Probabilistic Prediction Calibration using Brier Score

Zhenfeng Lin

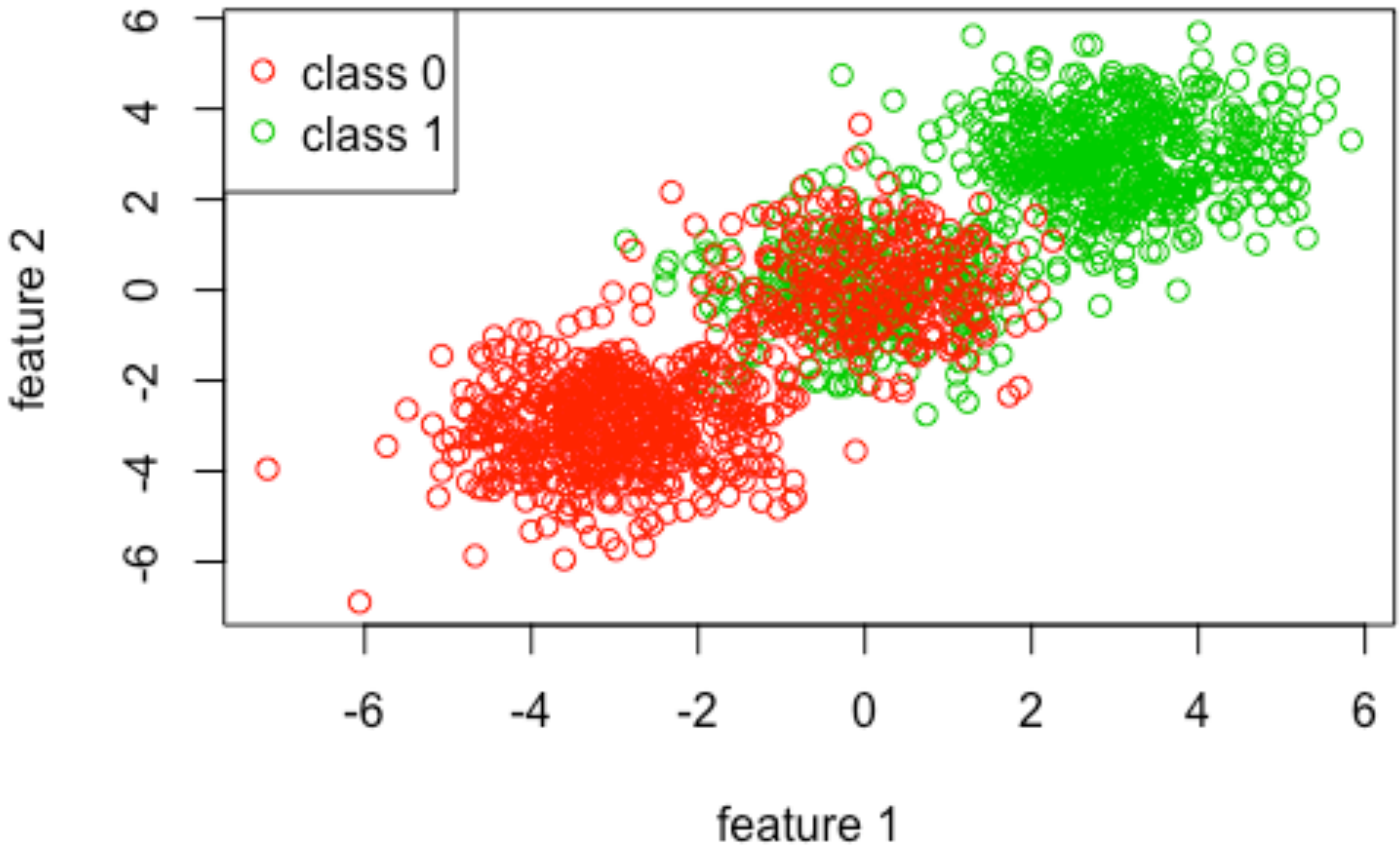
11/30/2016

Texas A&M University

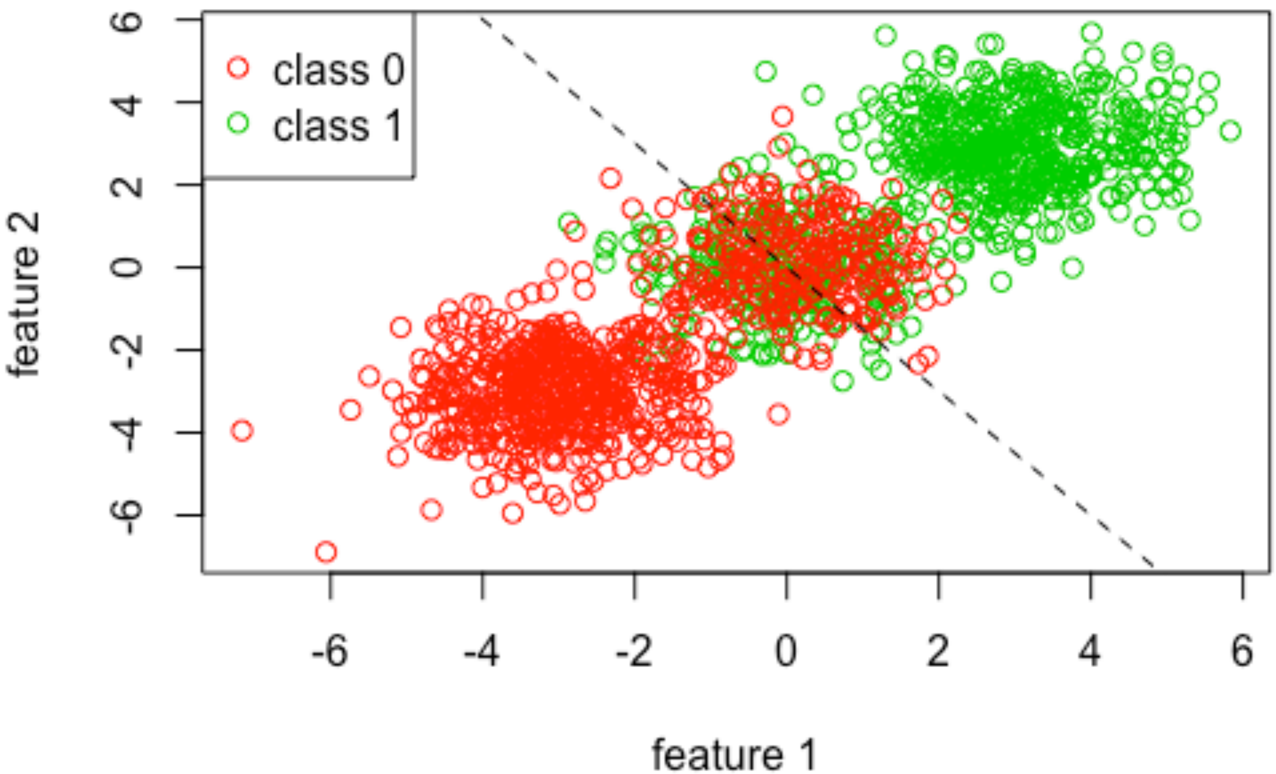
Outline

- Motivation
- Notation
- Brier Score Decomposition
- Reliability Diagram
- Decision Rule (threshold)
- Multi-thresholds
- Transformation

Motivation



Use Naïve Bayes Classifier:



predicted	actual 0	actual 1
0	619	131
1	131	619

Classification rate ≈ 0.83

Notation

Using the notation from [1]

- Feature/sample X
- Response/Label/Class $C \in \{0, 1\}$
- Probabilistic prediction $p(C \mid X)$
- Brier score

$$BS = \frac{1}{n} \sum_{i=1}^{n} \left(p(C \mid x_i) - c_i \right)^2$$

- Note: Brier's (1950) [2] original definition

$$BS = \frac{1}{n} \sum_{i=1}^{n} \sum_{c_i=1}^{C} \left(p(C \mid x_i) - c_i \right)^2$$

Brier score's two-component decomposition

Given a predicted probability t

- Set of samples yield t

$$R_t = \left\{ x_i : p(C = 1 \mid x_i) = t \right\}$$

- Frequency at t

$$\pi_t = \# R_t / n$$

- Observed probability at t

$$p(c \mid t) := p(C = 1 \mid t) = \frac{1}{\# R_t} \sum_{x_i \in R_t} I(c_i = 1)$$

Then Brier score can be rewritten as [3,4]

$$
\begin{aligned}
BS &= \int_0^1 \pi_t \left[p(c \mid t)(t-1)^2 + \left(1 - p(c \mid t)\right) t^2 \right] dt \\
&= \underbrace{\int_0^1 \pi_t \left(t - p(c \mid t)\right)^2 dt}_{\text{Calibration}} + \underbrace{\int_0^1 \pi_t p(c \mid t)\left(1 - p(c \mid t)\right) dt}_{\text{Refinement}}
\end{aligned}
$$

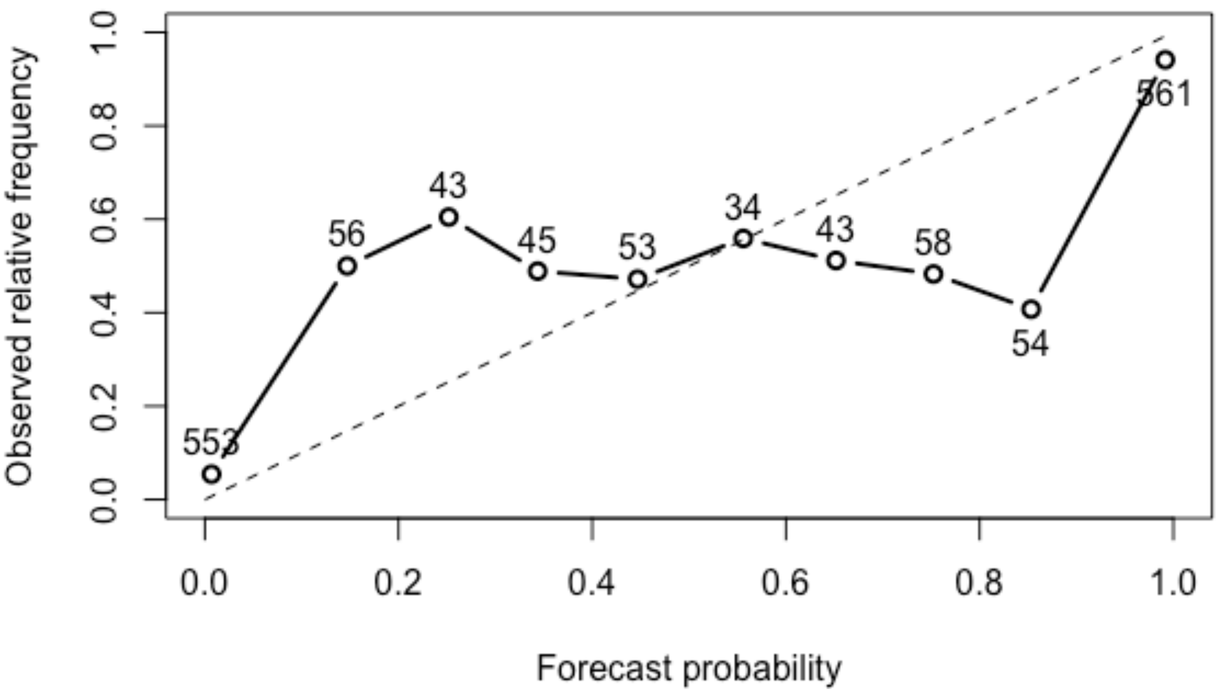
- "Calibration"(a.k.a. "Reliability") term indicates how close is the assessment to the frequency in reality.
- "Refinement" term scores the usefulness of each forecast.
- Note: MSE = Bias^2 + Var

- Discrete version

$$BS = \frac{1}{n}\sum_{k=1}^{K} n_k (t_k - o_k)^2 + \frac{1}{n}\sum_{k=1}^{K} n_k o_k (1 - o_k)$$

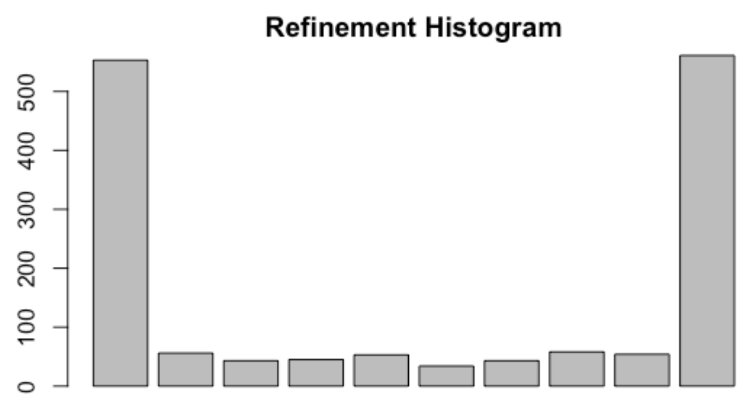
where we partition $[0,1]$ into K bins, and within *k*-th bin, n_k is the number of predictions, t_k is usually midpoint of the bin, o_k is the observed relative frequency.

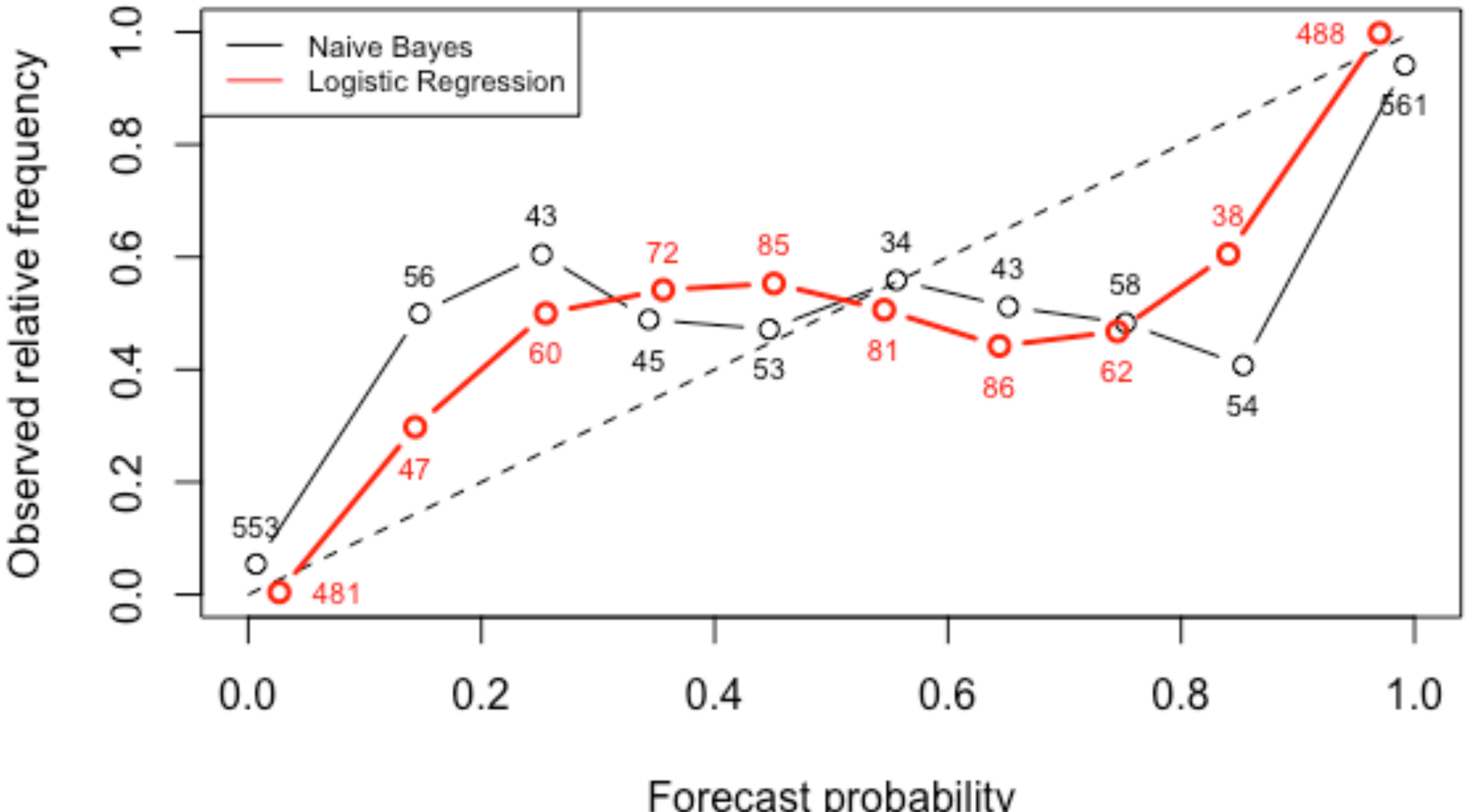
Reliability Diagram



More diagonal, the better

More extreme, the better





Naive Bayes
Logistic Regression
Observed relative frequency
Forecast probability
0.0
0.2
0.4
0.6
0.8
1.0
553
481
56
47
43
60
45
72
53
85
34
81
43
86
58
62
38
54
488
561

Decision Rule (threshold)

- Simply take threshold α

$$\hat{c} = \begin{cases} 0, & \text{if } t \le \alpha; \\ 1, & \text{o.w.} \end{cases}$$

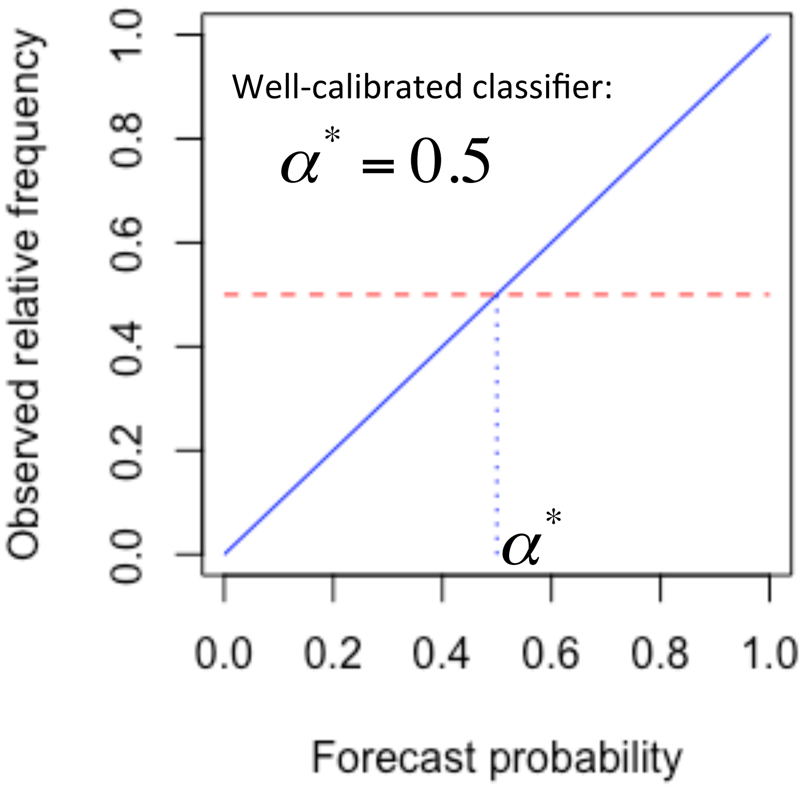
and usually take $\alpha = 0.5$

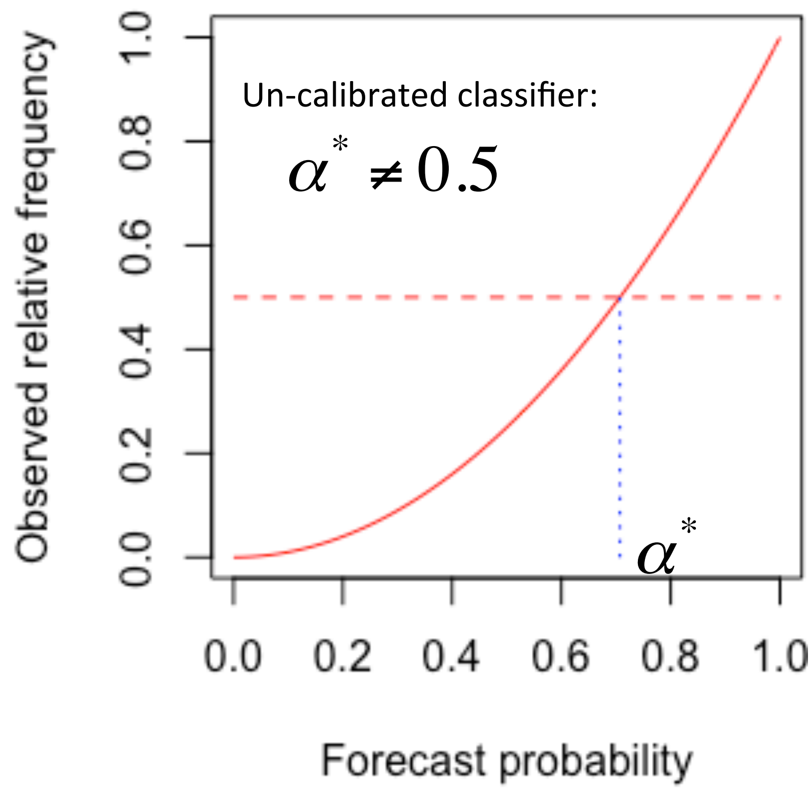
- Classification error

$$P_{error}(\alpha) = \int_0^\alpha p(c \mid t)\pi(t)\,dt + \int_\alpha^1 \big(1 - p(c \mid t)\big)\pi(t)\,dt$$

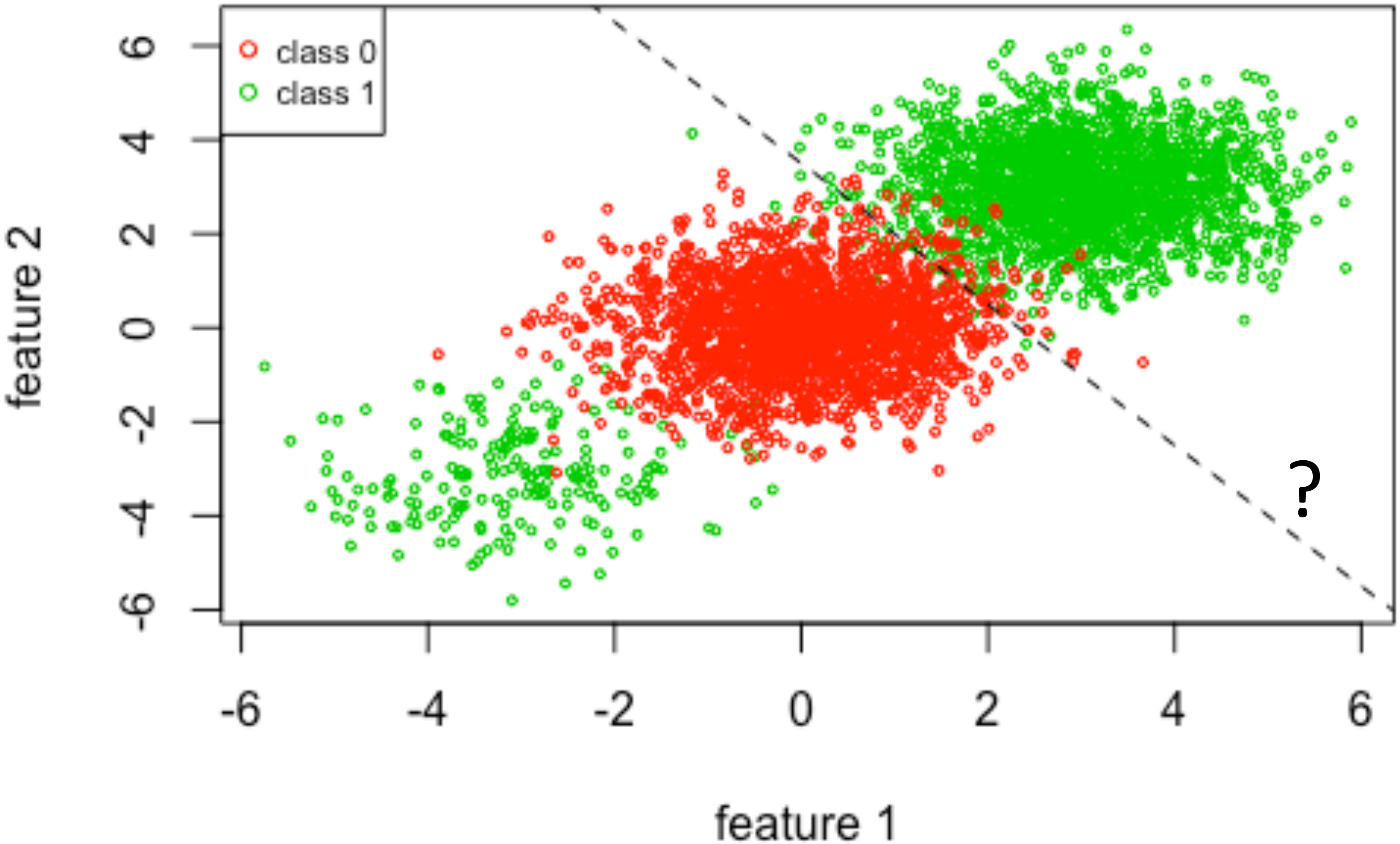
then optimal threshold $\alpha^* = \arg\min_\alpha P_{error}(\alpha)$ s.t.

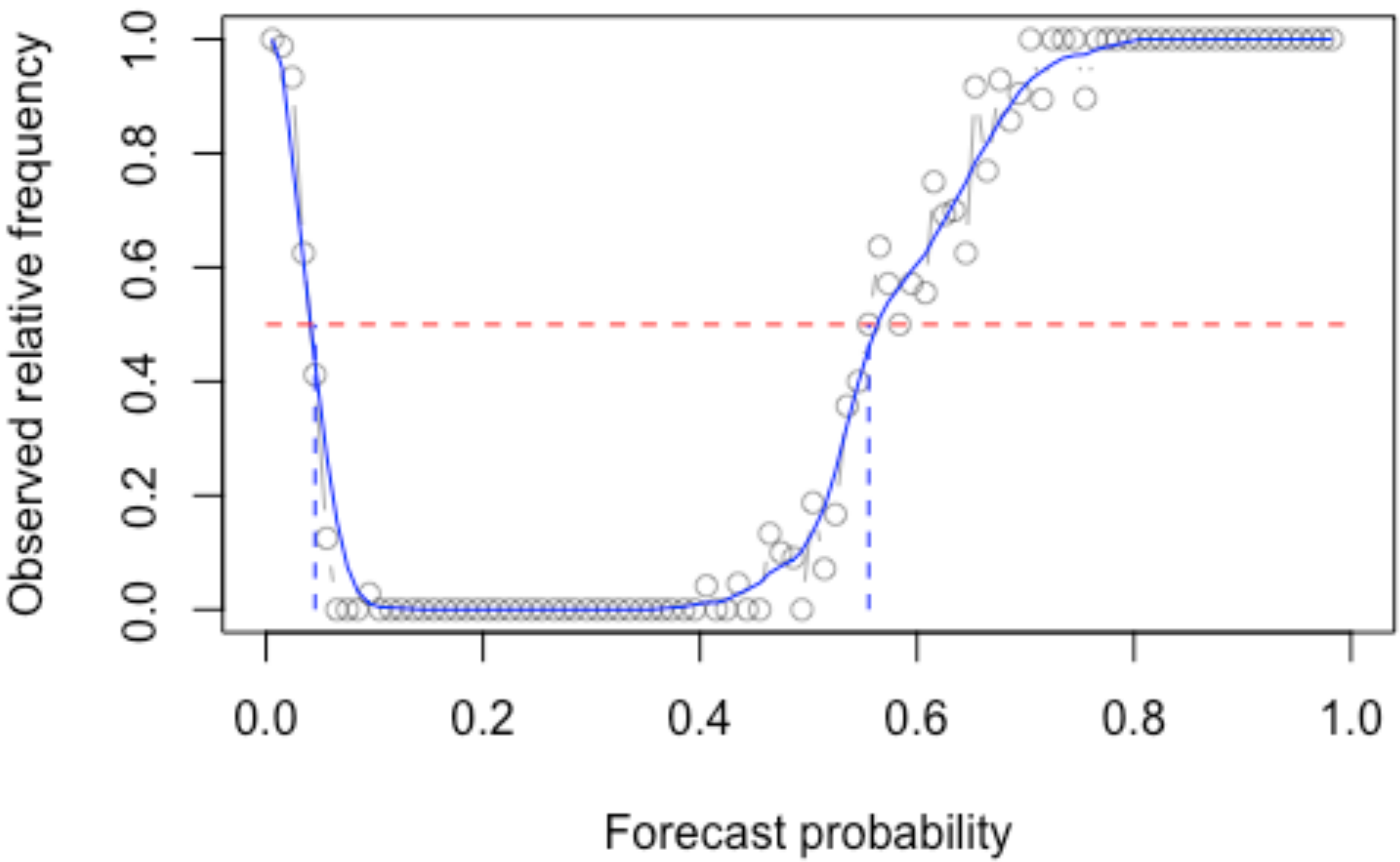
$$p(c \mid \alpha^*) = 0.5$$

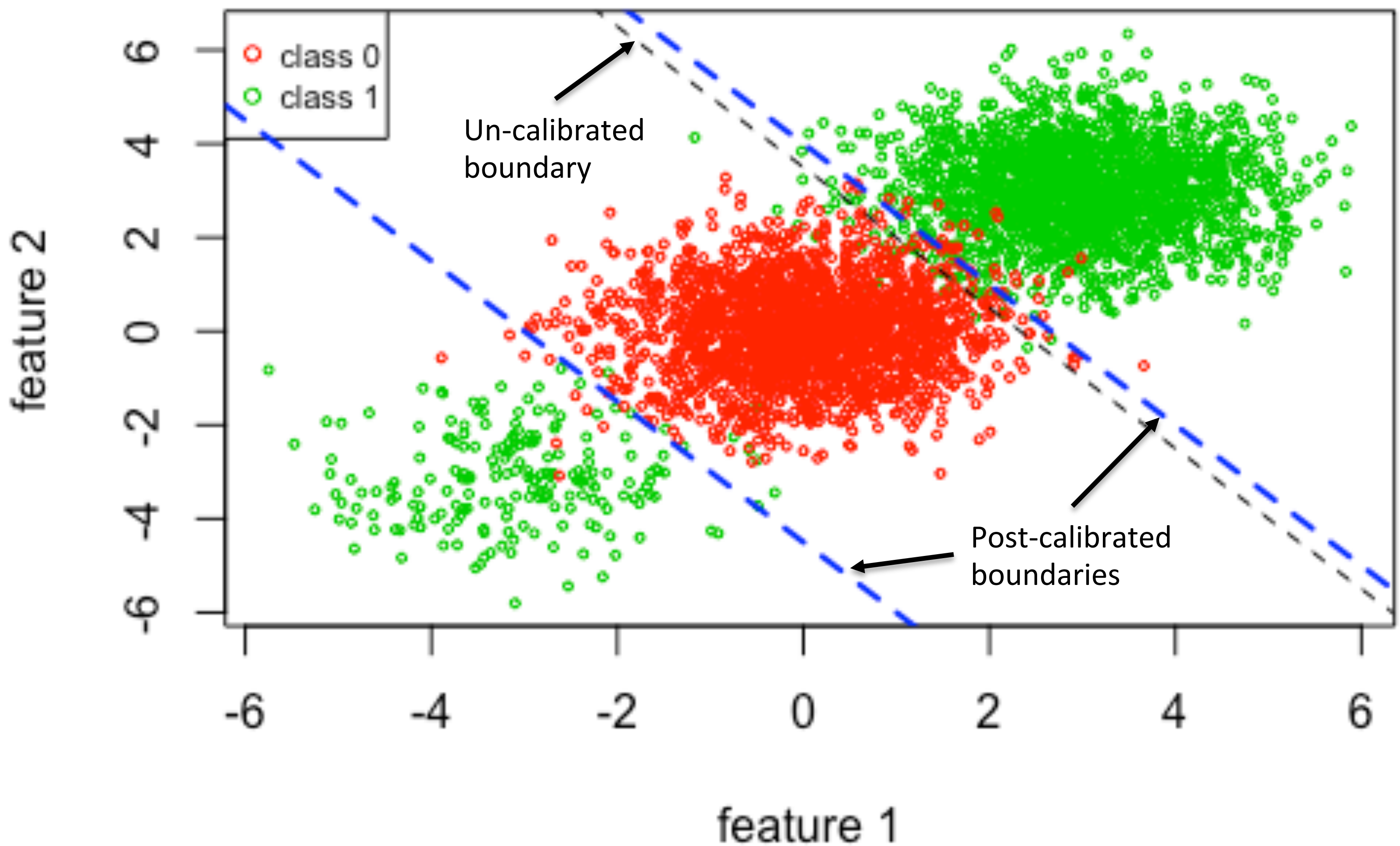

Well-calibrated classifier:
$\alpha^* = 0.5$
α^*
Observed relative frequency
0.0
0.2
0.4
0.6
0.8
1.0
Forecast probability


Un-calibrated classifier:
$\alpha^* \neq 0.5$
α^*
Observed relative frequency
0.0
0.2
0.4
0.6
0.8
1.0
Forecast probability

Multi-thresholds




Observed relative frequency
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Forecast probability


class 0
class 1
Un-calibrated
boundary
Post-calibrated
boundaries
feature 1
feature 2
-6
-4
-2
0
2
4
6
-6
-4
-2
0
2
4
6

Un-calibrated boundary:

	actual	
predicted	0	1
0	1895	207
1	105	1993

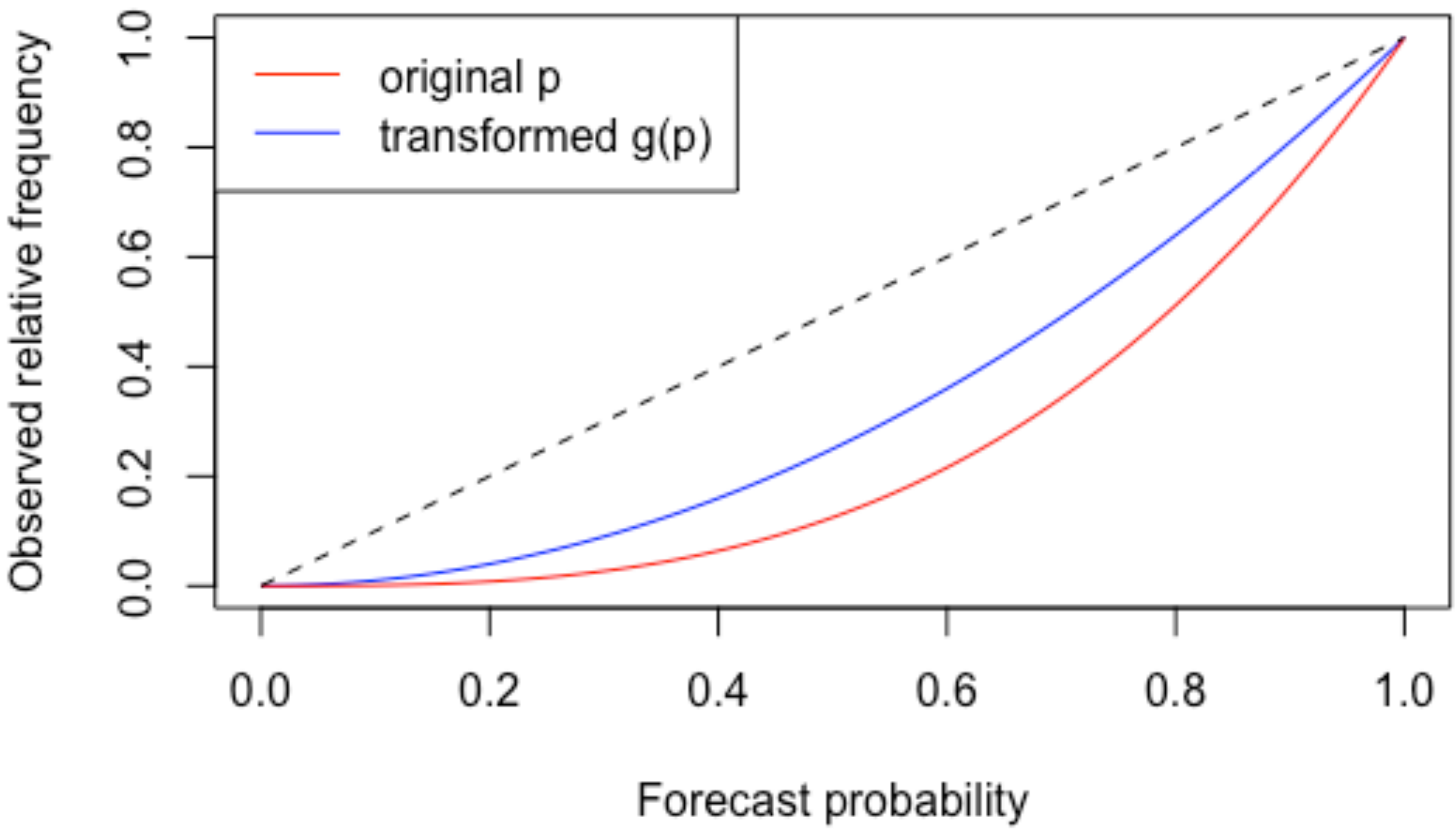
Classification rate ≈ 0.93

Post-calibrated boundaries:

	actual	
predicted	0	1
0	1939	31
1	61	2169

Classification rate ≈ 0.98

Transformation



- Platt scaling:
  - a method Platt (1999) [5] used to transform SVM outputs from $[-\infty,+\infty]$ to posterior probabilities.
  - It's particularly effective for max-margin methods.
- Isotonic regression:
  - a method Zadrozny and Elkan (2001, 2002) [7,8] used to calibrate predictions from Naïve Bayes, SVM and decision tree models.
  - Niculescu-Mizil etc (2005) [6] showed that it works better than Platt scaling.

Platt scaling, essentially, is a sigmoid transformation

$$g(p \mid a, b) = \frac{1}{1 + \exp(ap + b)}$$

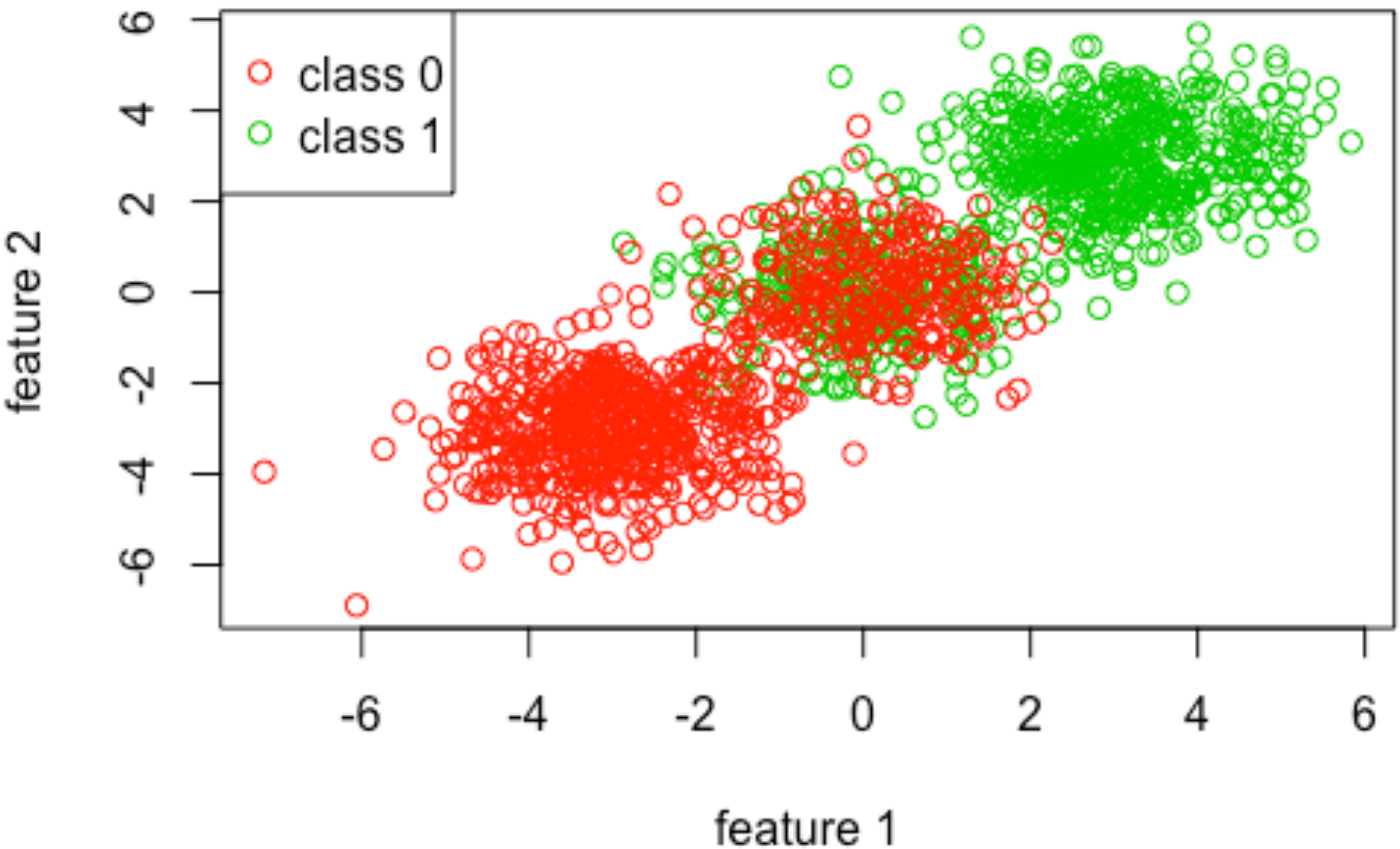
- It can be implemented via logistic regression
- To avoid over-fitting, some of training data are reserved to learn parameters of $g(p \mid a, b)$

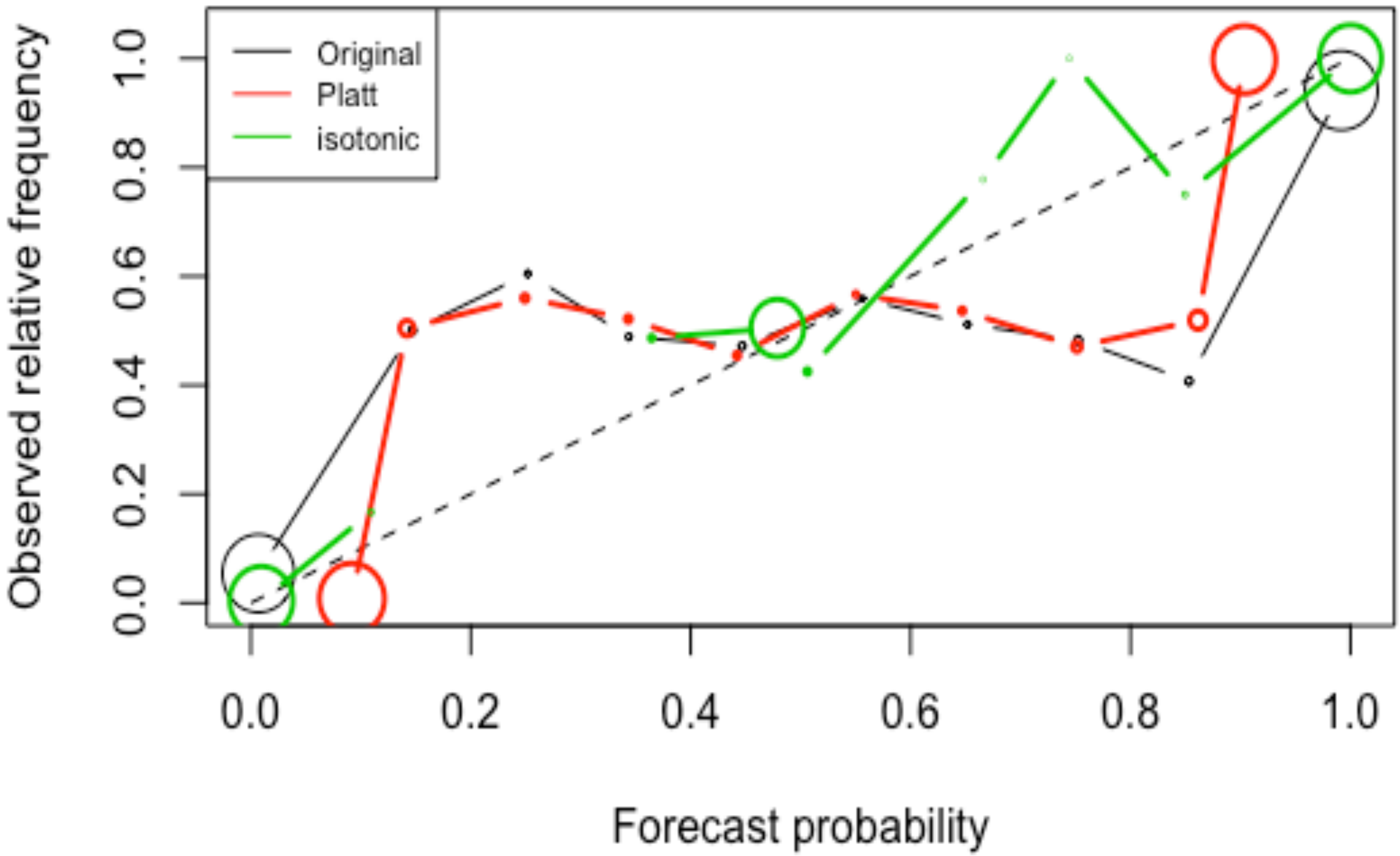
Isotonic regression solves problem: given a sequence of data points $y_1, \ldots, y_n$, how to best summarize this by a monotone sequence $\beta_1, \ldots, \beta_n$

Formally,

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_i)^2 \quad \text{subject to } \beta_1 \leq \cdots \leq \beta_n$$

- Unique solution exists, which can be obtained by pool adjacent violators algorithm (PAVA)
- If skillfully programmed, PAVA is $O(n)$

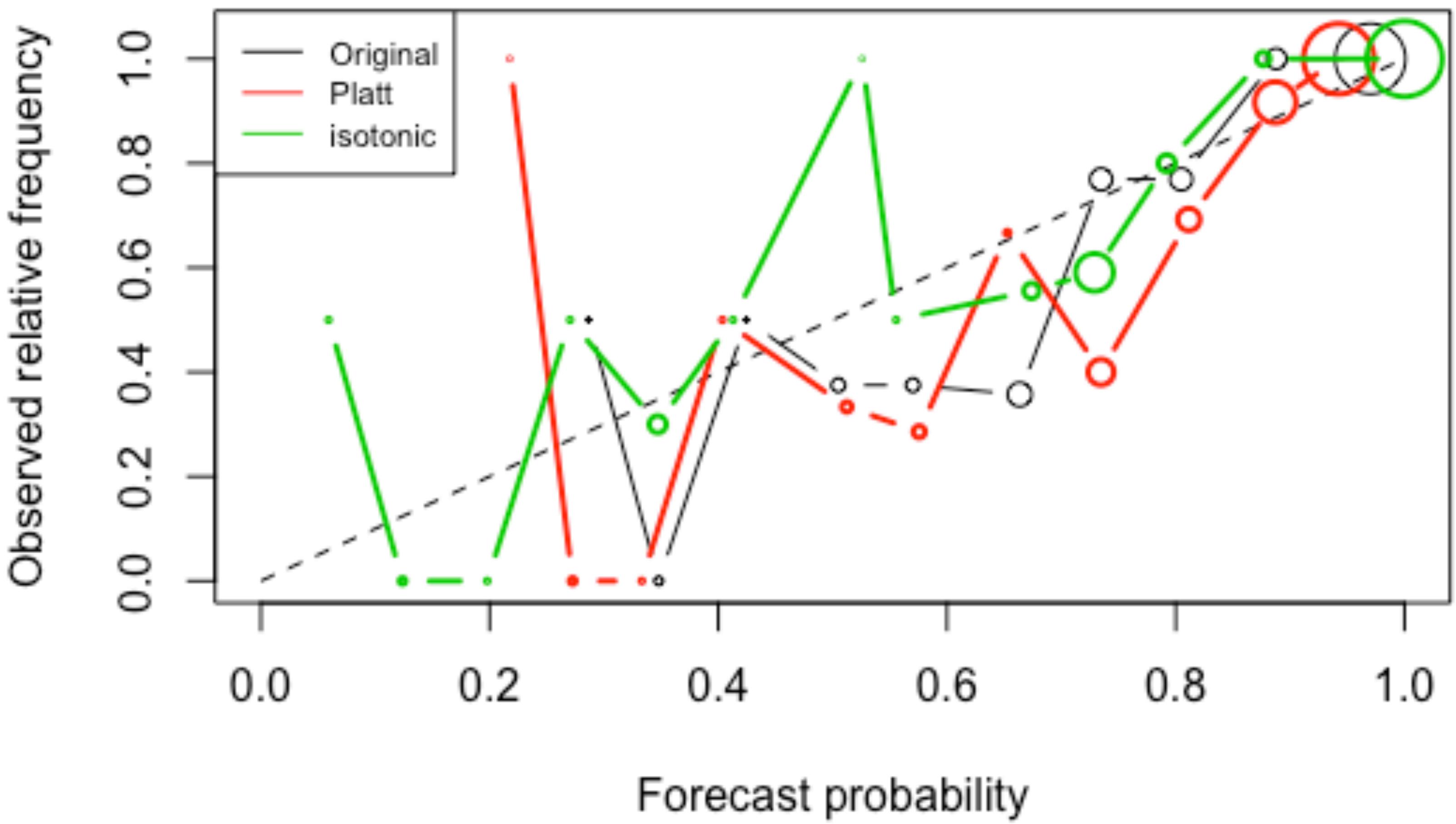

class 0
class 1
feature 2
feature 1
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6


Original
Platt
isotonic
Observed relative frequency
0.0
0.2
0.4
0.6
0.8
1.0
Forecast probability

A data from Faraway's paper (2016) [9]:

	Source	Blazar	CV
Training	CRTS	124	458
Testing	CSS	32	86

- Linear discriminant analysis (lda) is used


Original
Platt
isotonic
Observed relative frequency
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Forecast probability

Additional topics

- Multi-class case [2,8,10]
- More scoring rules [11,12]
- Bayesian binning [13]: a new transformation
- Brier curve [14]

Reference

[1] Ira Cohen and Moises Goldszmidt. (2004) Properties and benefits of calibrated classifiers. Springer. 3202 125–136.

[2] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78, 1–3.

[3] DeGroot, M., Fienberg, S. (1983) The comparison and evaluation of forecasters. The statistician 32 12-22.

[4] Murphy, A. H. (1973). A new vector partition of the probability score. Journal of Applied Meteorology. 12 (4): 595–600.

[5] Platt, J. (1999). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. Advances in Large Margin Classifiers. 61–74.

[6] Niculescu-Mizil, A., and Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of the International Conference on Machine Learning , 625–632.

[7] Zadrozny, B., & Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naïve bayesian classifiers. ICML. 609–616.

[8] Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. KDD. 694–699.

[9] Faraway J. etc. (2016) Modeling lightcurves for improved classification. JSADM. 9 1-11.

[10] Hamill, T.M., 1997: Reliability Diagrams for Multicategory Probabilistic Forecasts. Wea. Forecasting, 12, 736-741.

[11] Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102, 359–378.

[12] Richmond V. etc. (2008) Scoring Rules, Generalized Entropy, and Utility Maximization. OR. 56 1146-1157.

[13] Naeini, M.P. etc (2015) Obtaining Well Calibrated Probabilities Using Bayesian Binning. Proc Conf AAAI Artif Intell. 2901–2907.

[14] Hernandez-Orallo, J. etc (2011). "Brier curves: a new cost-based visualisation of classifier performance" (PDF). Proceedings of the 28th International Conference on Machine Learning (ICML-11). 585–592.

감사합니다 Natick

Grazie Danke Ευχαριστίες Dalu Obrigado

Thank You Köszönöm

Tack

Спасибо Dank Gracias

谢谢 Merci Seé

ありがとう