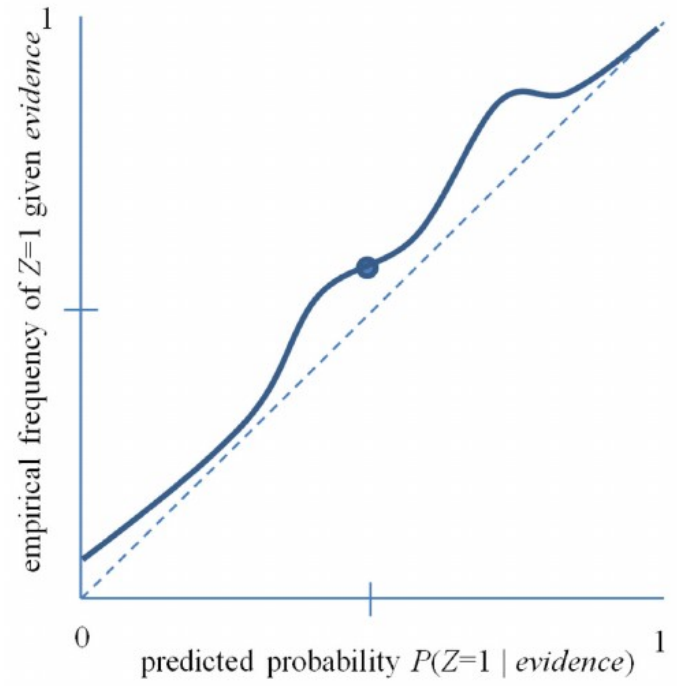

Figure 1.
The solid line shows a calibration (reliability) curve for predicting $Z = 1$. The dotted line is the ideal calibration curve.

Related Work

$$Score(M) = P(M) \cdot P(\mathscr{D}|M) \quad (1)$$

- BBQ is an extension of simple histogram binning method (Zadrozny and Elkan 2001). The authors added capability to consider different binning models (different number of bins) and their combinations under a Bayesian frame work (Heckerman, Geiger, and Chickering, 1995).

- The generated Bayesian score is provides further insight into Bayesian network structure and is used for combining binning models.

Related Work

$$Score(M) = P(M) \cdot P(\mathscr{D}|M) \quad (1)$$

- The marginal likelihood, P(D | M), has a closed form solution under the following 3 conditions (Heckerman, Geiger, and Chickering, 1995):

  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.

Related Work

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

- Marginal likelihood closed form, (Heckerman, Geiger, and Chickering, 1995)
- 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.

BBQ

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

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.

B: number of bins

$$B \in \{\tfrac{\sqrt[3]{N}}{C},\ldots,C\sqrt[3]{N}\},$$

Calibration Measures

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

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

- Where:
  - o_i: true fraction of positive instances in the 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

	SVM	Hist	Platt	IsoReg	BBQ
AUC	0.50	0.84	0.50	0.65	0.85
ACC	0.48	0.78	0.52	0.64	0.78
RMSE	0.50	0.39	0.50	0.46	0.38
ECE	0.28	0.07	0.28	0.35	0.03
MCE	0.52	0.19	0.54	0.58	0.09

(a) SVM Linear

	SVM	Hist	Platt	IsoReg	BBQ
AUC	1.00	1.00	1.00	1.00	1.00
ACC	0.99	0.99	0.99	0.99	0.99
RMSE	0.21	0.09	0.19	0.08	0.08
ECE	0.14	0.01	0.15	0.00	0.00
MCE	0.35	0.04	0.32	0.03	0.03

(b) SVM Quadratic Kernel

Friedman Test with Holm's Post-Hoc Procedure

- Friedman test: non-parametric version of ANOVA test
- In statistics, this technique is used to compare survey and reject null hypothesis.

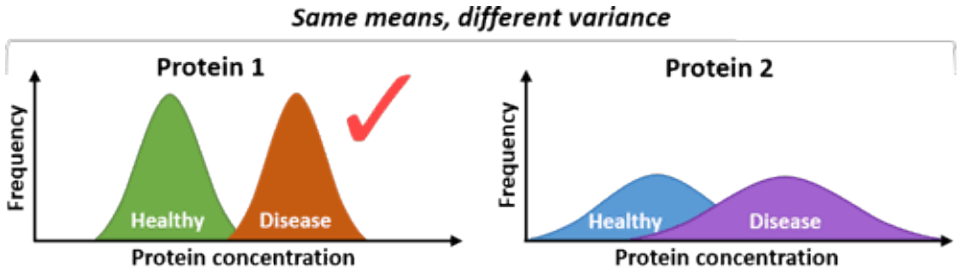

Figure 1. Overlapping histogram plots for concentrations of protein 1 in different populations.

Figure 2. Overlapping histogram plots for concentrations of protein 2 in different populations.

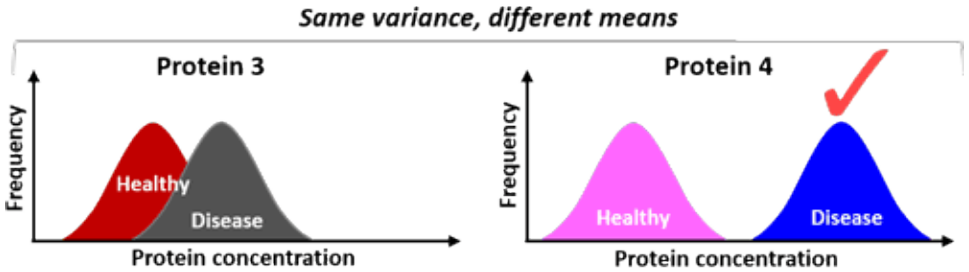

Figure 3. Overlapping histogram plots for concentrations of protein 3 in different populations.

Figure 4. Overlapping histogram plots for concentrations of protein 4 in different populations.

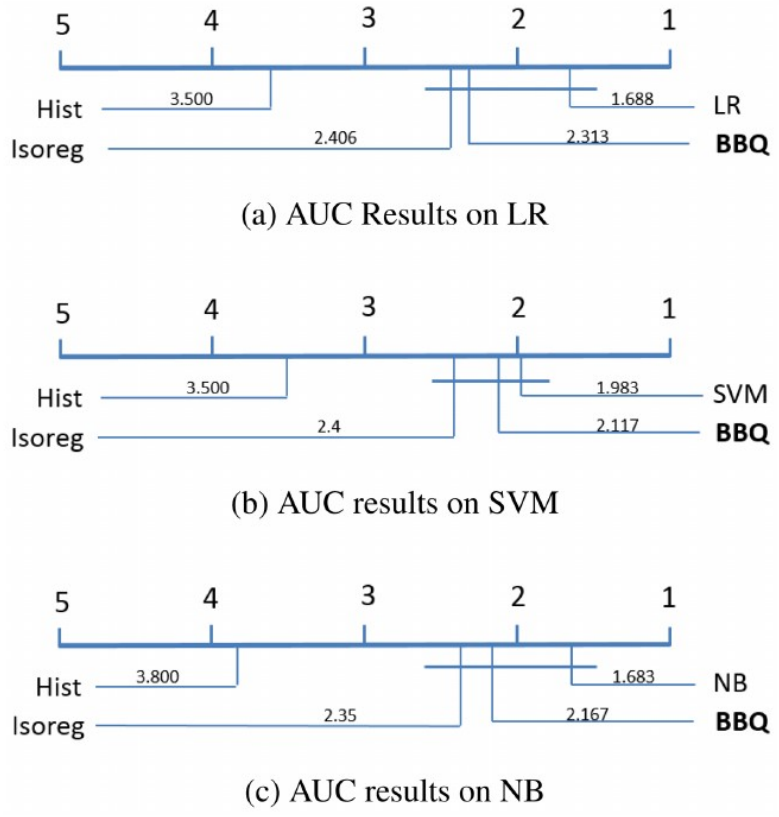

Figure 3.
Performance of each method in terms of average rank of AUC on the real datasets. All the methods which are not connected to BBQ by the horizontal bar are significantly different from BBQ (using Friedman test followed by Holm's step-down procedure at a 0.05 significance level).

› Thank you!