Robust Artificial Intelligence

Reading Seminar; Tsinghua University

Thomas G. Dietterich, Oregon State University

Goal: Trustworthy Classifiers

- The predictions of the classifier should be correct with high probability
- Cases:
  - Closed world, iid data
  - Open world, iid data
- Cases not considered:
  - Changing world (concept drift, distribution change, covariate shift, etc.)
  - Adversaries

Key idea: A classifier should have a model of its own competence

- Given query x_q and classifier f, $\mathrm{Comp}(f, x_q) = 1$ if the classifier is competent to classify x_q and 0 otherwise
- The coverage of f is the fraction of queries for which f is competent:
  - $\mathrm{Cov}(f) = P_{\mathcal{X}}[\mathrm{Comp}(f, x_q) = 1]$
- We want to find f that maximizes coverage while guaranteeing competence

Notation

- Input space $\mathcal{X}$ of dimension d
- Output space $\mathcal{Y} = \{1, \dots, K\}$ classes
- True joint distribution $P(x, y) = P(x)P(y|x)$
- Training data $\{(x_1, y_1), \dots, (x_N, y_N)\}$ drawn from $P(x, y)$
- Fitted function $f: \mathcal{X} \mapsto \Delta_K$ the K-dimensional probability simplex
- $f(x) = [\hat{p}(y = 1|x), \dots, \hat{p}(y = K|x)]$ class probability vector
- $\hat{y} = \arg\max_k \hat{p}(y = k|x)$ predicted class

- $I[u]$ is 1 if u is true and 0 otherwise
- Some classifiers do not output probabilities (e.g., SVMs), but we will ignore this in our notation

Types of Competence Models

- Calibrated probability models
  - The predicted probability equals the true probability $\hat{p}(y|x) = P(y|x)$
- Competence Region models
  - Define a region of competence, $\mathcal{X}_{comp} \subseteq \mathcal{X}$ such that $\forall x \in \mathcal{X}_{comp}$, $\hat{y}$ is correct with probability $1 - \epsilon$
  - $\mathcal{X}_{comp}$ is usually defined by thresholding the predicted probability or some other confidence function
    - If $\hat{p}(\hat{y}|x) \geq \tau$ then output $\hat{y}$; else abstain
    - If $\mathrm{conf}(x) \geq \tau$ then output $\hat{y}$; else abstain
- Conformal prediction
  - Output a set $C(x)$ such that with probability $1 - \epsilon$, $y \in C(x)$ for all (x, y)

Meeting 1: Calibrated Probabilities

- Reasons for Creating Calibrated Probabilities
- Reason 1: Rational Decision Making
  - If $L(k, k')$ is the loss received if $y = k$, then the expected loss of predicting k' is
    - $\sum_k P(y = k|x)L(k, k')$
  - We can choose k' to minimize this expected loss
  - We can consider other decisions including abstention. Let $L(k, abstain)$ be the cost of abstaining
    - E.g., Cost of asking a person to make the decision

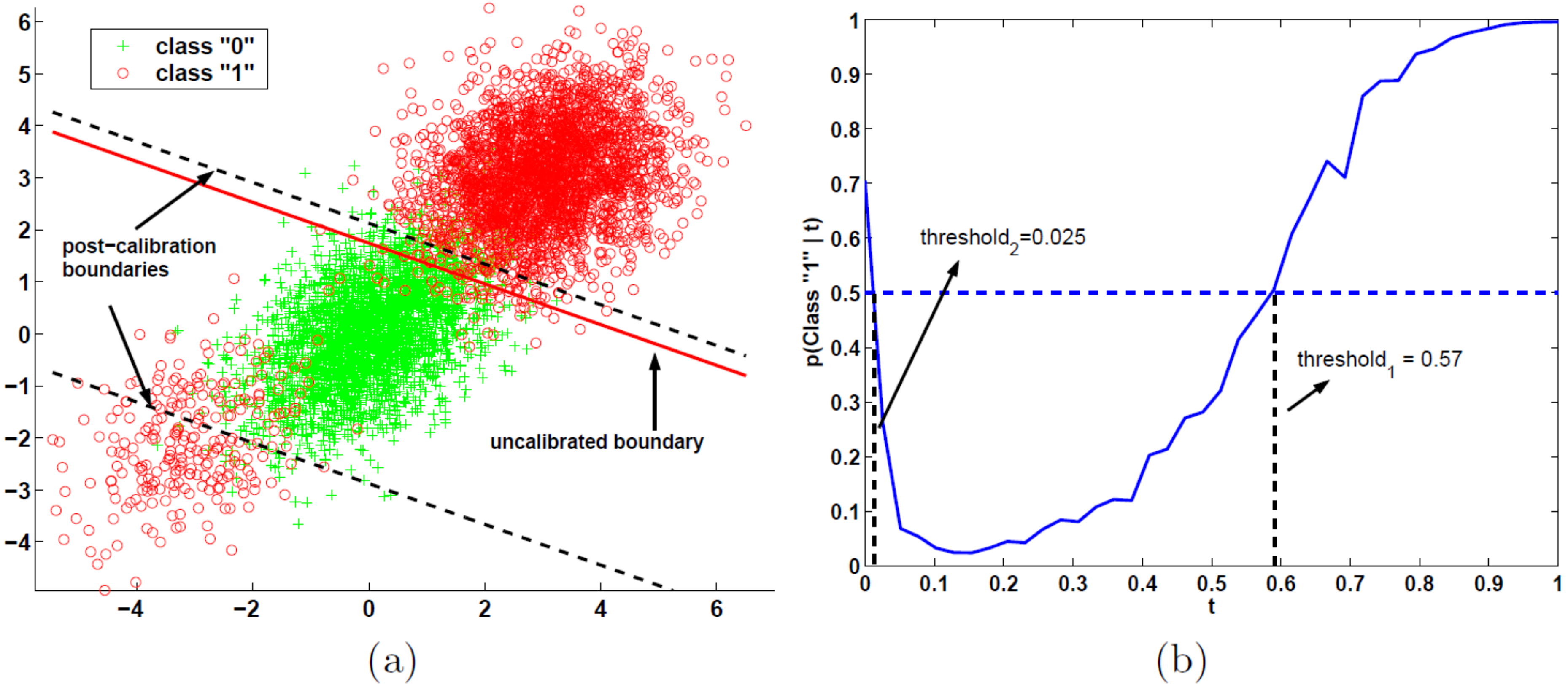
Reason 2: Interpretability

- People can understand a probability statement like $P(y = k|x) = 0.8$ better when the probability is well-calibrated

Reason 3: System Integration

- It is easier to integrate multiple AI subsystems if they all work with well-calibrated probabilities
- Examples:
  - Fusing multiple sensors
  - Combining evidence from multiple sources

Reason 4: Improved Accuracy



Cohen & Goldszmidt

Linear classifier => Non-monotonic calibration curve. Calibration can make it monotonic!

Improved Accuracy (2)

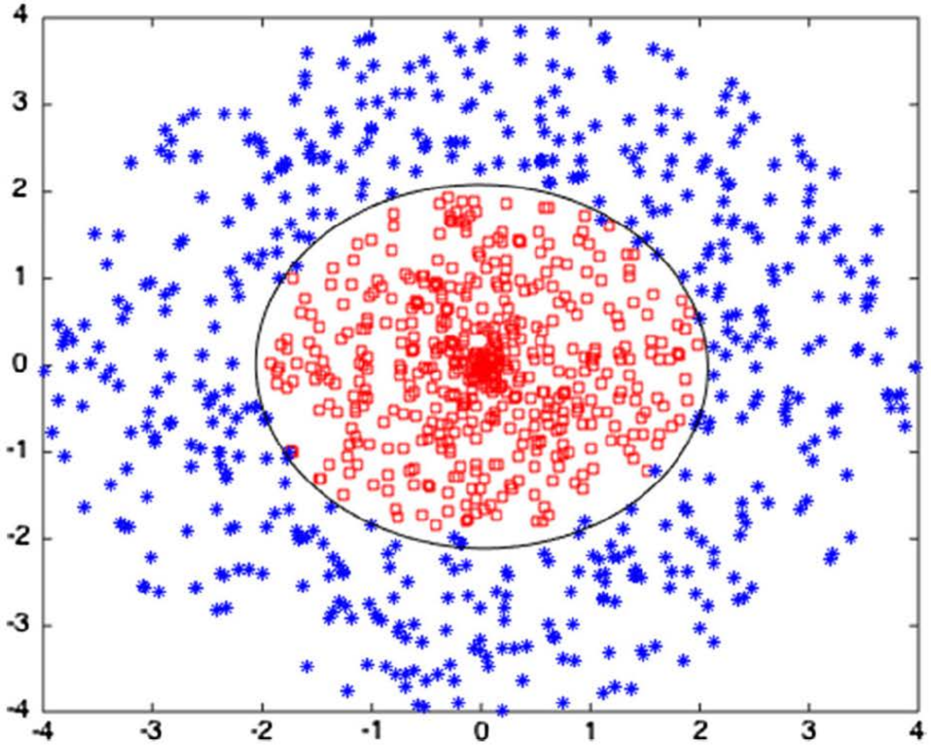


Fig. 3 Scatter plot of the simulated data. The two classes of the binary classification task are indicated by the red squares and blue stars. The black oval indicates the decision boundary found using SVM with a quadratic kernel (colour figure online)

	SVM	IsoRegC	BBQ	ENIR
(a) SVM linear kernel				
AUC	0.52	0.65	0.85	0.85
ACC	0.64	0.64	0.78	0.79
RMSE	0.52	0.46	0.39	0.38
ECE	0.28	0.35	0.05	0.05
MCE	0.78	0.60	0.13	0.12

Of course using a quadratic kernel gives AUC = 1.0 and ACC = 0.99

Improved Accuracy (3)

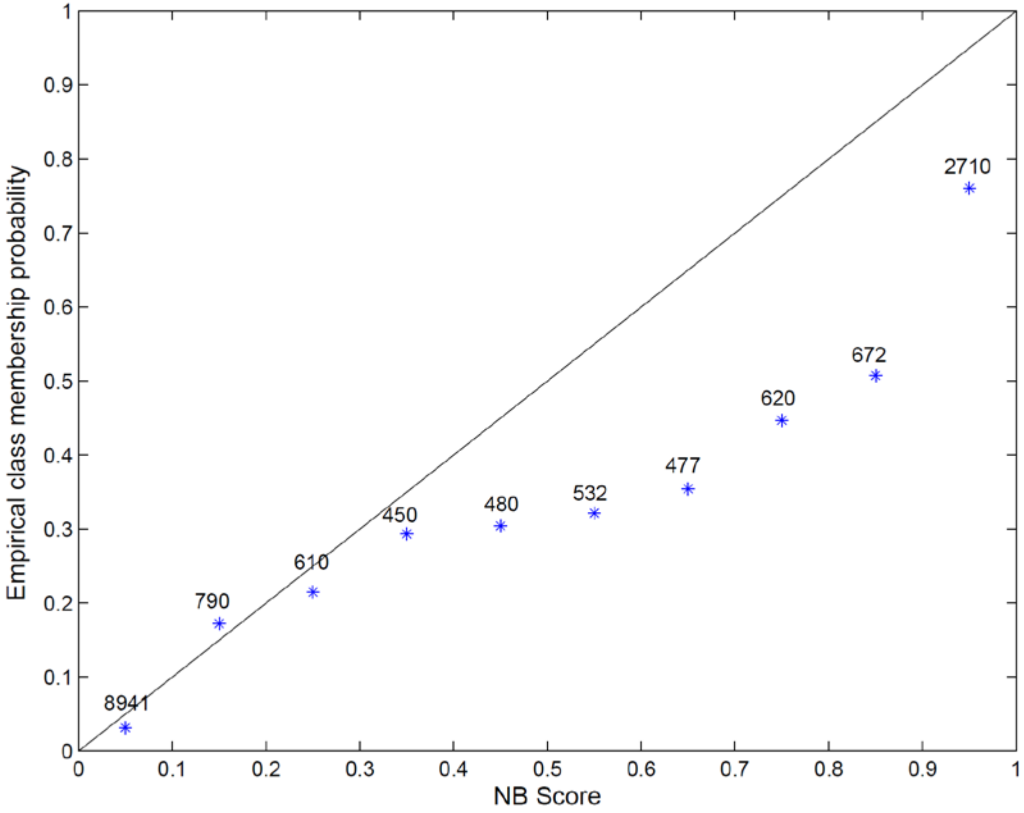
- In two-class problems, $\arg\max_k \hat{p}(y = k|x)$ is equivalent to $\hat{p}(y = 1|x) > 0.5$
- Calibration can find the true 0.5 rather than an optimistic or pessimistic 0.5
- In multiclass problems, the effect can be stronger, especially for rare classes which can be very poorly calibrated

- Not that AUC will be unchanged if $\hat{p}(y = k|x)$ is monotonically increasing with $P(y = k|x)$

Calibrated Classifiers

- A function f is well-calibrated if $\hat{p}(y = k|x) = P(y = k|x)$
- Given a “calibration set” of data points and a classifier, we can compute a reliability diagram
  - Divide $[0,1]$ into M bins (often $M = 10$). Bins may be of equal width or of equal quantiles according to $\hat{p}(\hat{y}|x)$
  - For bin $b \in \{1, \ldots, M\}$, let B_b be the set of points whose probability scores $\hat{p}(\hat{y}|x)$ belong in bin B_b
  - $\hat{p}(B_b) = \frac{1}{|B_b|} \sum_{x \in B_b} \hat{p}(\hat{y}|x)$. This is the average predicted probability of the points in B_b
  - $\hat{P}(B_b) = \frac{1}{|B_b|} \sum_{x \in B_b} I[\hat{y} = y]$. This is the fraction of predictions that are correct.
- Calibration score
  - $\sum_{b=1}^{M} \left[\hat{p}(B_b) - \hat{P}(B_b)\right]^2$ the summed squared calibration error

Reliability Diagram (Naïve Bayes; ADULT)



Zadrozny & Elkan, 2002

Calibration Score and the Brier Score

- The Brier Score is a proper scoring rule for probabilistic models
  - $BrierScore = \frac{1}{N}\sum_i (\hat{p}(\hat{y}_i|x_i) - I[\hat{y}_i = y_i])^2$
- It can also be written in terms of the bins as
  - $BrierScore = \frac{1}{M}\sum_b P_x(B_b)\left[\hat{p}(B_b) - \hat{P}(B_b)\right]^2 + \frac{1}{M}\sum_b P_x(B_b)\left[\hat{P}(B_b)\left(1 - \hat{P}(B_b)\right)\right]$
  - Here $\hat{P}(B_b)$ is $|B_b|/N$
- The first term is the Calibration Score
- The second term is called the “Refinement Score”. It is minimized when $\hat{P}(B_b)$ is near 0 or 1.
- So a classifier that minimizes the BrierScore seeks to be well-calibrated *and* highly certain

Improving Calibration does not necessarily Improve Refinement

- A classifier can be well-calibrated but useless
    - Suppose 70% of the calibration data points belong to class 1
    - Then always predict $\hat{y} = 1$ with $\hat{p}(\hat{y}) = 0.7$
    - This is perfectly calibrated but useless
    - Note that the Refinement score will be large
        - $0.7 \times 0.3 = 0.21$

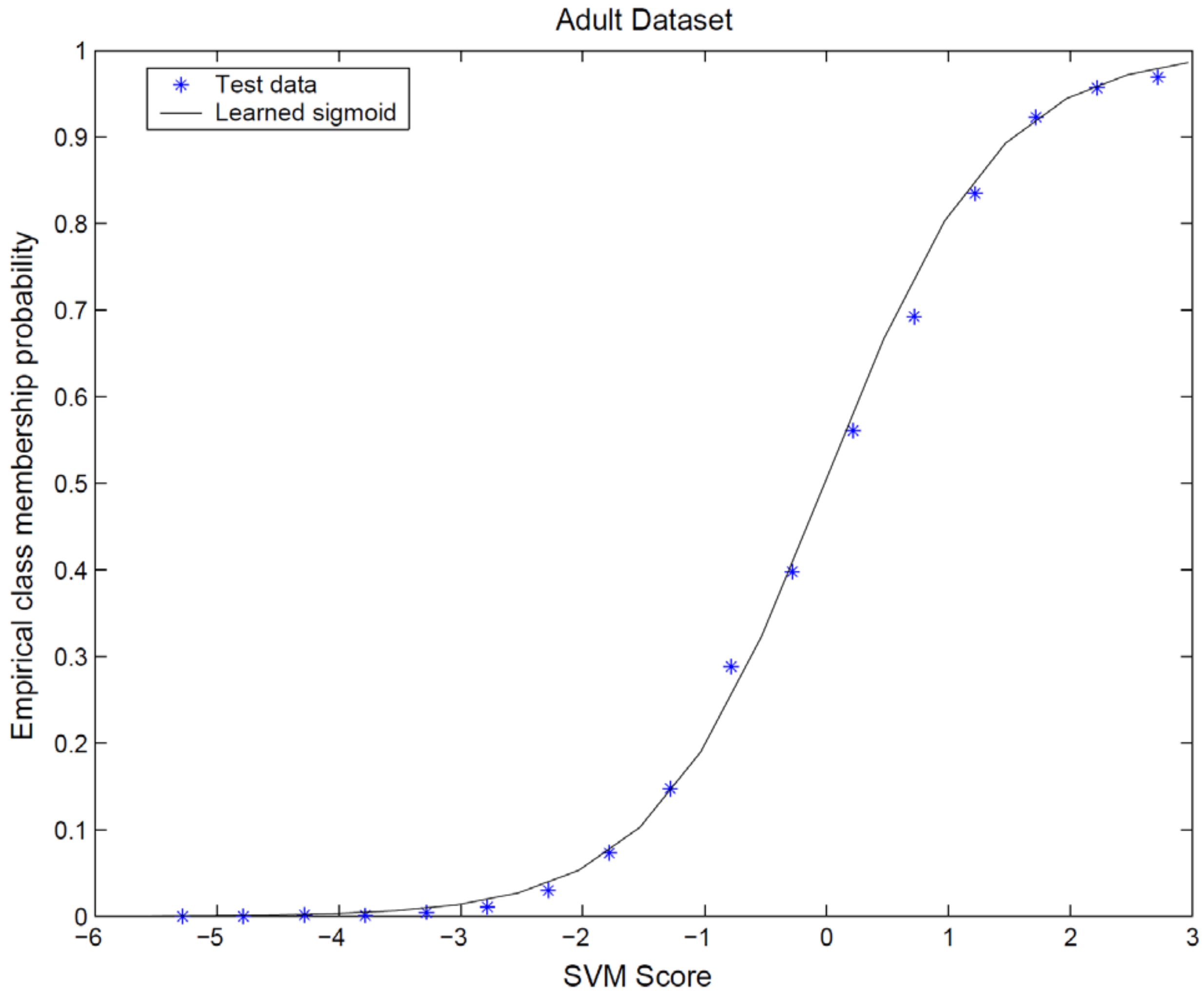
Calibration Method 1: Binning

- Fit a function g to map $\hat{p}$ to P and then replace $\hat{p}$ with $g(\hat{p})$
- “training data” consist of
  - $(\hat{p}_i, I[\hat{y} = y_i])$ pairs
- Fixed-width Bins
  - Sort the data by $\hat{p}$
  - Let $B_1, \ldots, B_M$ each be of width $\frac{1}{M}$
  - Estimate $\hat{P}(B_b)$ for each bin
  - $g(\hat{p}) = \hat{P}(B_b)$ for the bin B_b containing $\hat{p}$
- Quantile Bins
  - Define the bins so that each bin contains $\frac{1}{M}$ of the training data

Calibration Method 2: Platt Scaling

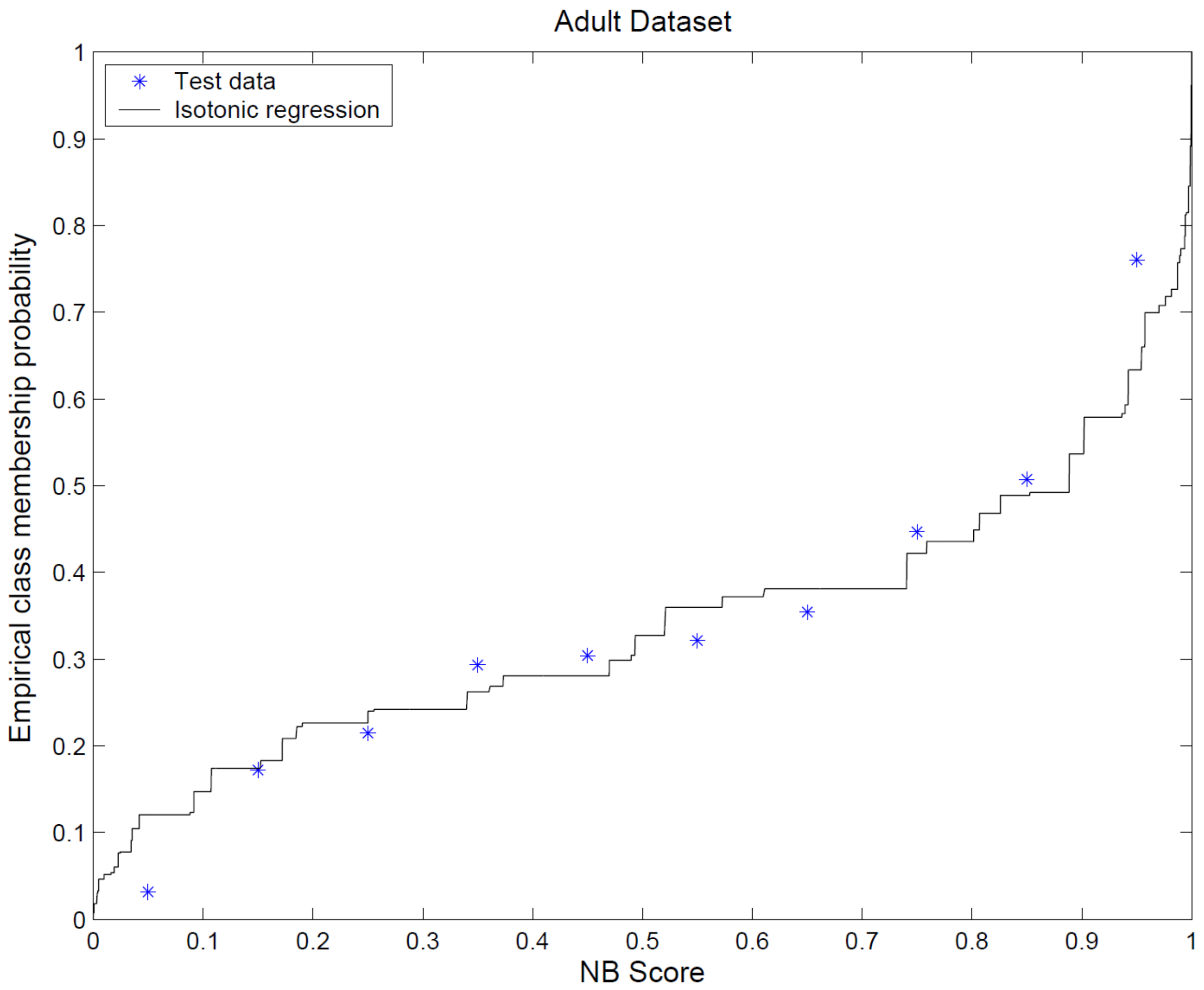
(Platt, 1999)

- $g(\hat{p}; a, b) = \frac{1}{1+e^{a+b\hat{p}}}$
- Logistic regression with a single "feature" ($\hat{p}$)



Method 3: Isotonic Regression

- Find the function g that is monotonically increasing from 0 to 1 and minimizes the Brier Score
- Pool-Adjacent Violators Algorithm
  - Ayer, et al. (1955)
  - Robertson, Wright, & Dykstra (1988)



PAV

Ayer, M., Brunk, H., Ewing, G., Reid, W., Silverman, E. (1955)

- Input: $(\hat{p}_i, y_i)$ sorted in ascending order by $\hat{p}_i$
- Initialize $\widehat{m}_{i,i} = y_i; w_{i,i} = 1$
- While $\exists i \; s.t. \widehat{m}_{k,i-1} \geq \widehat{m}_{k,i}$
  - $w_{k,l} \coloneqq w_{k,i-1} + w_{i,l}$
  - $\widehat{m}_{k,l} \coloneqq \frac{w_{k,i-1}\widehat{m}_{k,i-1} + w_{i,l}\widehat{m}_{i,l}}{w_{k,l}}$
  - Insert $\widehat{m}_{k,l}$ in place of $\widehat{m}_{k,i-1}$ and $\widehat{m}_{k,i}$
- Output the function
  - $\widehat{m}(\hat{p}) = \widehat{m}_{i,j}$ for $\hat{p} \in \left(\hat{p}_i, \hat{p}_j\right]$

Method 4: Regularized Isotonic Regression

- Isotonic Regression can be rewritten as the solution to the following problem
- Choose $\hat{P}_i$ to minimize
  - $\frac{1}{2}\sum_{i=1}^{N}\left(\hat{P}_i - \hat{p}_i\right)^2 + \lambda\sum_{i=1}^{N-1}\left(\hat{P}_i - \hat{P}_{i+1}\right)I\left[\hat{P}_i > \hat{P}_{i+1}\right]$ subject to $\lambda = +\infty$
- Tibshirani, Hastie & Tibshirani (2011) developed mPAVA, which constructs the complete regularization path from $\lambda = 0$ to $\lambda = \infty$
  - Efficient algorithm that produces a sequence of “near isotonic” regression models $g_1, \dots, g_t, \dots$
- ENIR (Ensemble of Near Isotonic Regressions; Naeini & Cooper, 2018) computes the BIC score of each g_t, normalizes these scores, and then computes the weighted average of the models to obtain g

Method 5: Other Flexible Models

- Splines (Lucena, 2018 arxiv 1809.07751)
- Piecewise linear functions via a tree-based decomposition (Leathart, Frank, Holmes, Pfahringer, 2017)
- Gaussian Processes (Song, Kull, Flach, 2018)

Methods for Multiclass Classifiers

- Method 1: Normalized one-vs-rest calibration
  - For each class k, learn a binary calibration function g_k based on a one-vs-rest classifier
  - Define $g\left(\hat{p}(y = 1|x), \dots, \hat{p}(y = K|x)\right)$ as follows
    - Let the predicted probability for class k be

$$\frac{g_k\left(\hat{p}(y = k|x)\right)}{\sum_{k'} g_{k'}\left(\hat{p}(y = k'|x)\right)}$$

Multiclass Method 2: Softmax Temperature Tuning (Guo et al, 2017)

- Let $\mathbf{z} = z_1, \ldots, z_K$ be the final layer outputs of a DNN (prior to the softmax)
- Define $\hat{p}(y = k|x) = \frac{\exp \frac{z_k}{T}}{\sum_{k'} \exp \frac{z_{k'}}{T}} = \sigma_{SM}\left(\frac{\mathbf{z}}{T}\right)$
- Adjust T to fit the calibration data

Multiclass Methods 3 and 4: Generalized Platt Scaling

- Matrix Scaling
  - Learn a matrix $\mathbf{W}$ and vector $\mathbf{b}$ to fit $\sigma_{SM}(\mathbf{W}\mathbf{z}_i + \mathbf{b})$ to a 1-hot encoding of y_i
- Vector Scaling
  - Matrix scaling with $\mathbf{W} = \mathrm{diag}(\mathbf{w})$

Experiments 1: Niculescu-Mizil & Caruana

- Insights
  - Max-margin methods push $\hat{p}$ toward 0.5
  - Naïve Bayes pushes $\hat{p}$ toward 1.0
  - Calibration flattens out this distribution
  - Max-margin methods are fit well by logistic regression (Platt scaling), which also needs relatively little data
  - Isotonic Regression works well with Naïve Bayes but usually requires more calibration data

Boosted Trees

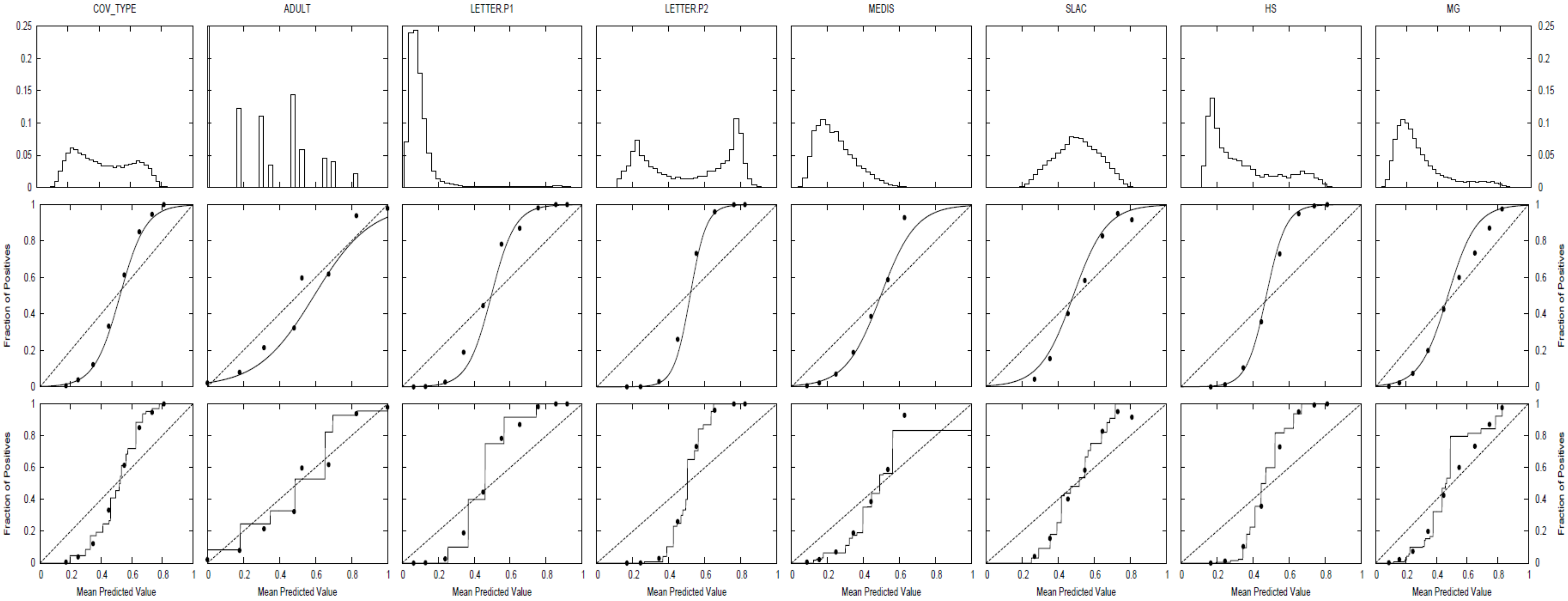


Figure 1. Histograms of predicted values and reliability diagrams for boosted decision trees.

Boosted Trees after Platt Calibration

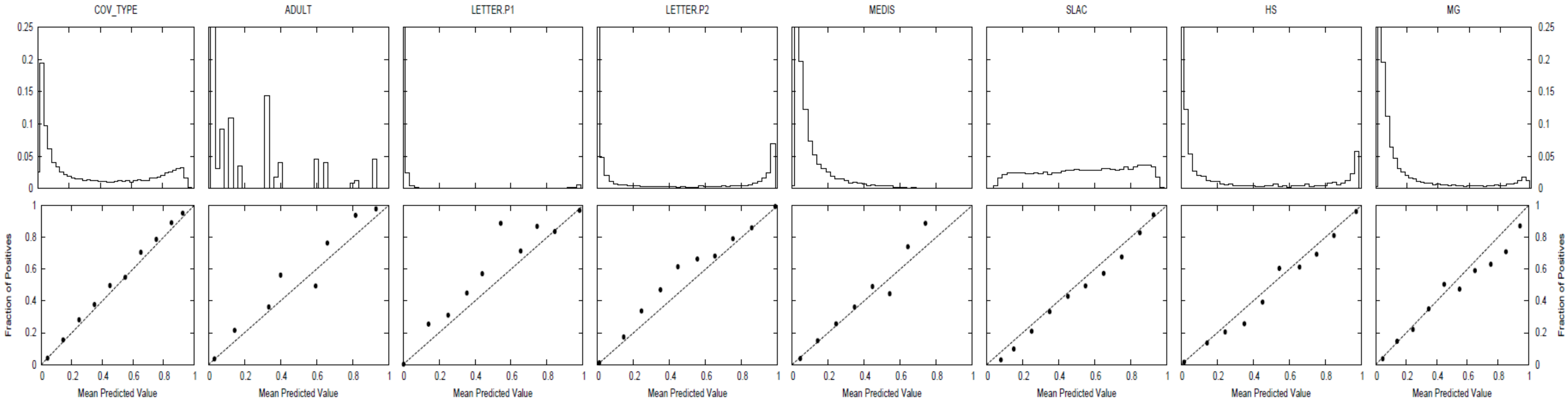


Figure 2. Histograms of predicted values and reliability diagrams for boosted trees calibrated with Platt's method.

Boosted Trees after Isotonic Regression Calibration

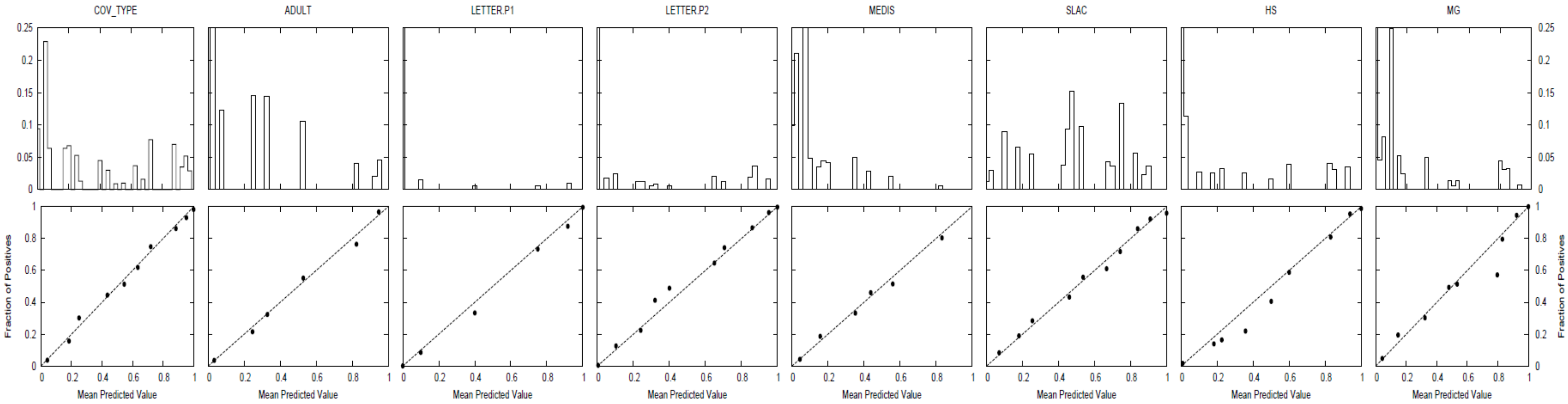


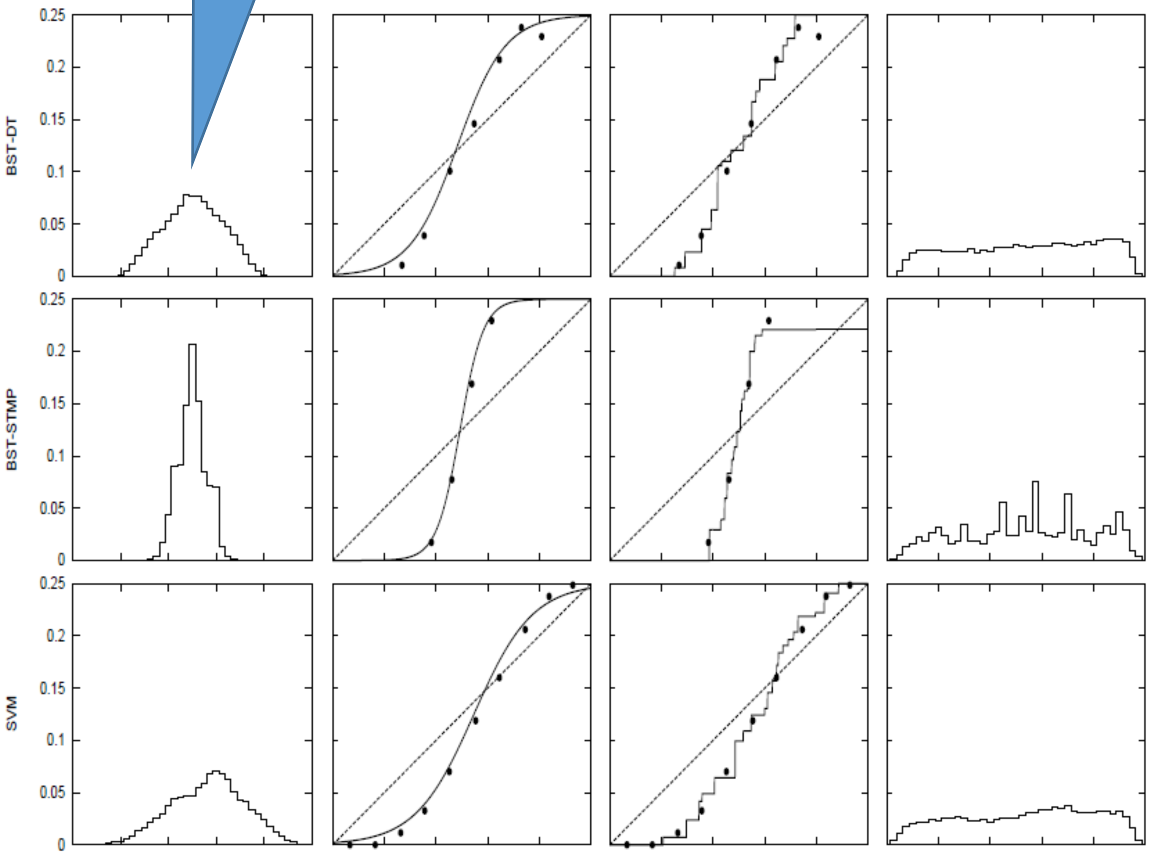
Figure 3. Histograms of predicted values and reliability diagrams for boosted trees calibrated with Isotonic Regression.

10 Different Learning Algorith

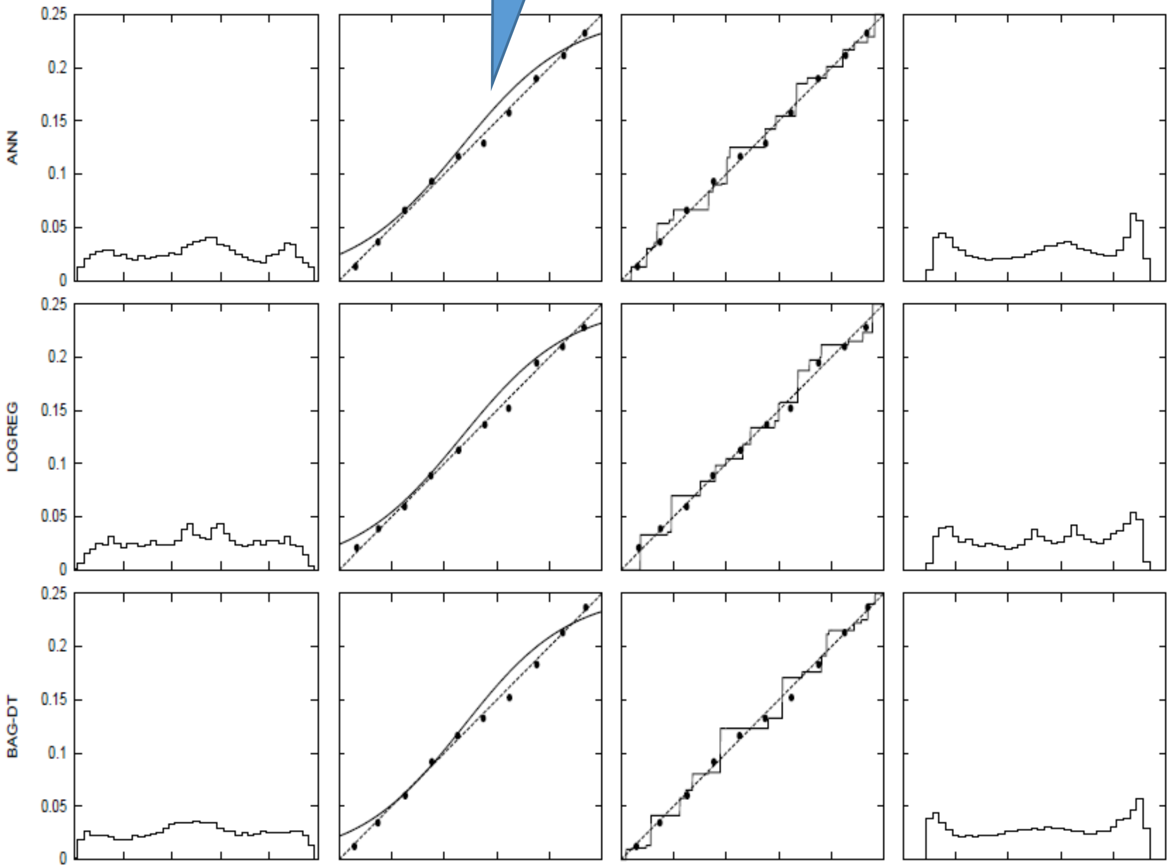
$\hat{p}$ concentrated in the middle

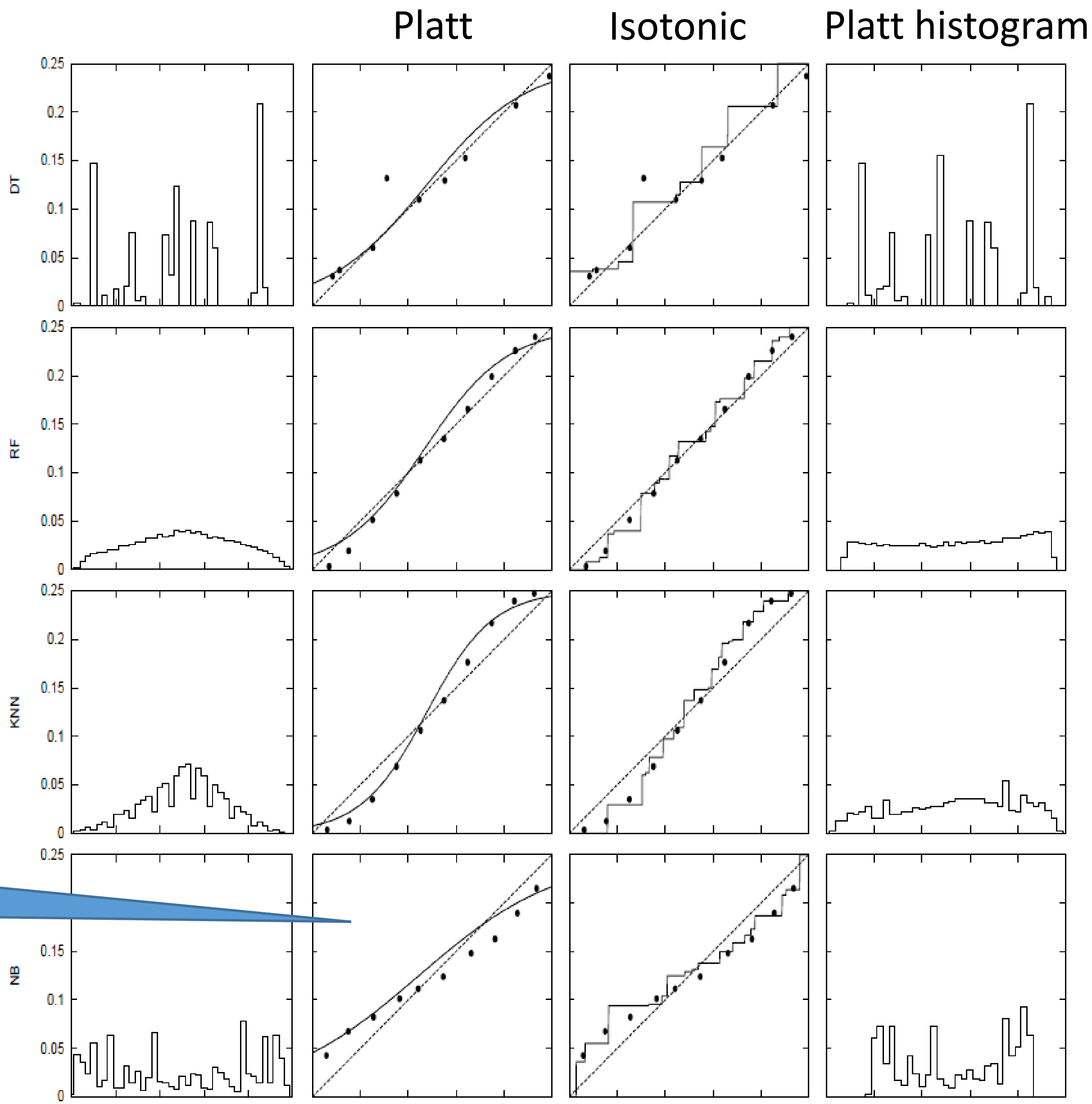
Already well-calibrated

Platt Isotonic Platt histogram



Platt Isotonic Platt histogram

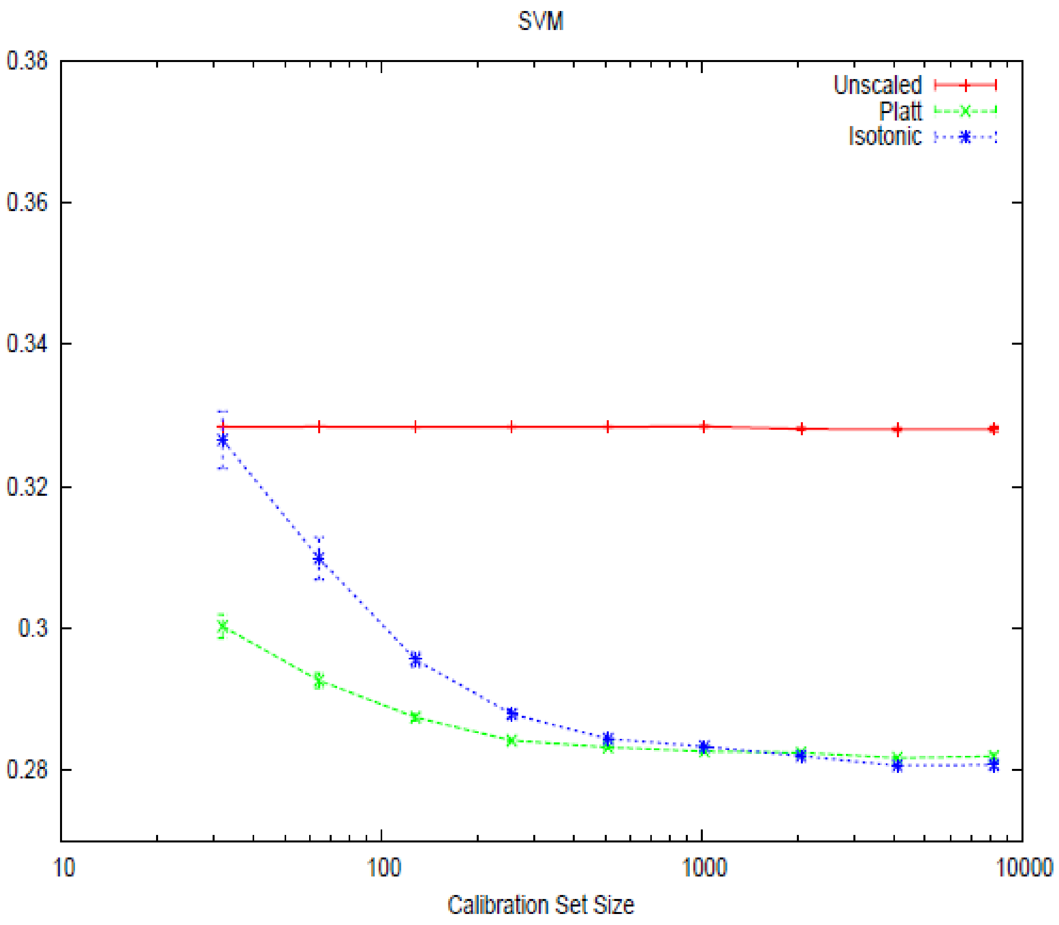




Sigmoid is not a good model for NB

Figure 6. Histograms and reliability diagrams for SLAC.

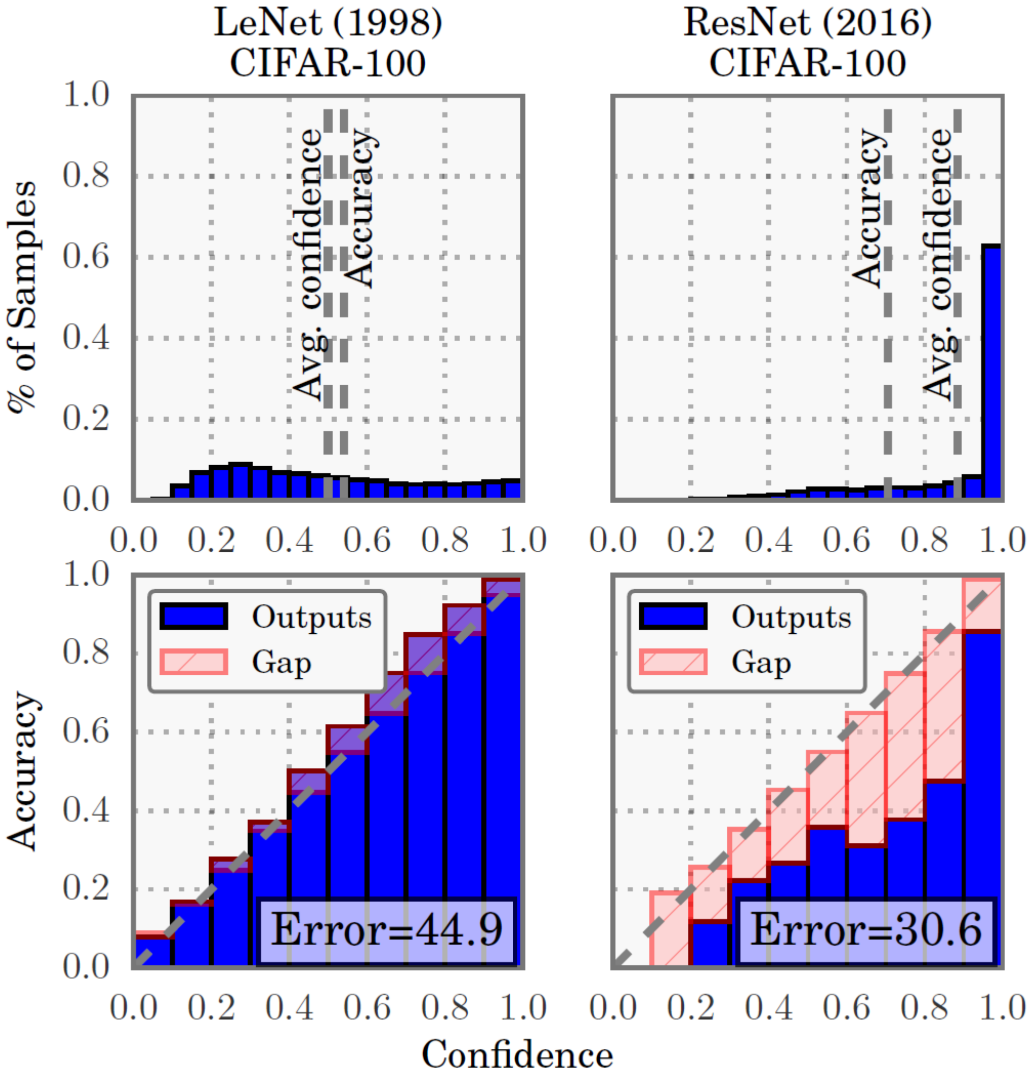
How big does the calibration set need to be?



Platt: 500; Isotonic: 8000?

Platt: 500; Isotonic: 4000

Experiments 2: Guo, Pleiss, Sun & Weinberger



ResNet is much more confident

ResNet over-confident!

What are the causes of bad calibration?

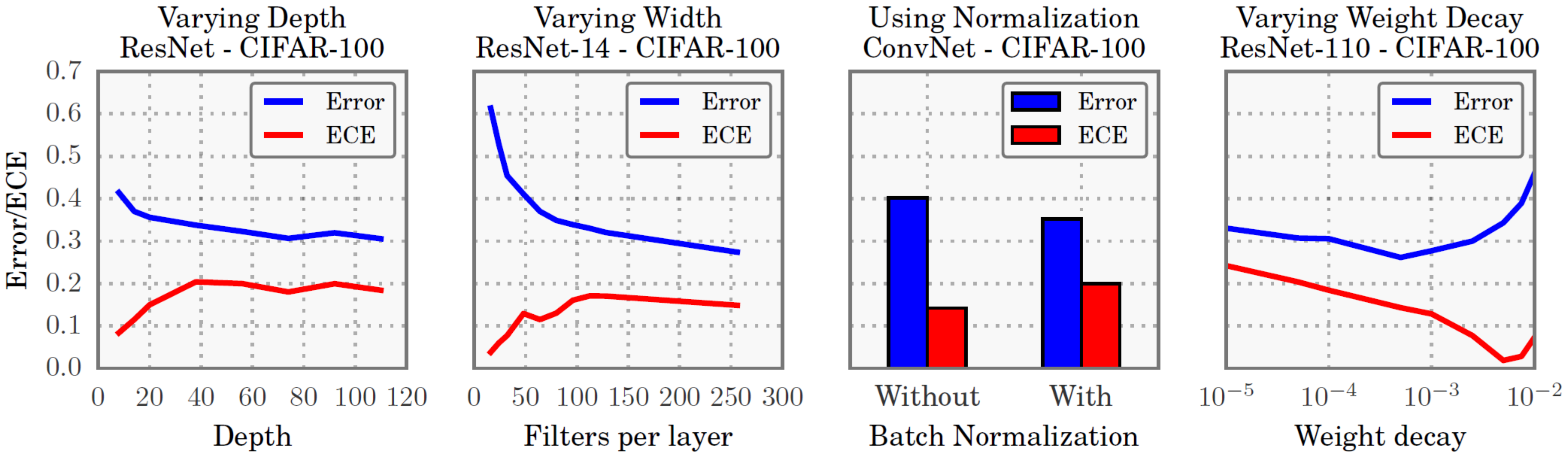
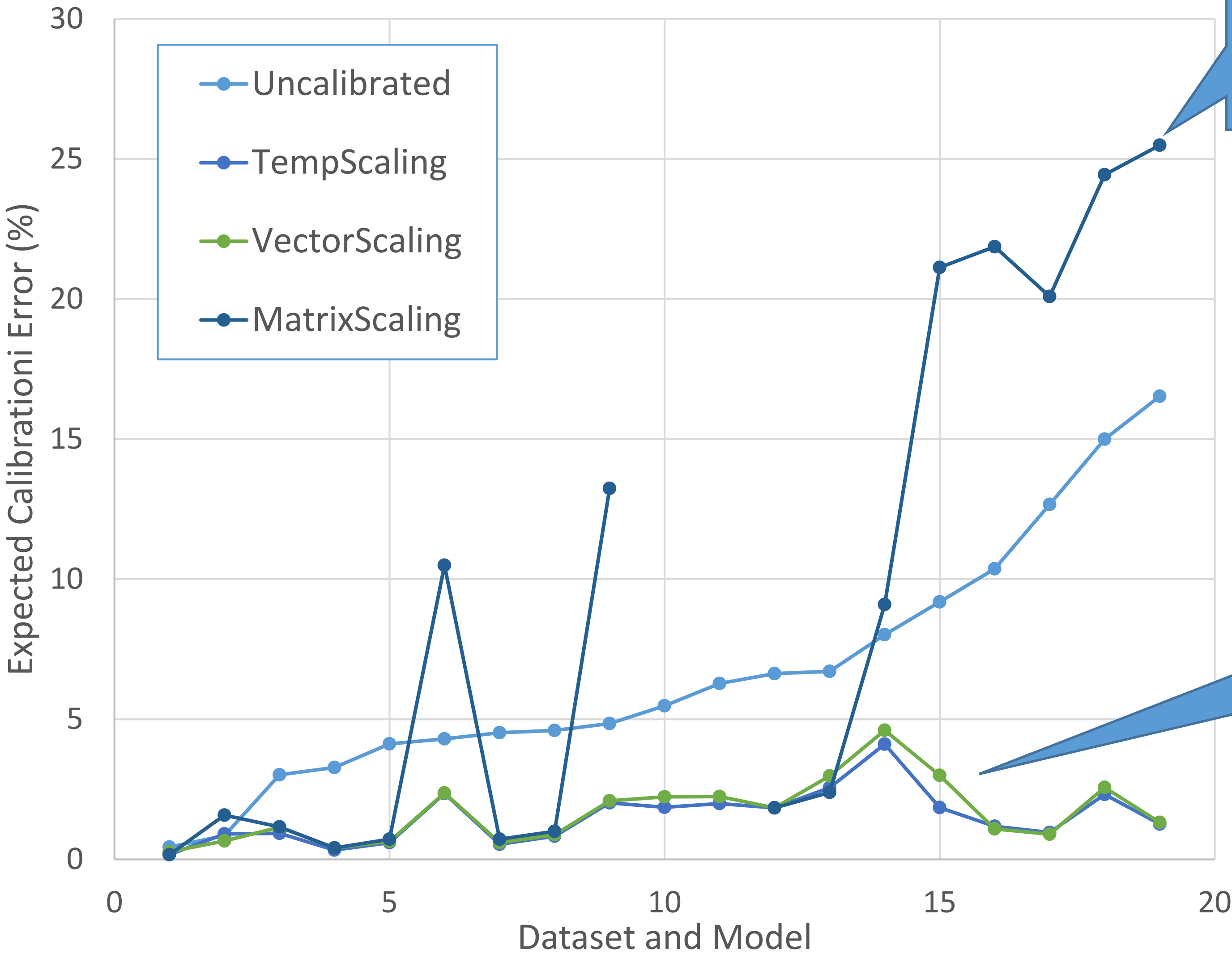


Figure 2. The effect of network depth (far left), width (middle left), Batch Normalization (middle right), and weight decay (far right) on miscalibration, as measured by ECE (lower is better).

Note: ECE = mean absolute calibration error $\sum_b \frac{|B_b|}{N} \left|\hat{P}(B_b) - \hat{p}(B_b)\right|$

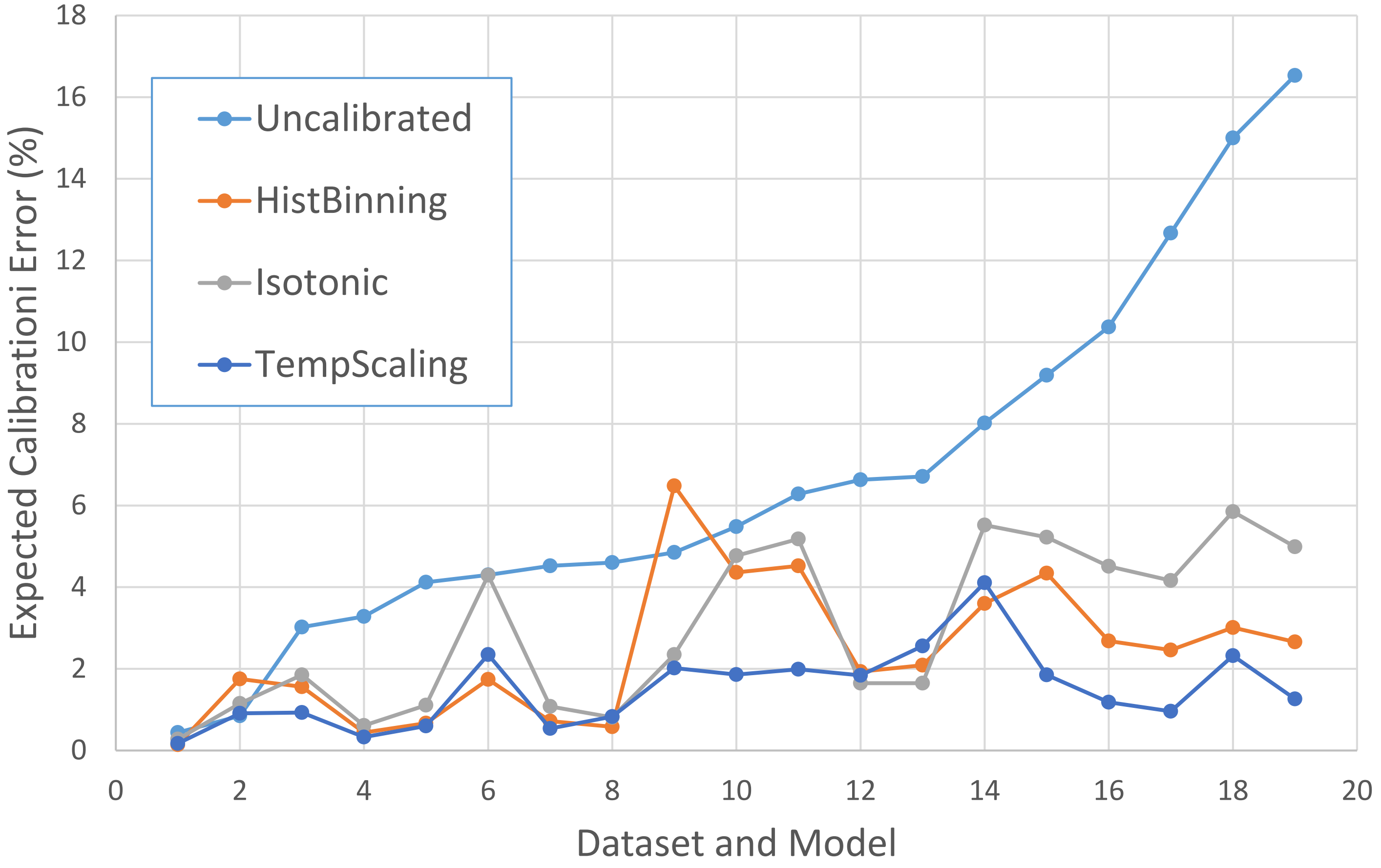
Comparison on Multiple Tasks and Architectures



Matrix Scaling is often very bad

Vector Scaling == Temp Scaling

Comparison against other methods



Insights and Questions

- The simple Temperature Calibration model works well and works better than more complex generalizations of Platt Scaling
- Temperature Calibration can be derived as the solution to a maximum entropy optimization problem
  - Maximize entropy of $\hat{P}$ subject to (a) $\hat{P}$ is a probability and (b) the sum of true class logits == mean value of all logits weighted by $\hat{P}$
  - Not clear by (b) makes sense
- Why didn’t they compare against Platt Scaling each class separately and then normalizing?

Experiment 3: Naeini & Cooper (ENIR)

- 40 UCI and LibSVM benchmark datasets
- Classifiers: Naïve Bayes, Logistic Regression, SVM
- Hyperparameters tuned via 10x10-fold cross-validation
- Calibration Algorithms:
  - Isotonic Regression (IsoRegC)
  - BBQ: Bayesian Quantile Binning (ensemble of quantile bin models)
  - ENIR: Ensemble of Near Isotonic Regressions

- Calibration reuses the training data
- No comparison against Platt scaling or other model-based approaches
- Metrics:
  - AUC = area under ROC curve
  - ACC = accuracy
  - RMSE = square root of the Calibration score
  - ECE = expected absolute calibration error
  - MCE = maximum absolute calibration error

Percentage Change

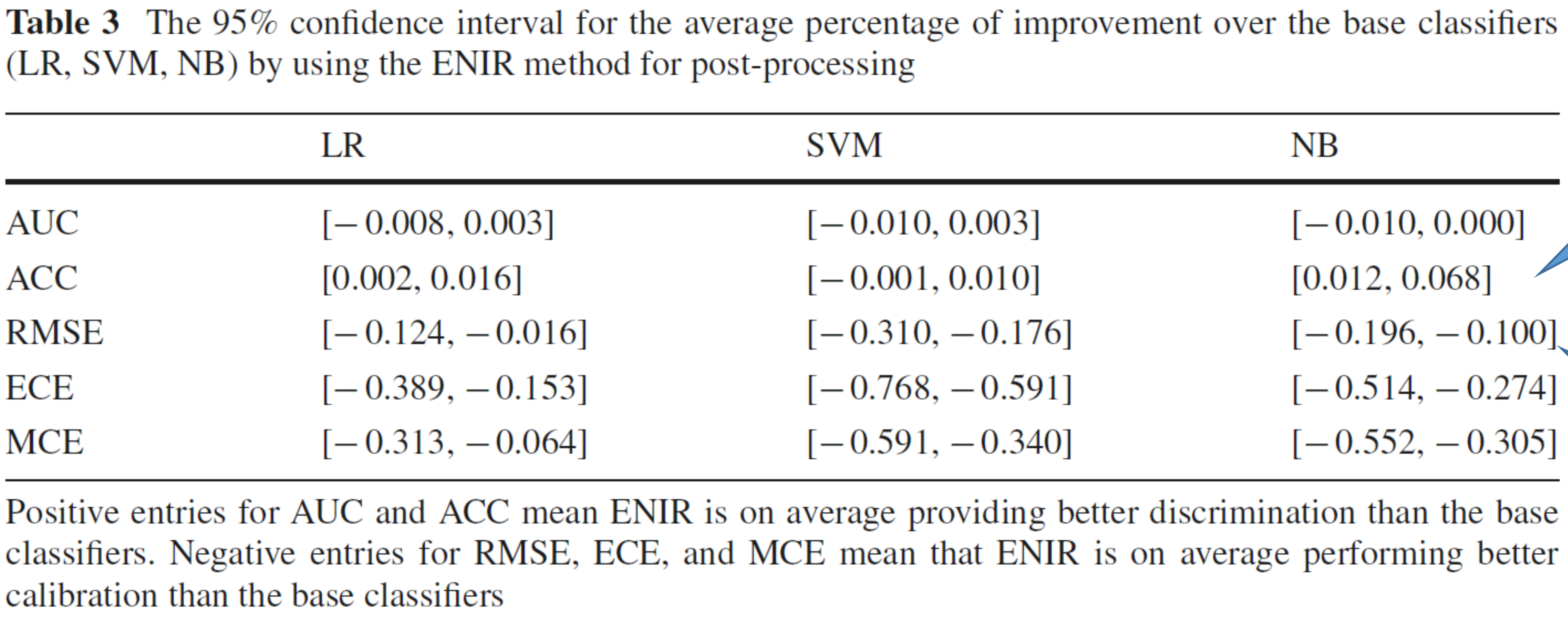


Table 3 The 95% confidence interval for the average percentage of improvement over the base classifiers (LR, SVM, NB) by using the ENIR method for post-processing

	LR	SVM	NB
AUC	[−0.008, 0.003]	[−0.010, 0.003]	[−0.010, 0.000]
ACC	[0.002, 0.016]	[−0.001, 0.010]	[0.012, 0.068]
RMSE	[−0.124, −0.016]	[−0.310, −0.176]	[−0.196, −0.100]
ECE	[−0.389, −0.153]	[−0.768, −0.591]	[−0.514, −0.274]
MCE	[−0.313, −0.064]	[−0.591, −0.340]	[−0.552, −0.305]

Positive entries for AUC and ACC mean ENIR is on average providing better discrimination than the base classifiers. Negative entries for RMSE, ECE, and MCE mean that ENIR is on average performing better calibration than the base classifiers

Naïve Bayes & LR ACC always improves

Calibration Metrics always improve

Accuracy improvements probably result from better thresholding

Logistic Regression

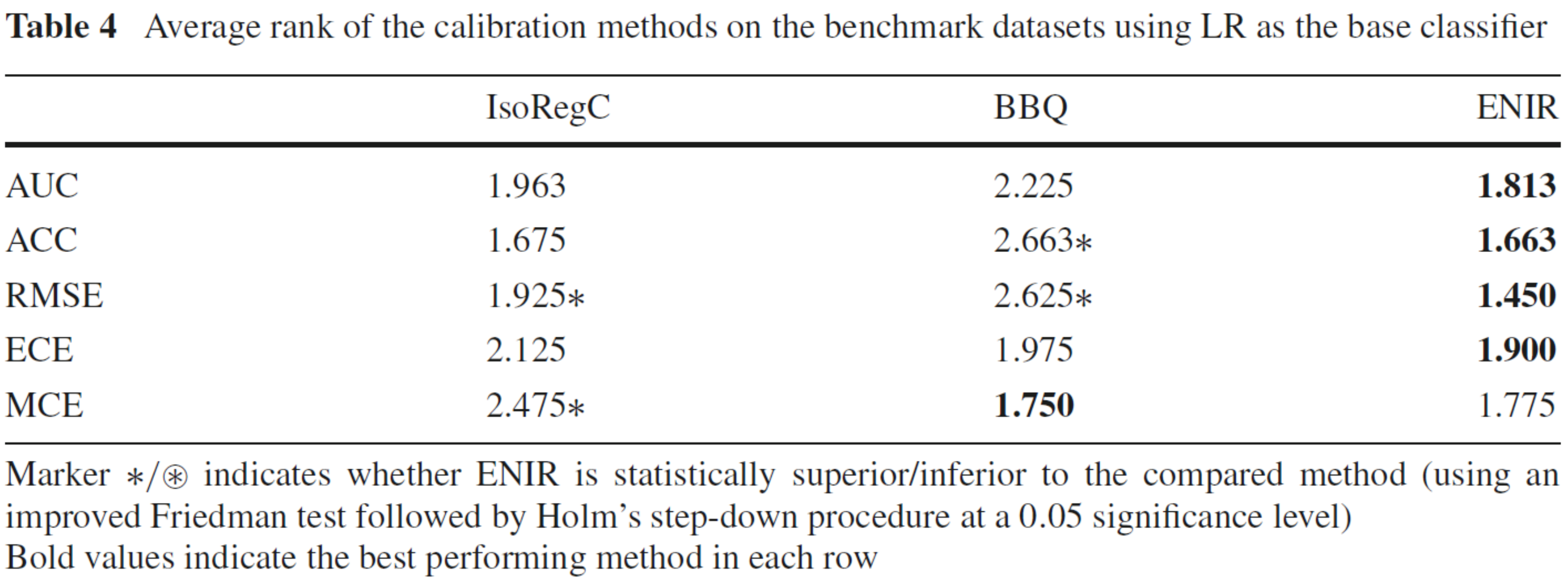


Table 4 Average rank of the calibration methods on the benchmark datasets using LR as the base classifier

	IsoRegC	BBQ	ENIR
AUC	1.963	2.225	**1.813**
ACC	1.675	2.663∗	**1.663**
RMSE	1.925∗	2.625∗	**1.450**
ECE	2.125	1.975	**1.900**
MCE	2.475∗	**1.750**	1.775

Marker ∗/⊛ indicates whether ENIR is statistically superior/inferior to the compared method (using an improved Friedman test followed by Holm's step-down procedure at a 0.05 significance level)
Bold values indicate the best performing method in each row

No significant change in AUC, ACC, ECE
Significant change in RMSE
Trend looks good
AUC, ACC should not improve unless calibration curve is non-monotonic

SVMs

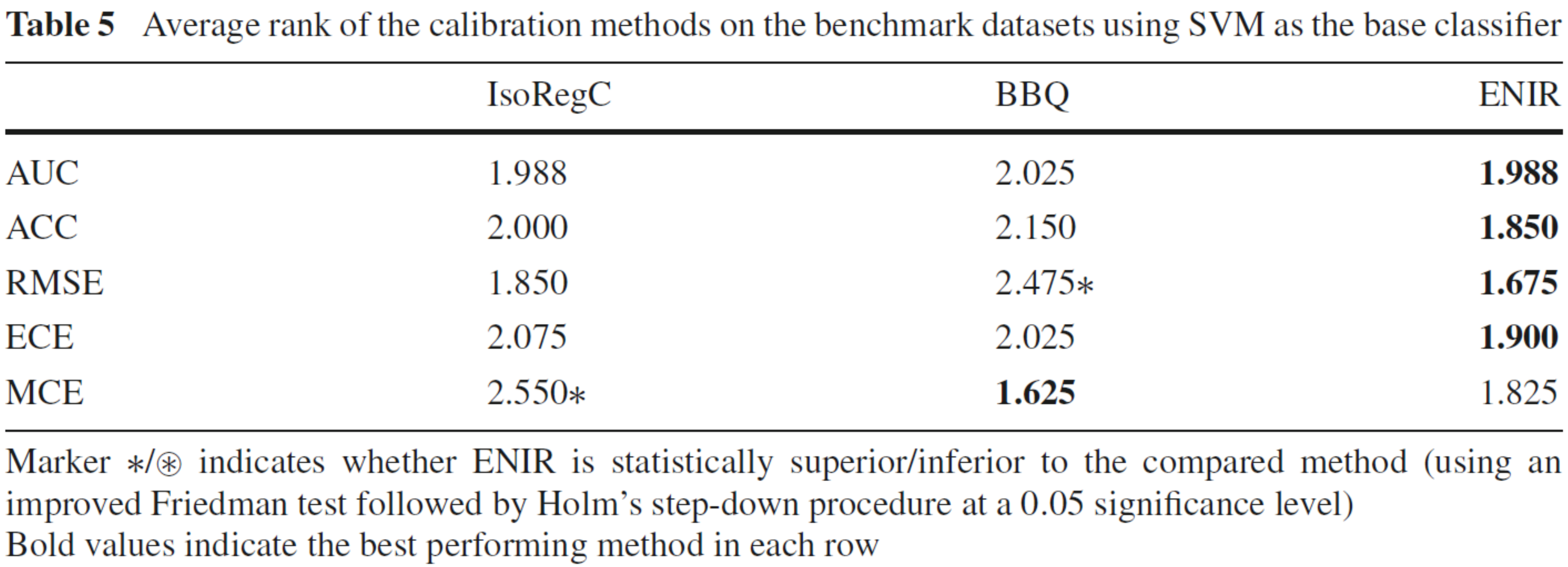


Table 5 Average rank of the calibration methods on the benchmark datasets using SVM as the base classifier

	IsoRegC	BBQ	ENIR
AUC	1.988	2.025	**1.988**
ACC	2.000	2.150	**1.850**
RMSE	1.850	2.475∗	**1.675**
ECE	2.075	2.025	**1.900**
MCE	2.550∗	**1.625**	1.825

Marker ∗/⊛ indicates whether ENIR is statistically superior/inferior to the compared method (using an improved Friedman test followed by Holm's step-down procedure at a 0.05 significance level)
Bold values indicate the best performing method in each row

No significant change in AUC, ACC, RMSE, ECE

Naïve Bayes

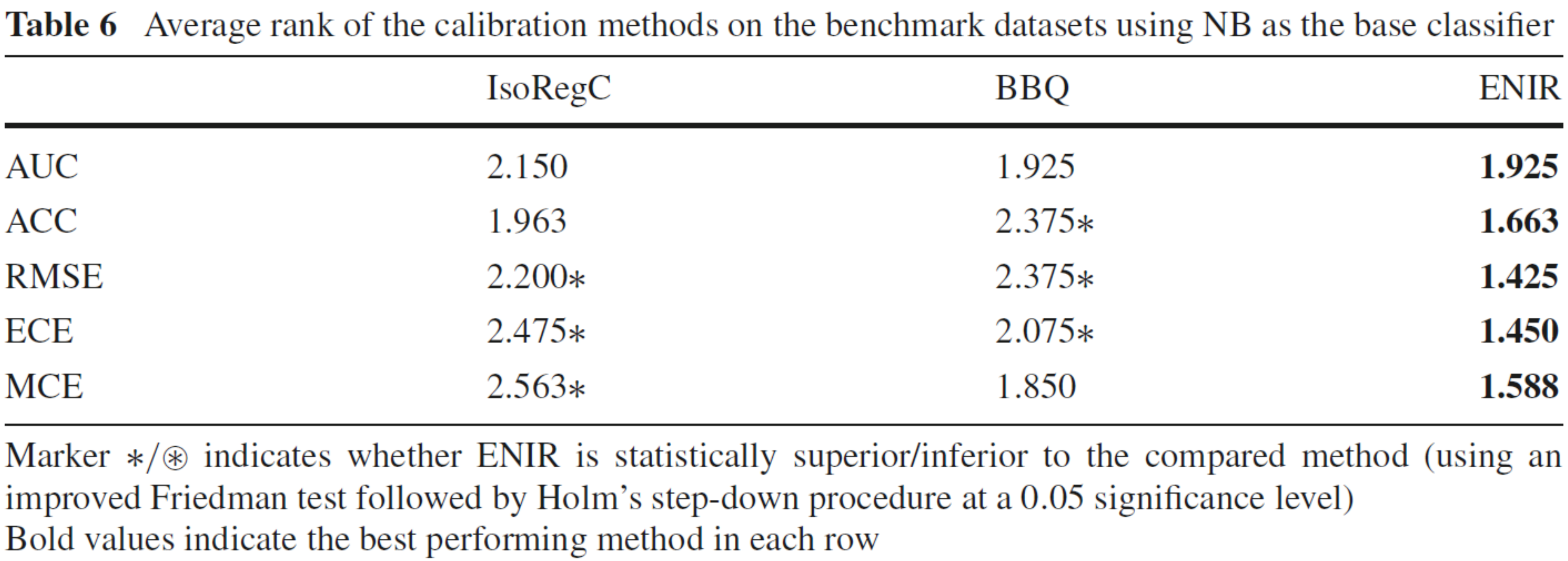


Table 6 Average rank of the calibration methods on the benchmark datasets using NB as the base classifier

	IsoRegC	BBQ	ENIR
AUC	2.150	1.925	**1.925**
ACC	1.963	2.375∗	**1.663**
RMSE	2.200∗	2.375∗	**1.425**
ECE	2.475∗	2.075∗	**1.450**
MCE	2.563∗	1.850	**1.588**

Marker ∗/⊛ indicates whether ENIR is statistically superior/inferior to the compared method (using an improved Friedman test followed by Holm's step-down procedure at a 0.05 significance level)
Bold values indicate the best performing method in each row

No significant change in AUC or ACC (surprisingly!)
Significant improvements in all calibration metrics (not surprisingly)

Insights and Questions

- Using a regularized version of Isotonic Regression does not improve accuracy or AUC compared to regular Isotonic Regression
- But it does improve measures of calibration

- The main advantage of regularizing should be to reduce the amount of calibration data that is needed, but the authors did not study this question

Summary of Miscalibration Behaviors

- Max Margin Methods (SVM, boosted trees):
  - $\hat{p}$ concentrates near 0.5
  - Sigmoid-shaped Reliability Diagram
  - Platt (logistic regression) model fits well, learns quickly
- Naïve Bayes and Deep Nets
  - $\hat{p}$ concentrates near 0 and 1; systematically optimistic
  - Sigmoid model fits NB poorly; Isotonic regression is better
  - Temperature Calibration worked better for Deep Nets
- Random Forests, Bagging, MLPs
  - Naturally well-calibrated except at extreme probabilities
  - Sigmoid model fits poorly
  - Need lots of calibration data to obtain any improvements

Closing Thoughts

- Do we care equally about all parts of the $\hat{p}$ space?
- For high-confidence predictions
  - We only care about large values of $\hat{p}$
- For anomaly detection
  - We only care about very small values of $\hat{p}$
- For stock market trading
  - We care about values of $\hat{p} = 0.5 + \epsilon$

Do we need to calibrate, or can we just threshold?

- That is the subject for Friday!
- Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. Journal of Machine Learning Research, 9, 371–421. Retrieved from http://arxiv.org/abs/0706.3188
- Cortes, C., DeSalvo, G., & Mohri, M. (2016). Learning with rejection. *Lecture Notes in Artificial Intelligence*, *9925 LNAI*, 67–82. http://doi.org/10.1007/978-3-319-46379-7_5
- Papadopoulos, H. (2008). Inductive Conformal Prediction: Theory and Application to Neural Networks. Book chapter. https://www.researchgate.net/publication/221787122_Inductive_Conformal_Prediction_Theory_and_Application_to_Neural_Networks

References

- Ayer, M., Brunk, H., Ewing, G., Reid, W., Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, 26(4):641–647, 1955. (PAV Algorithm)
- Cohen, I., Goldszmidt, M. (2004). Properties and benefits of calibrated classifiers. HP Labs Report HPL-2004-22.
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. *ArXiv*. Retrieved from http://arxiv.org/abs/1706.04599
- Lucena, B. (2018) Spline-Based Probability Calibration. https://export.arxiv.org/abs/1809.07751
- Leathart, T., Frank. E., Holmes, G., Pfahringer, B. (2017). Probability Calibration Trees. *Asian Conference on Machine Learning (2017).* JMLR: Workshop and Conference Proceedings 77:145-160.
- Naeini, Cooper, G. (2018). Binary classifier calibration using an ensemble of piecewise linear regression models. *Knowledge and Information Systems, 54(1)*: 151-170. https://arxiv.org/pdf/1511.05191.pdf
- Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning ICML '05*, (2005), 625–632. http://doi.org/10.1145/1102351.1102430
- Platt, J. (1999). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *Advances in Large Margin Classifiers* (pp. 61–74).
- Robertson, T., Wright, F., Dykstra, R. (1988). *Order Restricted Statistical Inference*, chapter 1. John Wiley & Sons, 1988.
- Song, H., Kull, M., Flach, P. (2018). Non-Parametric Calibration of Probabilistic Regression. https://arxiv.org/abs/1806.07690
- Tishirani, R. J., Hoefling, H., Tibshirani, R., (2011). Nearly isotonic regression. *Technometrics, 53(1)*:54–61, 2011
- Zadrozny, B., & Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naïve Bayesian classifiers. *ICML* (pp. 609–616).
- Zadrozny, B., Elkan, C. (2002). Transforming Classifier Scores into Accurate Multiclass Probability Estimates. *KDD* (pp. 694–699).