Binary Classifier Calibration Using a Bayesian Non-Parametric Approach

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht

SIAM Data Mining (SDM) 2015

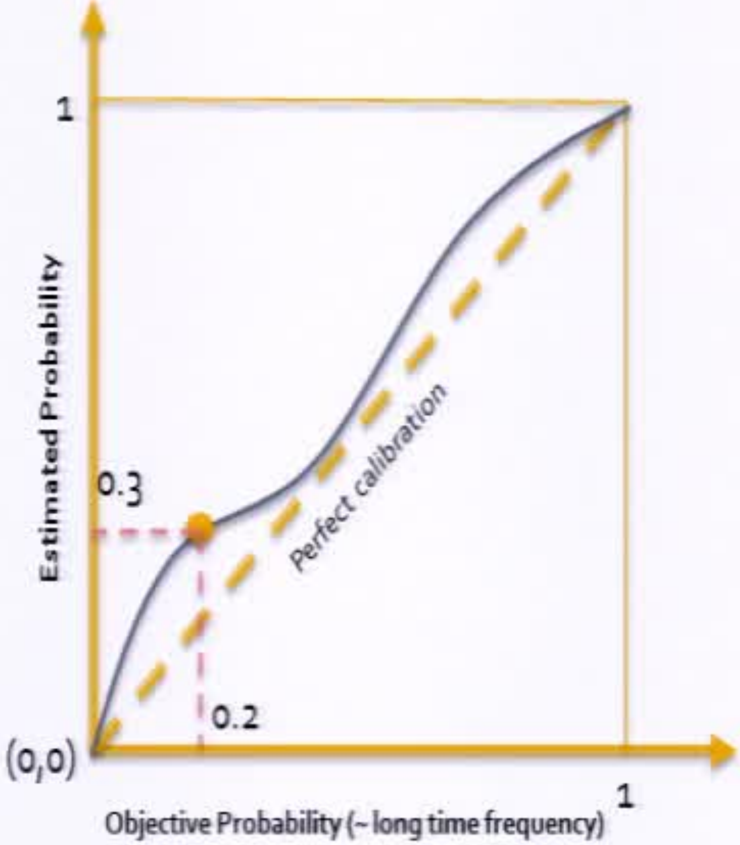
University of Pittsburgh

Talk overview

- Calibration problem
- Review of Existing Calibration Methods
- Bayesian Calibration Methods (SBB-ABB)
- Experimental Results
- Conclusions and Future Work

Problem Definition

- We have a set of **probabilistic predictions of a binary outcome**
- Probabilistic predictions are well-calibrated if the outcomes predicted to occur with probability *p* do occur about *p* fraction of the time



Motivation

- Accurate probability outputs are critical in decision making, outlier detection:
  - Science (e.g., determining which experiments to perform)
  - Medicine (e.g., deciding which therapy to give a patient)
  - Business (e.g., making investment decisions)
- Outputs of many classification models are either:
  - not probabilistic (e.g. SVM)
  - Or, they do not give a well-calibrated probabilistic output (e.g. Naïve Bayes, logistic regression)

Calibrated Classification Models

Methods for learning well-calibrated classification model

- **Calibration is built-in to the classification learning algorithm**
  - Can make the optimization harder

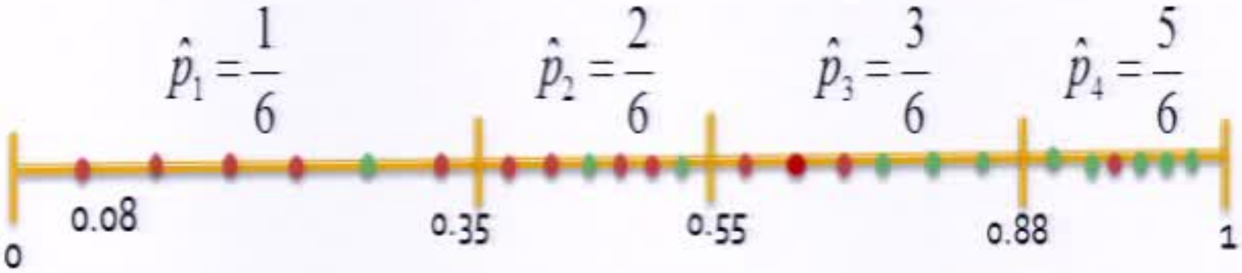
- **Optimized in a separate post-processing step**
  - Learn the classification model using an arbitrary loss function first
  - Calibrate the output in the post-processing step

Post-Processing Calibration Methods

- Platt's calibration method (John C Platt, 1999)

$$P(y=1 \mid p_{in}) = \frac{1}{1+\exp(a \times p_{in} + b)}$$

- Equal frequency histogram binning (B. Zadrozny and C. Elkan, 2001)


$\hat{p}_1 = \frac{1}{6}$
$\hat{p}_2 = \frac{2}{6}$
$\hat{p}_3 = \frac{3}{6}$
$\hat{p}_4 = \frac{5}{6}$
0
0.08
0.35
0.55
0.88
1

Post-Processing Calibration Methods

- Isotonic regression (B. Zadrozny and C. Elkan, 2002)
  - Fits a piecewise-constant non-decreasing function to model $P(y=1 \mid p_{in})$
  - Assumes the classifier ranks the instances correctly
- Adaptive Calibration of Predictions (ACP) (X. Jiang, et. al. 2012)
  - Finds 95% confidence interval (CI) around the predicted value
  - Use observed frequency of instances in the CI as calibrated probability
  - ACP is designed for LR

Bayesian binning

Our approach:

- Bayesian model selection
- Bayesian model averaging

over all possible histogram binning models induced by the training data.

Challenges:

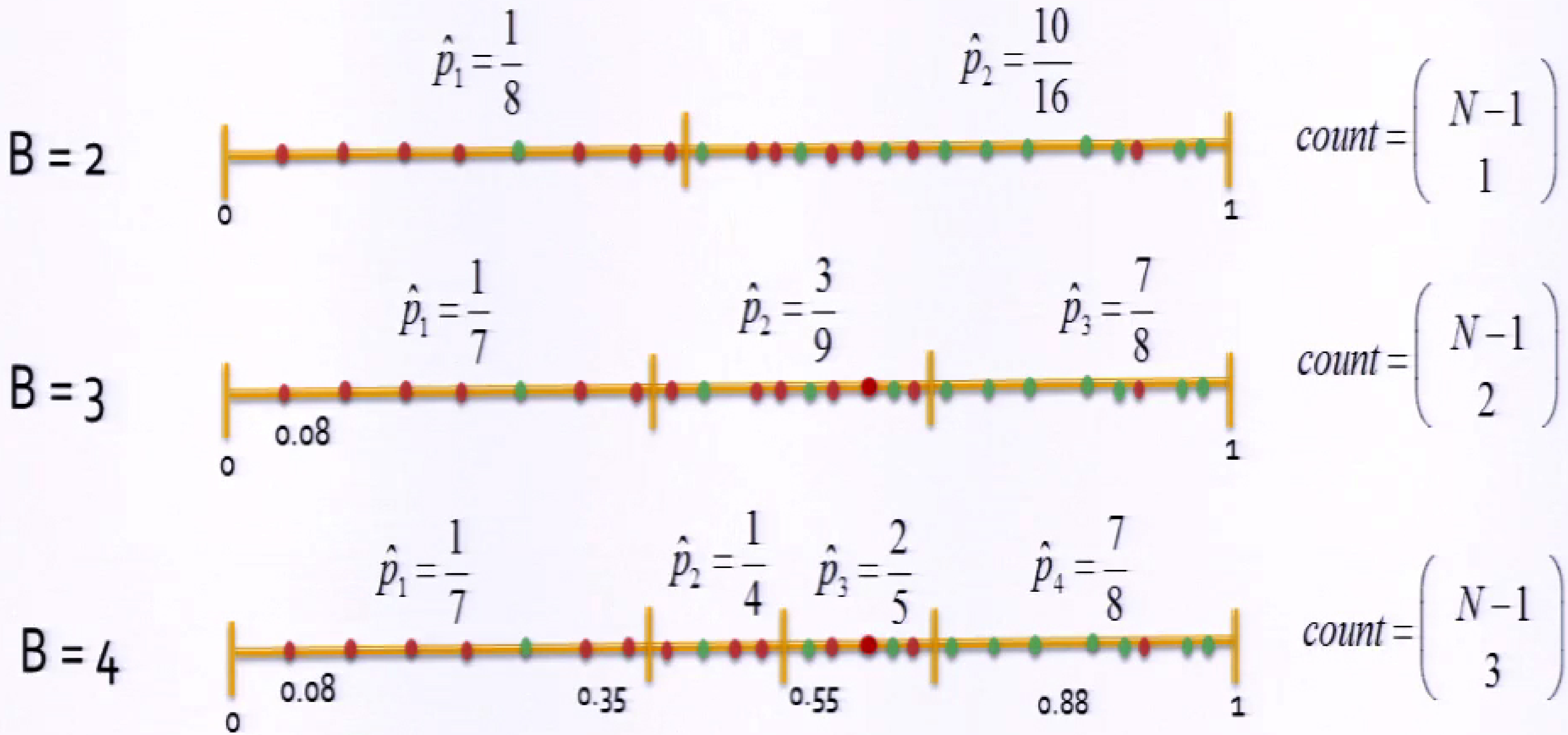
- The number of binning models is exponential in N: 2^N
- How to make the methods more efficient?

Solutions:

- use decomposable Bayesian scoring functions
- use dynamic programming

Bayesian Binning

- Let us assume *N* training data points
- The number of possible binnings of *N* is exponential in *N*



B = 2

$\hat{p}_1 = \frac{1}{8}$ $\hat{p}_2 = \frac{10}{16}$

$count = \binom{N-1}{1}$

B = 3

$\hat{p}_1 = \frac{1}{7}$ $\hat{p}_2 = \frac{3}{9}$ $\hat{p}_3 = \frac{7}{8}$

$count = \binom{N-1}{2}$

B = 4

$\hat{p}_1 = \frac{1}{7}$ $\hat{p}_2 = \frac{1}{4}$ $\hat{p}_3 = \frac{2}{5}$ $\hat{p}_4 = \frac{7}{8}$

$count = \binom{N-1}{3}$

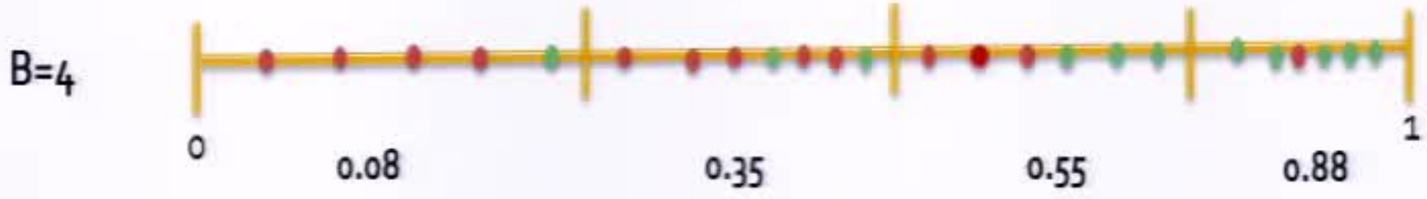
Binning model: preliminaries

- Let *D* denotes all training data sorted according to the input score/probability

  $D = \{(p^1{}_{in}, y_1),...,(p^N{}_{in}, y_N)\}, \;\; p^1{}_{in} \leq p^2{}_{in} ... \leq p^N{}_{in}$

  - p^i_{in} classifier scores for i^{th} instance
  - y_i the true class of i^{th} instance

- Let *M be the binning model*
  - *B*: denotes number of bins
  - *Pa*: denotes partitioning of D into B bins using bin boundaries
  - $\theta = \{\theta_1, \theta_2, ..., \theta_B\}$ parameters of Binomial distributions $\theta_b = P(y=1 \mid p_{in} \in Pa(b))$



Bayesian score

- Let *M* be a binning model
  - *B*: denotes number of bins
  - *Pa*: denotes partitioning of D into B bins using bin boundaries
  - $\theta = \{\theta_1, \theta_2, \ldots, \theta_B\}$ parameters of Binomial distributions

Bayesian score:

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

Model prior

Marginal likelihood of M

Decomposable score:

$$Score(M) = \prod_{b=1}^{B} Score(b, l, u)$$

Marginal Likelihood

$$P(D|M) = \int_{\theta} P(D|M,\theta)P(\theta|M)d\theta$$

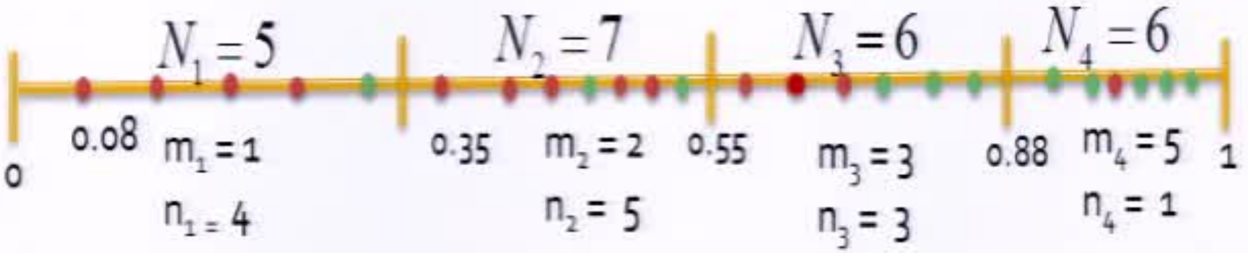
Assuming:

- all samples are i.i.d
- The distributions of the class variable for two different bins are independent
- The priors of these distributions are defined as $P(\theta_b \mid M) = Beta(\theta_b \mid \alpha_b, \beta_b)$

→ **Decomposable marginal likelihood:**

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

K2 score, BDeu score: different choices of prior parameters α_b, β_b



Decomposable model priors

Decomposable model prior:

$$P(M) = \prod_{b=1}^{B} P_{prior}(b,l,u)$$

Examples:

- **A uniform prior**: $P_{prior}(b,l,u)$ independent of the number of bins
- **Prior based on Poisson distribution** (Lustgarden et al 2011)

Let Prior(k) defines the prior probability of having a bin boundary between p_{in}^{k} and p_{in}^{k+1} :

$$Prior(k) = 1 - e^{-\lambda \frac{d(k,k+1)}{d(1,n)}}$$

Assuming the independence of partitioning boundaries, the prior probability of having a bin containing the training instances $\left\{p_{in}^{l_b}, p_{in}^{l_b+1}, \ldots, p_{in}^{u_b}\right\}$ will be calculated as:

$$Prior(u_b)\left(\prod_{k=l_b}^{u_b-1}(1 - Prior(k))\right)$$

Selection over Bayesian Binning (SBB)

Goal: find the binning model with the highest Bayesian score

Idea: Use score decomposability and the dynamic programming to find the best binning efficiently

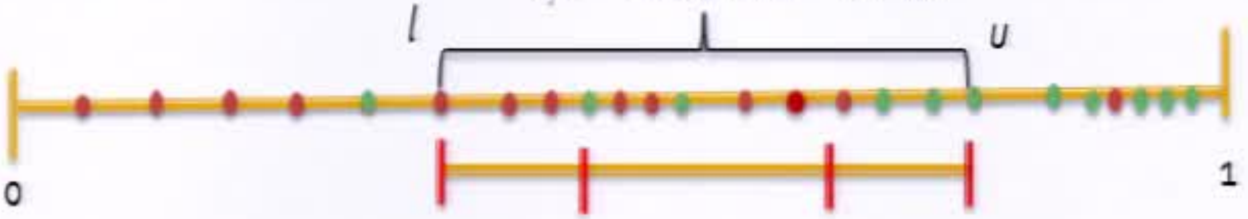
Notation:

$S_{l,u} = \{p_{in}^{l}, p_{in}^{l+1}, \ldots, p_{in}^{u}\}$ a subset of data points for indexes *l* and *u*

$M_{l,u}$ the optimal binning of $S_{l,u} = \{p_{in}^{l}, p_{in}^{l+1}, \ldots, p_{in}^{u}\}$

$v_{l,u}$ the Bayesian score for $M_{l,u}$

$S_{l,u} = \{p_{in}^{l}, p_{in}^{l+1}, \ldots, p_{in}^{u}\}$



$M_{l,u}$ $score_{l,u} = v_{l,u}$

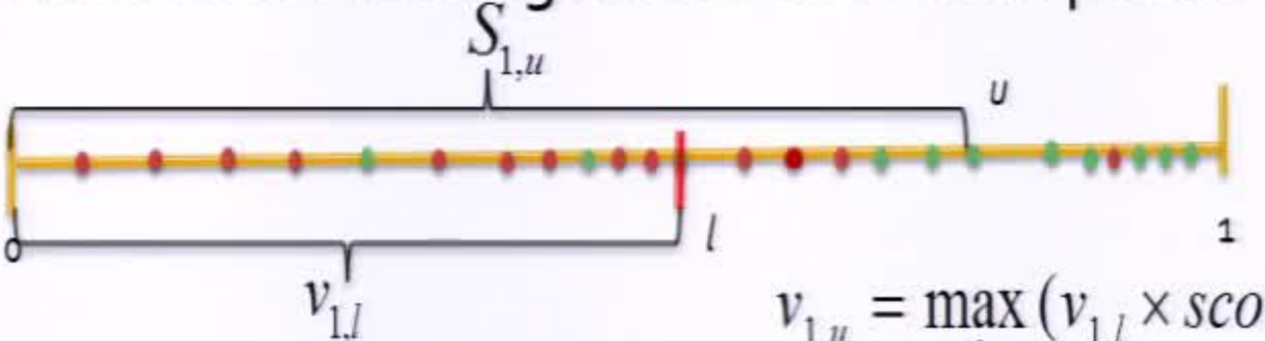
Selection over Bayesian Binning (SBB)

- **Dynamic programming algorithm:**
- Builds the best binning model starting from data with lower indexes
- O(N^2) time
- **Assume:** the best binning models for subsets $S_{1,1}, S_{1,2}, \ldots, S_{1,u-1}$

  have scores $v_{1,0}, v_{1,1}, \ldots, v_{1,u-1}$

  Then $v_{1,u} = \max_{l} (v_{1,l} \times score_{l,u})$

  and $M_{1,u}$ is defined by the optimal choice of l (or $M_{1,l}$)
- $M_{1,N}$ is the best binning model for N data points



$$v_{1,u} = \max_{l} (v_{1,l} \times score_{l,u})$$

Averaging over Bayesian Binning (ABB)

Algorithm: Offline step + Online prediction step

- both require O(N^2) time

Offline step:

- Forward step: sequentially add the contributions of many binning models $v_{1,1}, v_{1,2}, \ldots, v_{1,N}$ and their corresponding Bayesian scores (maximization replaced by summation)
- Backward step: Sequentially add the contributions of many binning models $v_{N,N}, v_{N-1,N}, \ldots, v_{1,N}$ and their scores
- Keep the results of both the forward and the backward step

Averaging over Bayesian Binning (ABB)

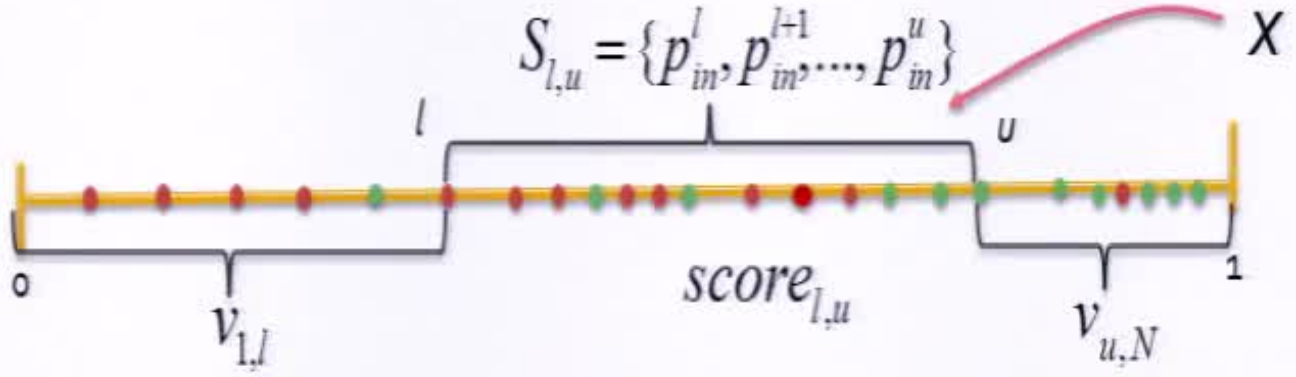
Online prediction step:

For any new data point *x* do the following steps:

- Find the index *k* so that $x \in [p_{in}^{k}, p_{in}^{k+1}]$

$$P(x) \propto \sum_{1 \le l \le k} \sum_{k+1 \le u \le N} (v_{1,l-1} \times Score_{l,u} \times v_{u+1,N} \times \hat{p}_{l,u}(x))$$

- $\hat{p}_{l,u}(x)$ is the frequency of positive instances located in the bin that contains all the training instances indexed by [*l*,…,*u*]



$$S_{l,u} = \{p_{in}^{l}, p_{in}^{l+1}, \ldots, p_{in}^{u}\}$$

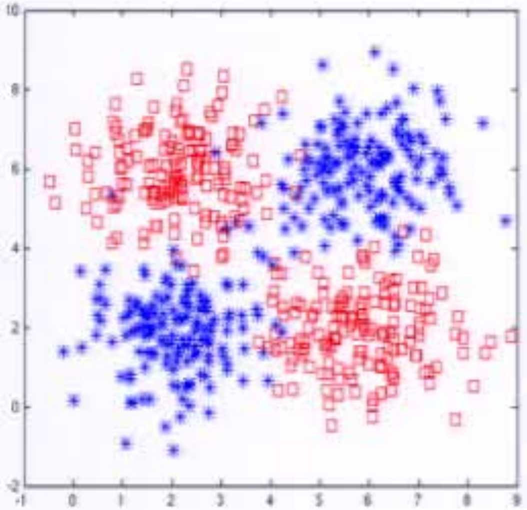
Computational Cost

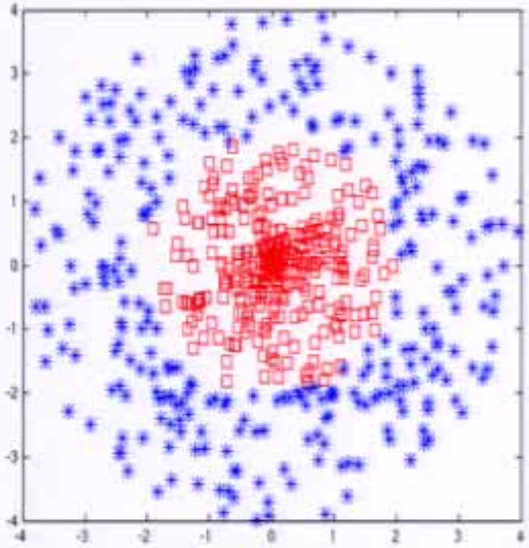
The main drawback of ABB is its time complexity at the test time. One can address this problem simply by caching the probabilities based on the required precision.

	Plat	Hist	IsoReg	ACP	SBB	ABB
Model optimization	O(TN)	O(NlogN)	O(NlogN)	O(NlogN)	$O(N^2)$	$O(N^2)$
Online prediction	O(1)	O(logB)	O(logB)	O(N)	O(logB)	$O(N^2)$

Experiments on Simulated Data: Setup

- **Experiments :**
  - 600 train/calibration
  - 600 test averaging over 10 runs
- **Two simulated datasets:**
  - Parity function data
  - Circular class data
- **Base model:**
  - Logistic Regression
  - Note: LR is not a good model for the data





Experiments: Evaluation metrics

- Discrimination measures: AUC, ACC
- Calibration measures: RMSE, Expected Calibration Error (ECE), Maximum Calibration Error (MCE)

- Partition the interval [0,1] into *K* intervals (K=10, equal frequency bins)
  - o_i is the true fraction of positive instances in the i^{th} interval
  - e_i is the mean of the classifier scores located inside the i^{th} interval
  - *P(i)* fraction of all the instances that fall into the i^{th} interval

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

Experiments on Simulated Data: Results

Parity function data

	LR	ACP	IsoReg	Platt	Hist	SBB	ABB
(Higher is better)							
AUC	0.497	**0.950**	0.704	0.497	0.931	0.914	**0.941**
ACC	0.510	**0.887**	0.690	0.510	0.855	**0.887**	**0.888**
(Lower is better)							
RMSE	0.500	**0.286**	0.447	0.500	0.307	0.307	**0.295**
MCE	0.521	**0.090**	0.642	0.521	0.152	0.268	**0.083**
ECE	0.190	**0.056**	0.173	0.190	0.072	0.104	**0.062**

Circular class data

	LR	ACP	IsoReg	Platt	Hist	SBB	ABB
(Higher is better)							
AUC	0.489	**0.852**	0.635	0.489	0.827	0.816	**0.838**
ACC	0.500	0.780	0.655	0.500	**0.795**	**0.790**	0.773
(Lower is better)							
RMSE	0.501	**0.387**	0.459	0.501	0.394	0.393	**0.390**
MCE	0.540	0.172	0.608	0.539	**0.121**	0.790	**0.146**
ECE	0.171	0.098	0.186	0.171	**0.074**	0.138	**0.091**

Notes

Platt's method: AUC is unchanged, poor calibration performance

Isotonic Regression: AUC can change (performance may improve), calibration is typically poor due to isotonicity

Histogram binning, ACP, SBB and ABB: can improve AUC

Experiments on Real Data: Setup

- Experiments on real world datasets
  - Community Acquired Pneumonia dataset (CAP)
  - UCI datasets: Adult, and SPECT datasets

- Three most commonly used classifiers:
  - Logistic Regression (LR)
  - Support Vector Machine (SVM)
  - Naïve Bayes (NB)

Current and Future Research

- Bayesian averaging over a subset of binning models
- Theoretical results on the quality of binning methods
  - Traditional histogram method:

    by setting $B = c\sqrt[3]{N}$ one can achieve perfect calibration in terms of ECEand MCE, without loosing any discrimination power in terms of AUC

    (preprint : http://arxiv.org/abs/1401.3390)
  - We work on extending the histogram binning theorems for ABB and SBB
- Calibration methods for multi-class classification

Thank You !

?