SCHOOL OF DATA ANALYSIS

NATIONAL RESEARCH UNIVERSITY

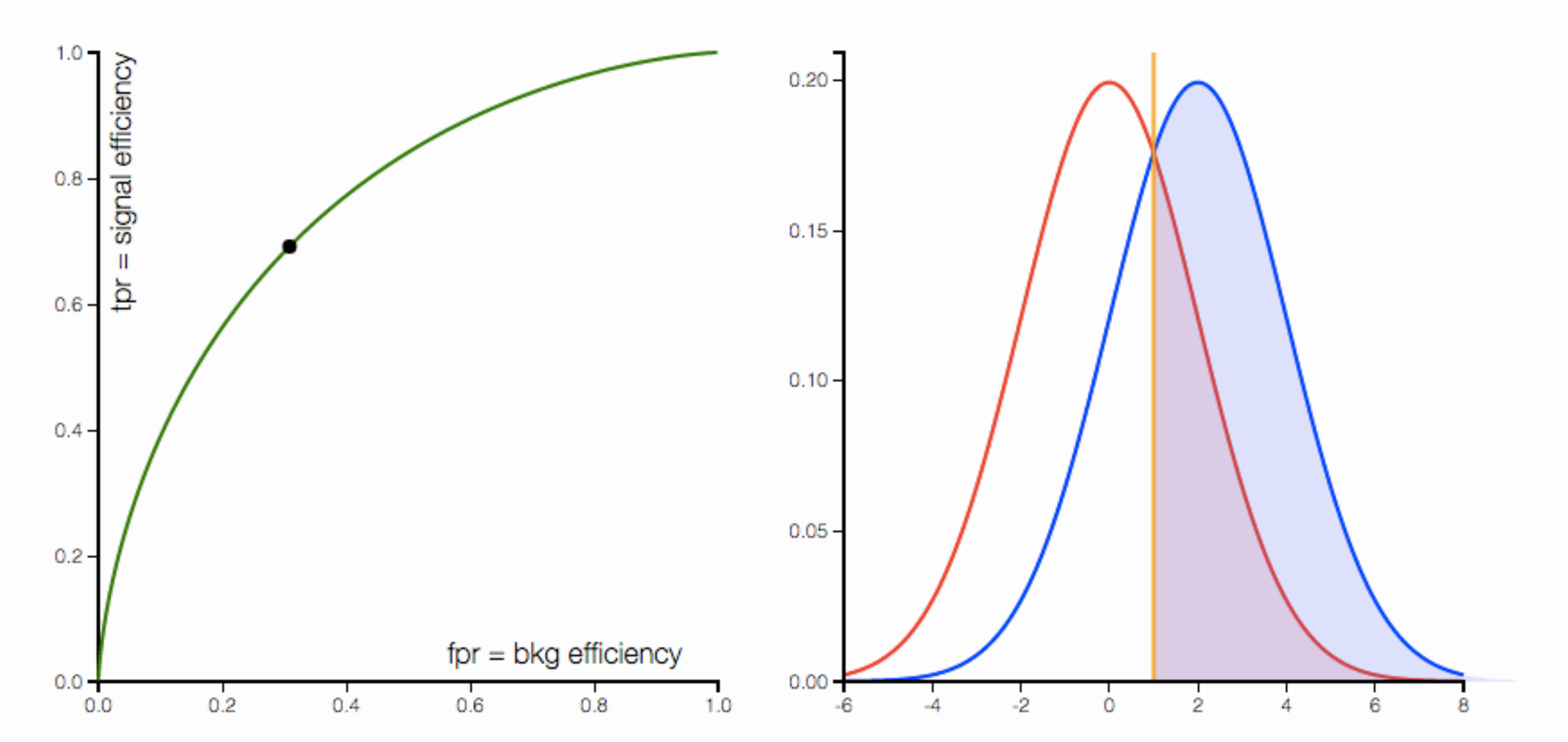
Classifier output calibration

YSDA, NRU HSE

Tatiana Likhomanenko

Introduction

ROC curve: output ranking



Applications

In the following areas we need to obtain probabilities for samples:

- science (e.g., determining which experiments to perform)
- medicine (e.g., deciding which therapy to give a patient)
- business (e.g., making investment decisions)
- weather forecasting
- game theory
- ad click prediction
- HEP

HEP applications

- Probability estimation for some physics processes requires true probabilities
- Combine information from different parts of the event within probabilistic model
- Probabilities are easier to manipulate

Probabilistic classifier is

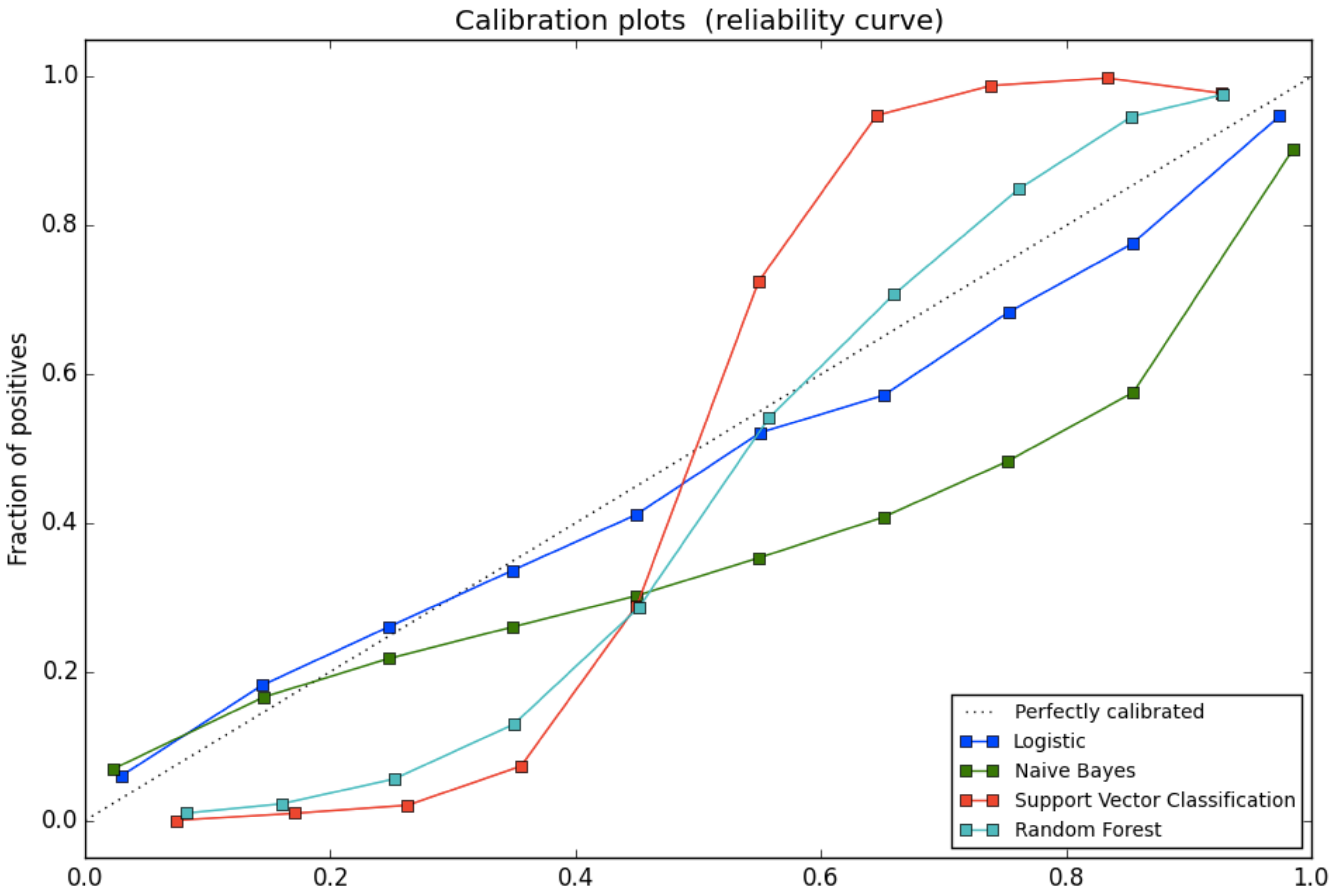
- predicts not only outputting the most likely class that the sample should belong to;
- is able to predict a probability distribution over a set of classes $P(\mathbf{y}|x)$;
- provides classification with a degree of certainty, which can be useful in its own right, or when combining classifiers into ensembles.

Which model is a probabilistic classifier?

- Naive Bayes, Logistic Regression and Multilayer Perceptrons (when trained under an appropriate loss function) are naturally probabilistic.
- Other models such as Support Vector Machines are not.
- Decision Trees and Boosting methods produce distorted class probability distributions [1].
- There are methods to turn them into probabilistic classifiers.

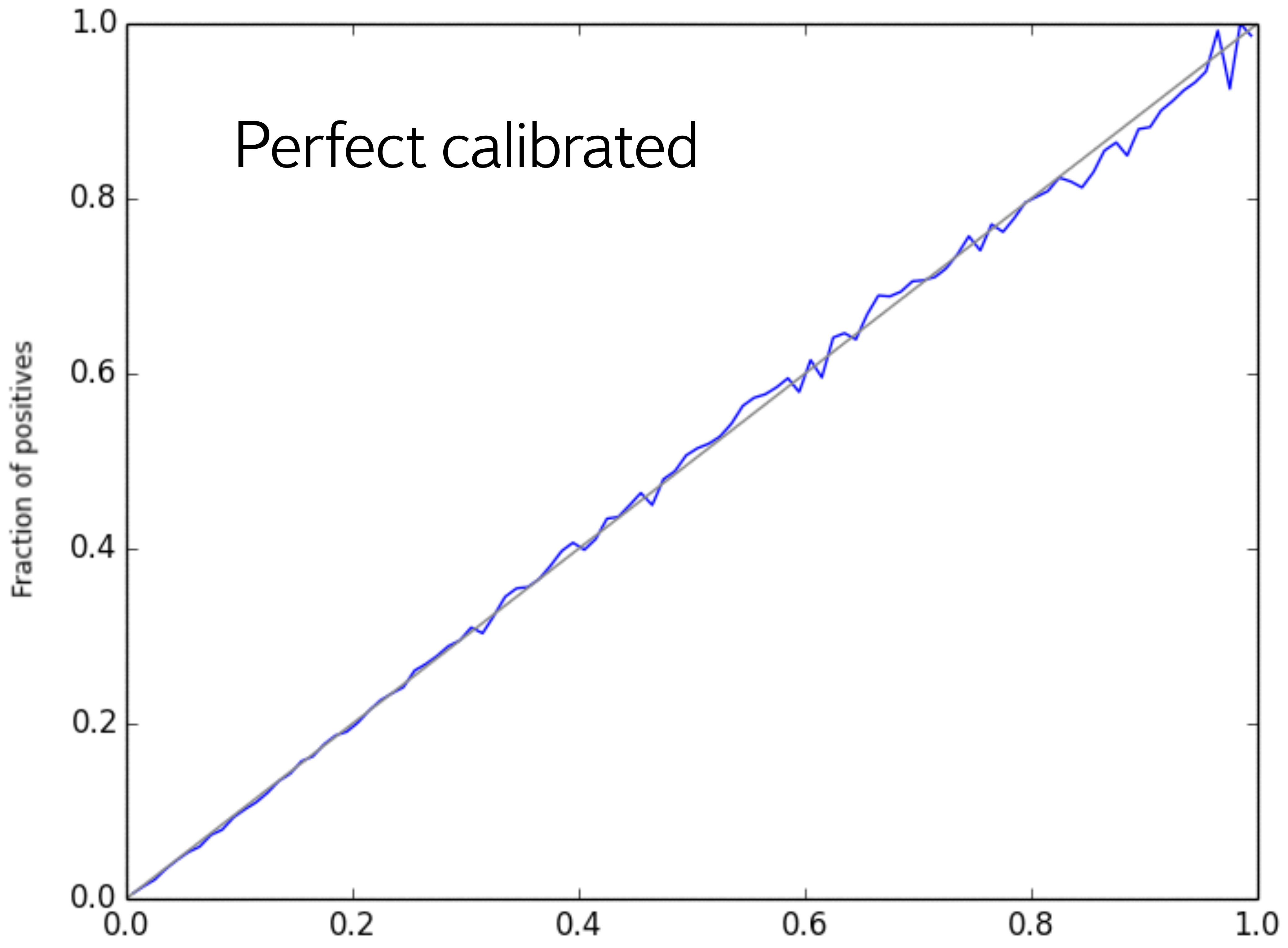
The transformation of the score returned by a classifier into a posterior class probability is called calibration

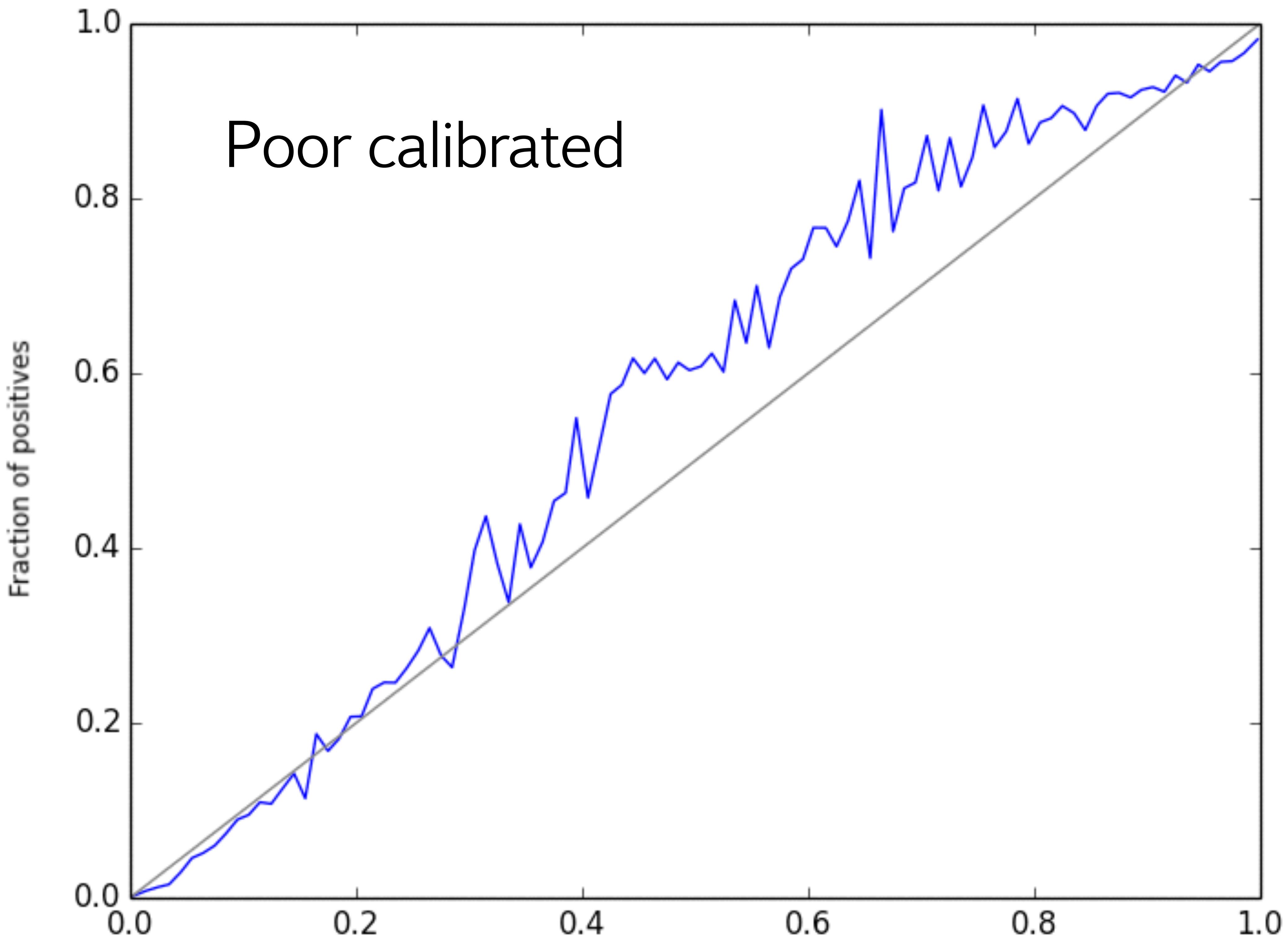
Examples



Sources and Descriprion

Examples


Perfect calibrated
Fraction of positives
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Poor calibrated
Fraction of positives
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Scoring rules

In decision theory, a score function, or scoring rule, measures the accuracy of probability predictions

Proper scoring rule

- Winkler and Murphy, 1968
- A scoring function will give a reward of $S(\mathbf{p}, \omega)$ if the ωth class occurs.
- A scoring rule, for which the highest expected reward is obtained by reporting the true probability distribution, is called proper.
- A scoring rule is strictly proper if it is uniquely optimized by the true probabilities.
- Strictly proper scoring rules remain strictly proper under linear transformation.
- The scoring rule S is local if $S(\mathbf{p}, \omega) = s(p_\omega)$ for some function s.

Strictly proper scoring rules

- The logarithmic scoring rule is a local strictly proper score (negative of surprisal):

$$L(\mathbf{p}, \omega) = \log_b (p_\omega), \qquad b > 0$$

- The quadratic scoring rule:

$$Q(\mathbf{p}, \omega) = 2\, p_\omega - ||\mathbf{p}||^2$$

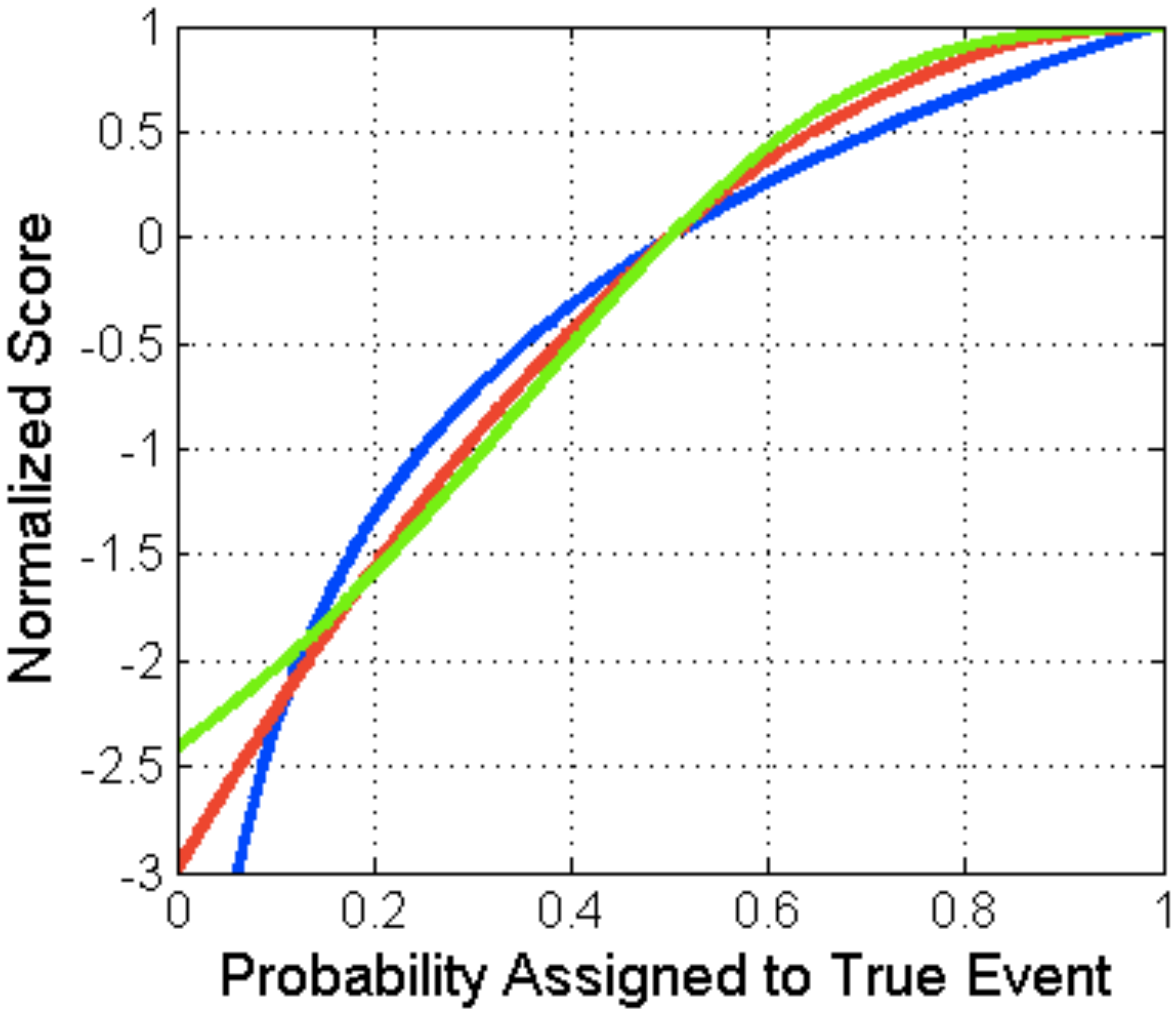
- The Brier score (should be minimized) obtained from quadratic by affine transform:

$$B(\mathbf{p}, \omega) = ||\, \mathbf{p} - \mathbf{I}_\omega \,||^2$$

- The spherical scoring rule:

$$S(\mathbf{p}, \omega) = p_\omega \,/\, ||\mathbf{p}||^2$$

Strictly proper scoring rules



logarithmic

spherical

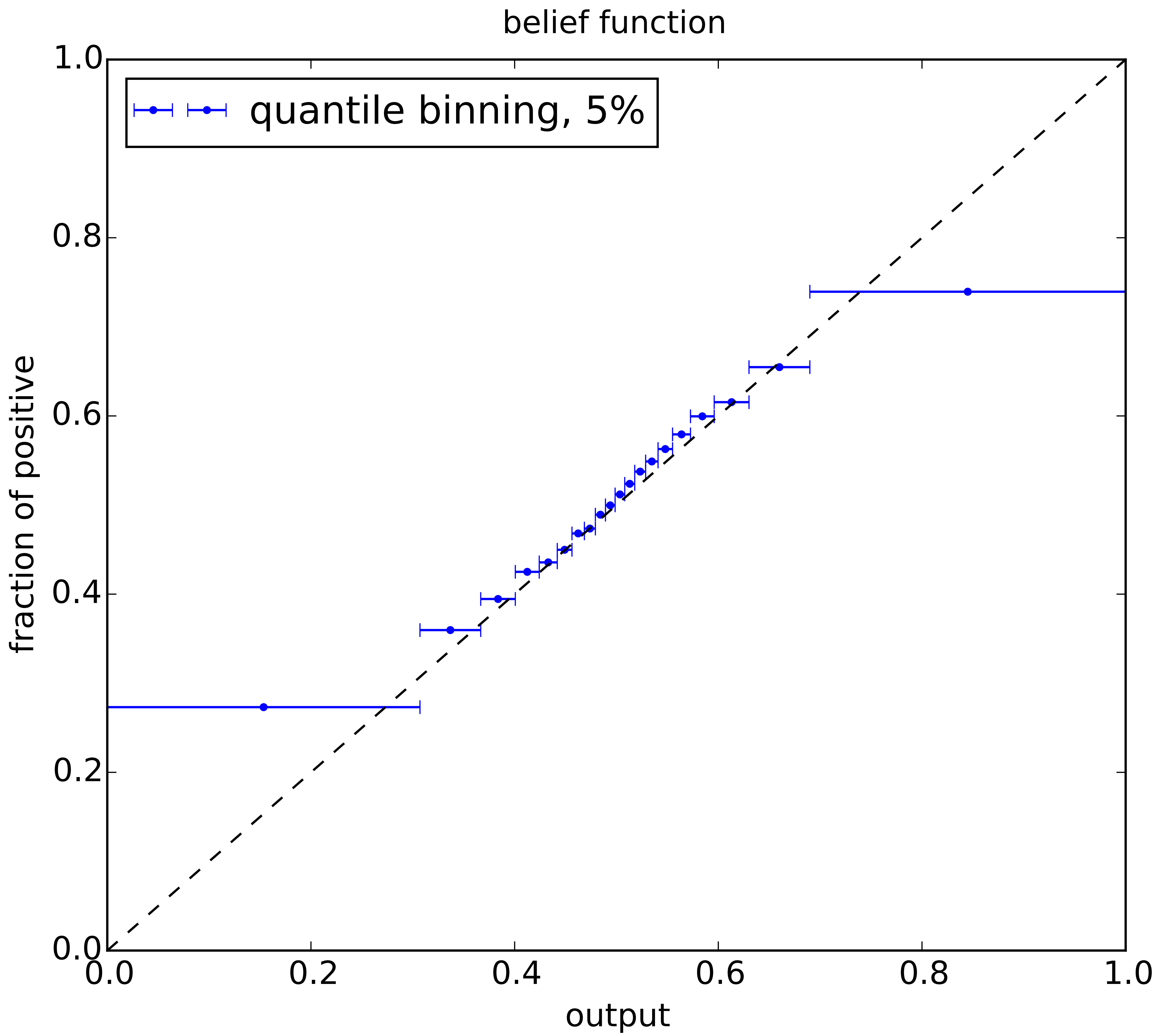
-Brier

Calibration approaches

Methods to calibrate classifier

- parametric
  - Platt scaling [2]
- non-parametric
  - quantile binning [3]
  - isotonic regression [4]

Quantile binning: calibration mapping



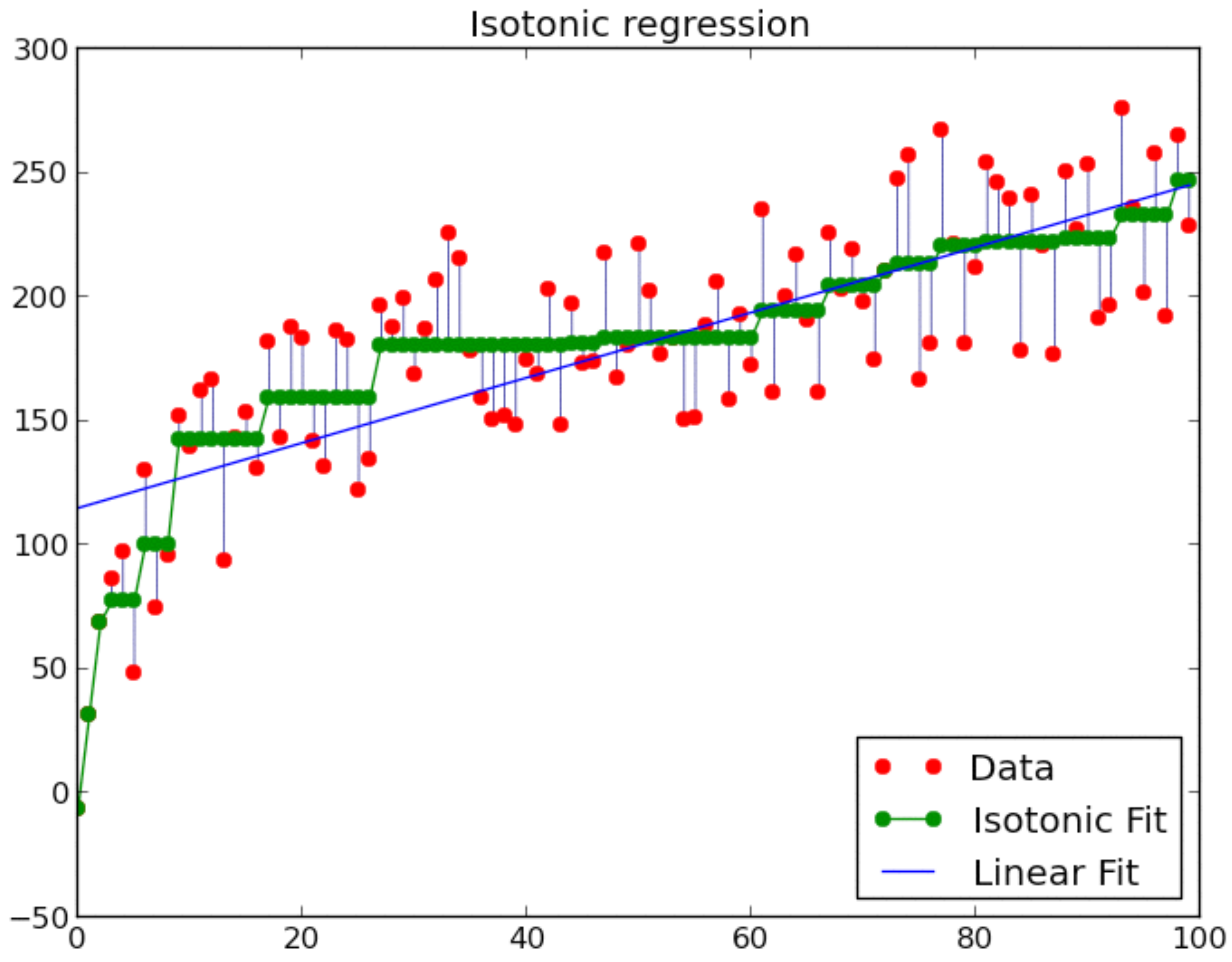
Quantile binning limitations

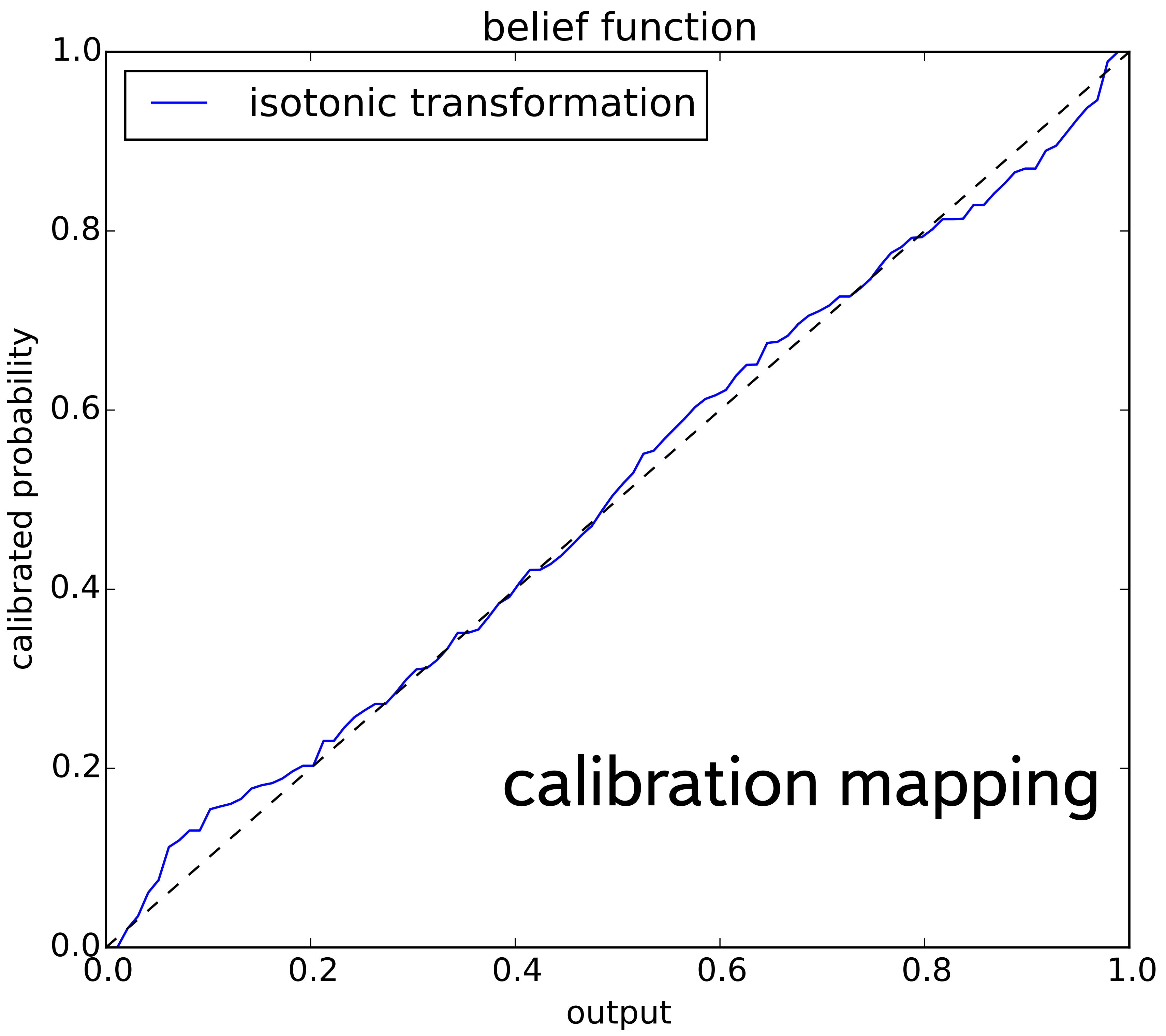
- bins map output into only N possibilities;
- fixed bin boundaries;
- which number of bins should be used?

Isotonic regression

- isotonic (monotonic) mapping
- generalizes a histogram binning model
- position of the bins boundaries are fitted
- optimizes the Brier score with isotonic restriction
- sometimes monotonicity assumption can be failed (ROC curve is not convex)

Isotonic regression





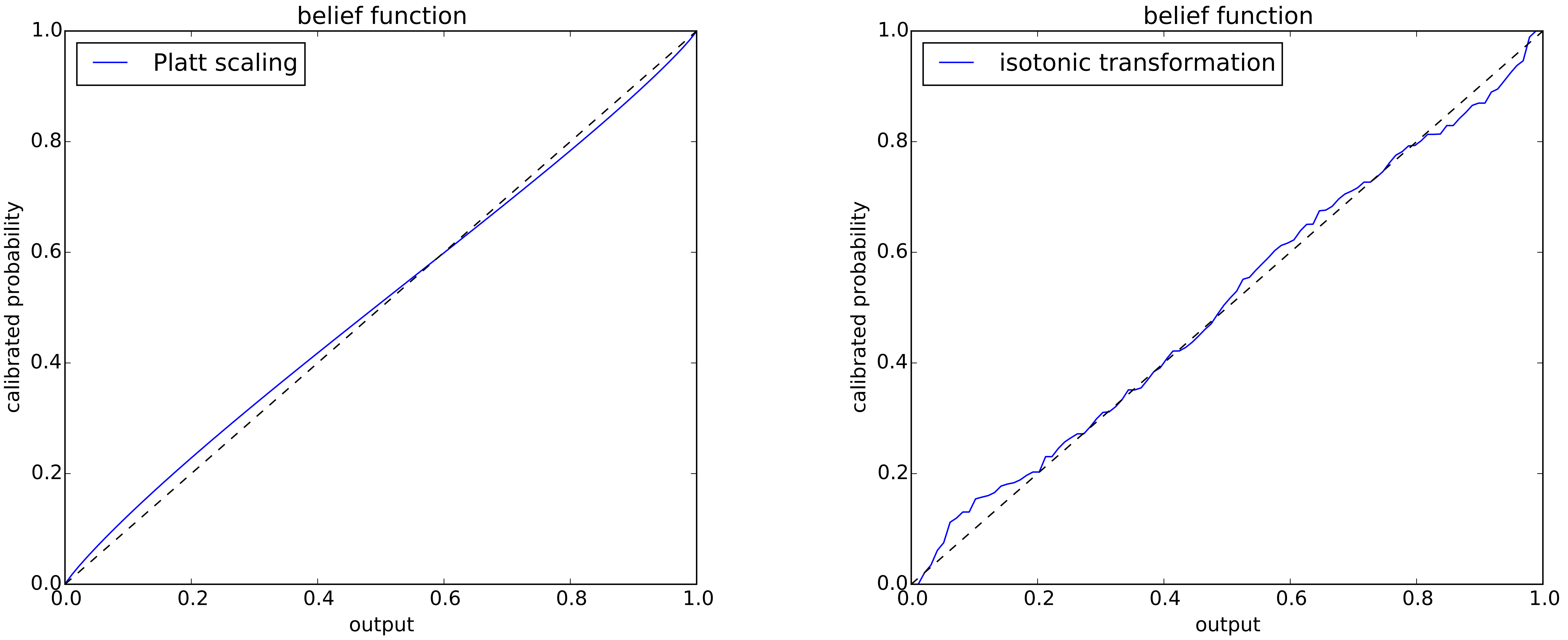
Sklearn implementation

Platt scaling (logistic regression)

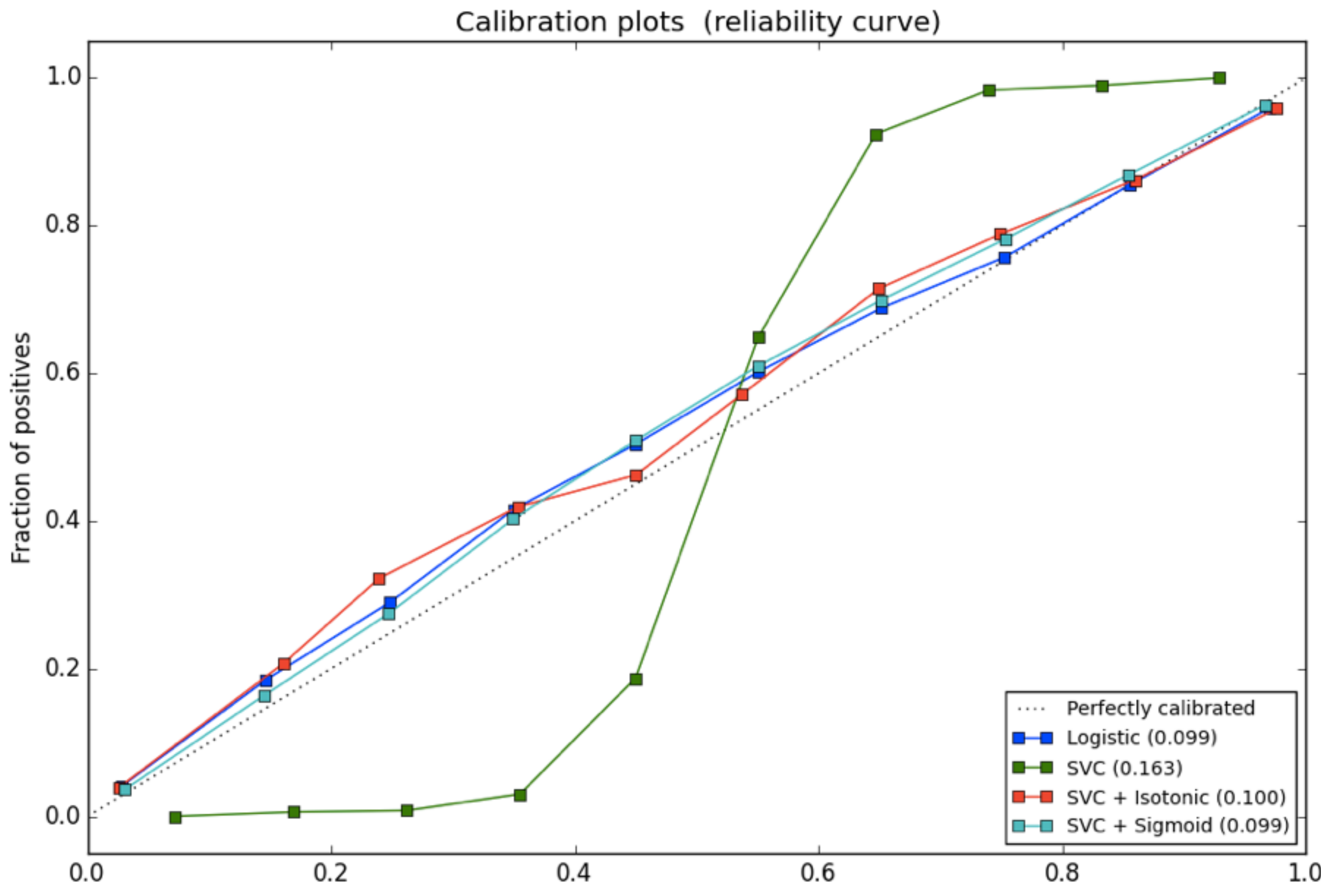
- sigmoid transformation
- learn affine transformation followed by sigmoid
- minimize logistic loss
- effective for SVMs
- not change the ranking (ROC curve stays the same)
- sigmoid function rarely fits the true distribution

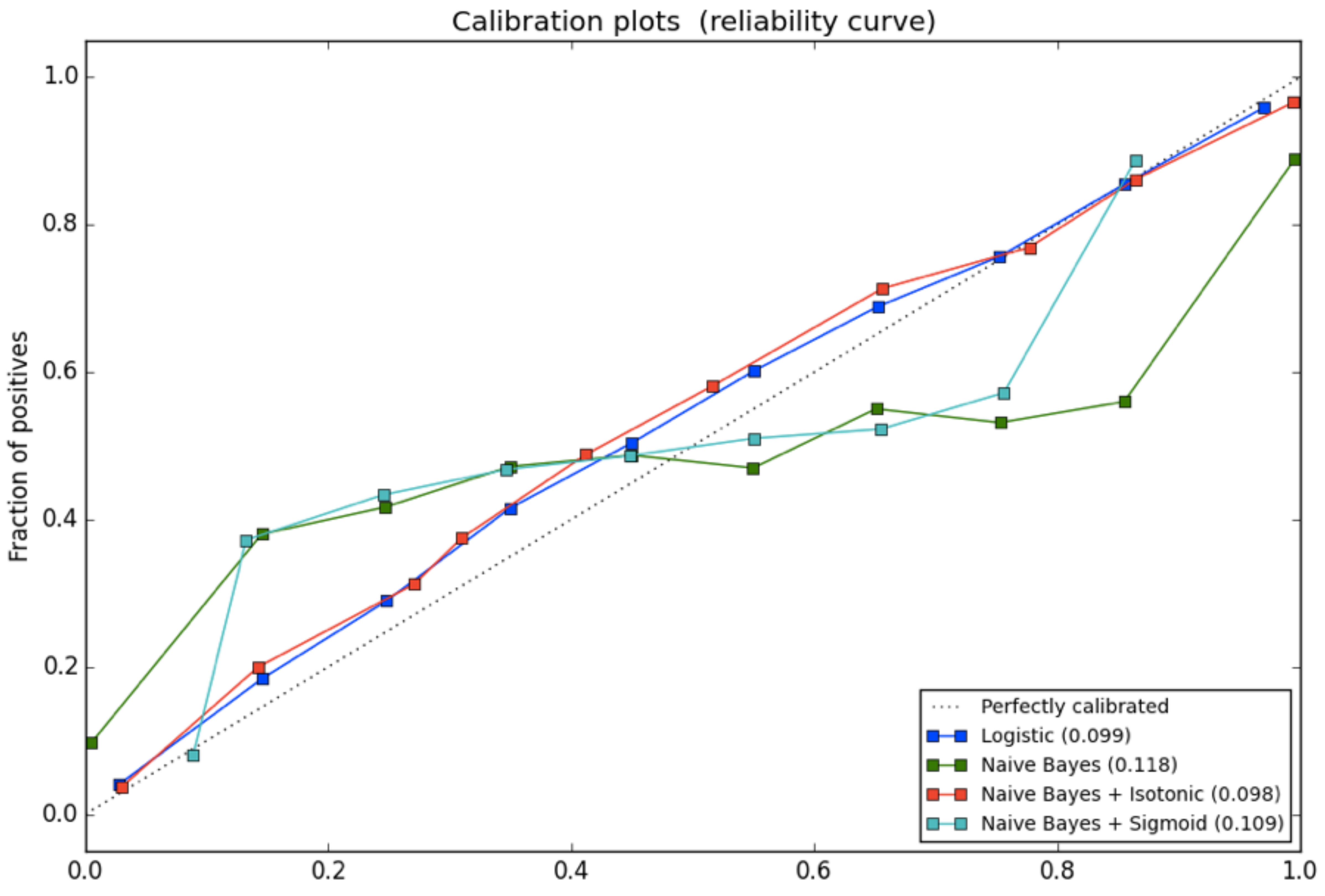
$$p_{true} = \frac{1}{1 + e^{-(Ap+B)}}$$

Platt scaling: calibration mapping



Examples: calibrated models





Recommendations

- Make sure you need probabilities :)
- Use Platt scaling and isotonic regression for calibration
- Use holdout to check your calibration rule
- Use logarithmic and Brier scorings to select optimal calibration rule (and model)

References

- Scoring Rules and Decision Analysis Education
- Some Comparisons among Quadratic, Spherical, and Logarithmic Scoring Rules
- Strictly Proper Scoring Rules, Prediction, and Estimation
- Obtaining Calibrated Probabilities from Boosting
- Blogpost about classifier's output calibration to probability
- Binary Classifier Calibration using an Ensemble of Near Isotonic Regression Models
- Venn-Abers predictors

Thanks for attention

Contacts

Likhomanenko Tatiana
researcher-developer

antares@yandex-team.ru, tatiana.likhomanenko@cern.ch