Machine Learning for Probabilistic Prediction

Quantitative Finance Webinar, Stony Brook University (11/11/2022)

Valery Manokhin, PhD, MBA, CFQ

Speaker Bio

- PhD in Machine Learning (2022) from Royal Holloway, University of London
- During PhD conducted research and published papers in probabilistic and conformal prediction. PhD supervised by Prof. Vladimir Vovk, the creator of Conformal Prediction (Prof. Vladimir Vovk is the last PhD student of Andrey Kolmogorov)
- Dr. Valery Manokhin holds a number of advanced MSc degrees including from the Moscow Institute of Physics and Technology (Physics/Math), UCL (Computational Statistics and Machine Learning), University of Sussex (Quant Finance) and an MBA from the University of Warwick
- Published in the leading machine learning journals, including ‘Neurocomputing’, ‘Journal of Machine Learning Research’ and ‘Machine Learning Journal’, also in the industry journals including ‘Frontiers in Energy Research’
- Created 'Awesome Conformal Prediction' - the most comprehensive professionally curated resource on Conformal Prediction (over 900 stars on GitHub). 'Awesome Conformal Prediction' has been featured at the leading conferences such as ICML and in Kevin Murphy's bestselling book 'Probabilistic Machine Learning: An Introduction'

Outline of this webinar

- Introduction to Probabilistic Prediction
- Probability Calibration
- Introduction to Conformal Prediction
- Conformal Prediction for Classification
- Conformal Prediction for Regression
- Conclusion

Why Probabilistic Prediction?

Machine Learning is primarily concerned with producing functions mapping objects onto predicted labels

Classical statistical techniques - for small scale, low-dimensional data

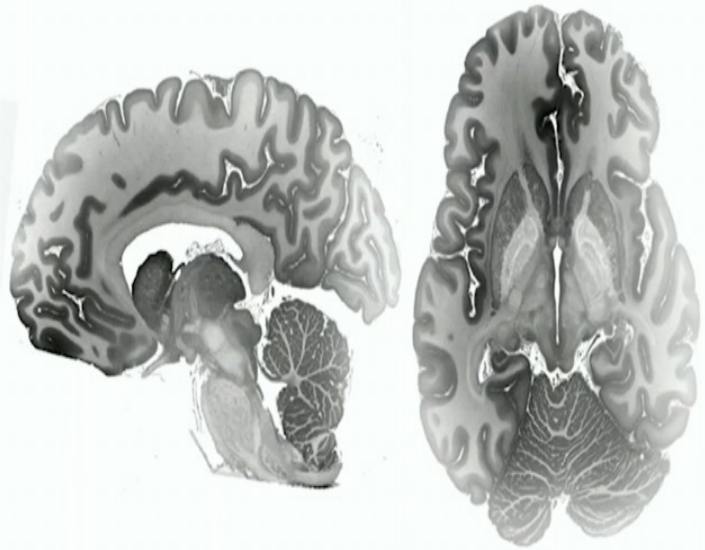
High-dimensional data does not necessarily follow well-known distributions and hence required new approaches (e.g. SVM)

Industry practitioners often focus on the point predictions, but overlook predictive uncertainty

Why Probabilistic Prediction?

- Picture from “Michael Jordan on Conformal Prediction”

High-dimensional Uncertainty Quantification



Decision-making algorithms **require uncertainty**

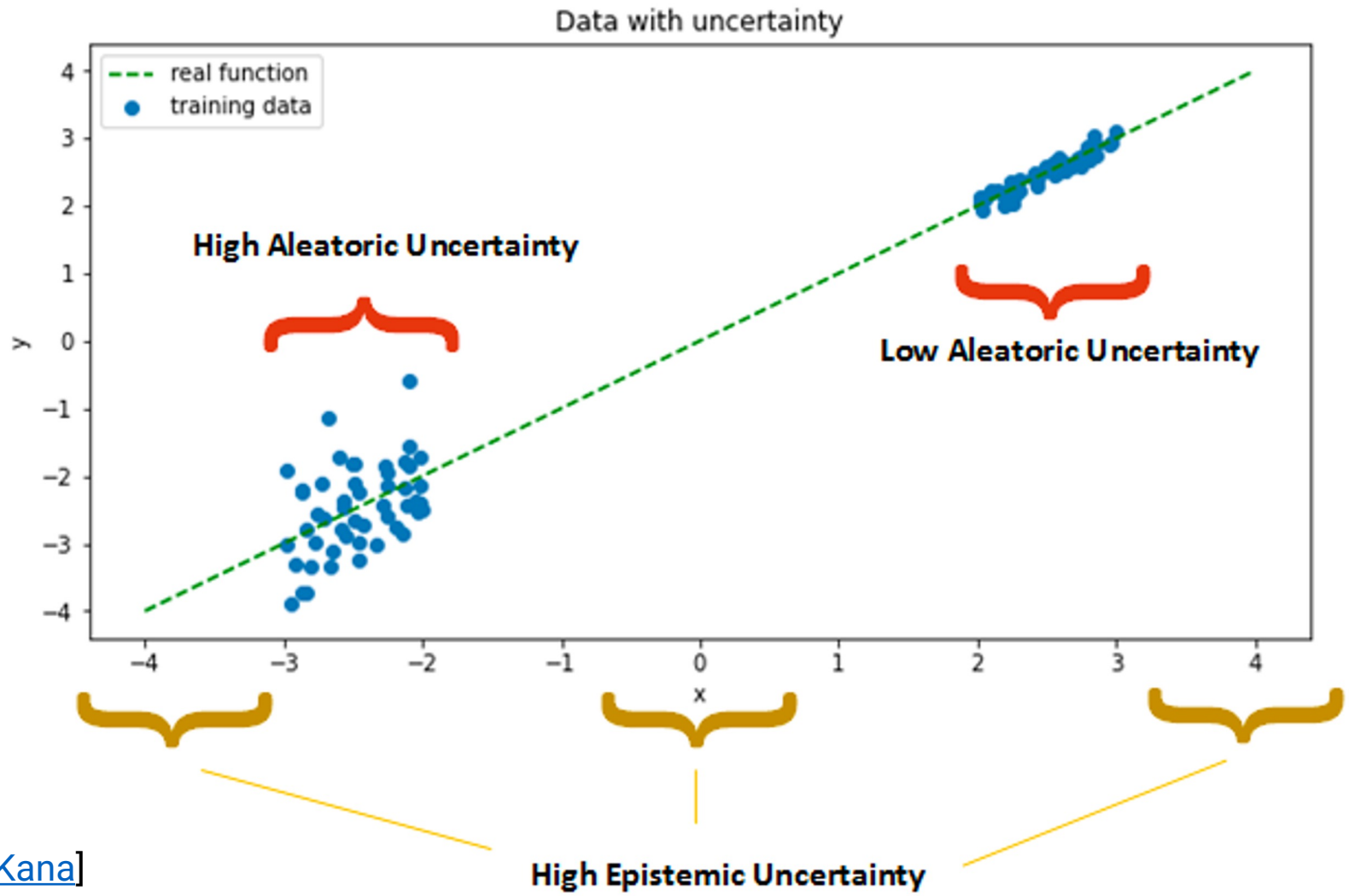
Need to rule-in or rule-out “bad” outcomes

Point predictions not enough

Why Probabilistic Prediction?

- Classification – insufficient to predict labels alone
- Prediction (non-time series) – producing Prediction Intervals is more valuable than only producing point predictions
- Time-series – producing probabilistic forecasts is way more valuable than producing point forecasts (demand planning, systematic trading)

The sources of uncertainty



[Image by Michel Kana]

Probability Calibration

- What is calibration
- How do we measure calibration
- How can we make calibrated predictions
- How can we calibrate predictions

Calibration – key concepts

$$P(model\ is\ correct \mid confidence\ is\ \boldsymbol{\alpha}) = \boldsymbol{\alpha}\ \textbf{(*)}$$

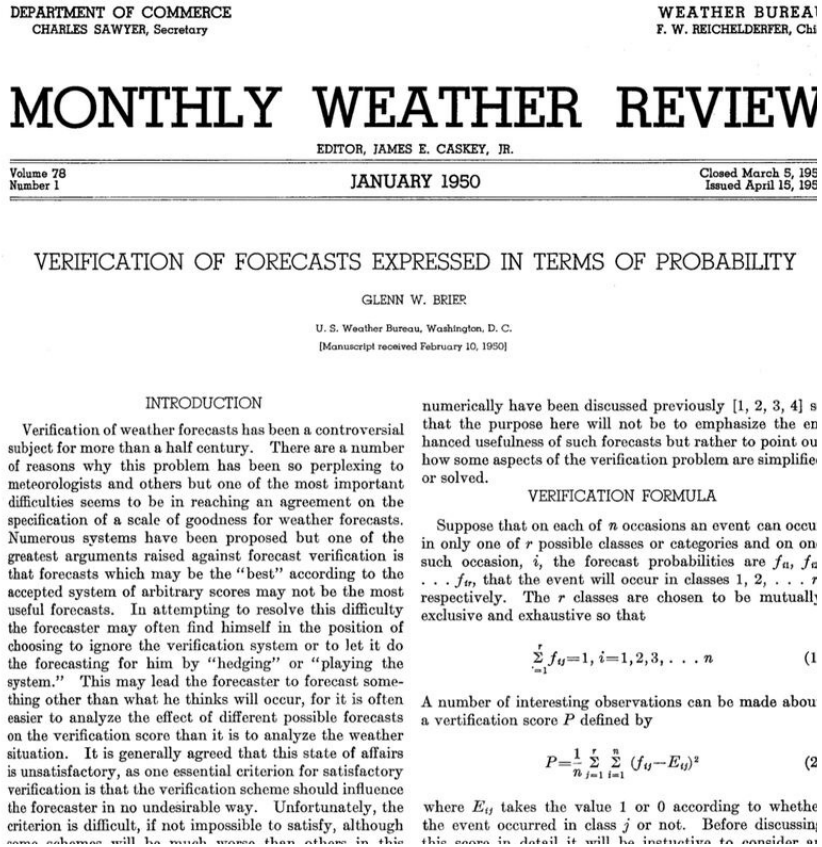
Suppose our model outputs confidence scores

We say our model is **calibrated** if (*) holds

In other words, ***α*-fraction** of all predictions with confidence α should be **correct**.

The history of probabilistic prediction ideas

- Weather forecasters were pioneers of many calibration concepts
- Glenn W. Brier proposed what has become known as the Brier score applicable to tasks in which predictions must assign probabilities to a set of mutually exclusive discrete outcomes or classes
- A forecast '70% chance of rain' should be followed by rain 70% of the time
- In general vector of probabilities should match empirical (observed) probabilities



DEPARTMENT OF COMMERCE
CHARLES SAWYER, Secretary

WEATHER BUREA
F. W. REICHELDERFER, Chi

MONTHLY WEATHER REVIEW

EDITOR, JAMES E. CASKEY, JR.

Volume 78
Number 1

JANUARY 1950

Closed March 5, 195
Issued April 15, 195

VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY

GLENN W. BRIER

U. S. Weather Bureau, Washington, D. C.

[Manuscript received February 10, 1950]

INTRODUCTION

Verification of weather forecasts has been a controversial subject for more than a half century. There are a number of reasons why this problem has been so perplexing to meteorologists and others but one of the most important difficulties seems to be in reaching an agreement on the specification of a scale of goodness for weather forecasts. Numerous systems have been proposed but one of the greatest arguments raised against forecast verification is that forecasts which may be the "best" according to the accepted system of arbitrary scores may not be the most useful forecasts. In attempting to resolve this difficulty the forecaster may often find himself in the position of choosing to ignore the verification system or to let it do the forecasting for him by "hedging" or "playing the system." This may lead the forecaster to forecast something other than what he thinks will occur, for it is often easier to analyze the effect of different possible forecasts on the verification score than it is to analyze the weather situation. It is generally agreed that this state of affairs is unsatisfactory, as one essential criterion for satisfactory verification is that the verification scheme should influence the forecaster in no undesirable way. Unfortunately, the criterion is difficult, if not impossible to satisfy, although

numerically have been discussed previously [1, 2, 3, 4] s
that the purpose here will not be to emphasize the en
hanced usefulness of such forecasts but rather to point ou
how some aspects of the verification problem are simplifie
or solved.

VERIFICATION FORMULA

Suppose that on each of n occasions an event can occu
in only one of r possible classes or categories and on on
such occasion, i, the forecast probabilities are f_{i1}, f_{i2}
$\ldots f_{ir}$, that the event will occur in classes 1, 2, $\ldots$ r
respectively. The r classes are chosen to be mutually
exclusive and exhaustive so that

$$\sum_{j=1}^{r} f_{ij}=1,\ i=1,2,3,\ldots n \qquad (1$$

A number of interesting observations can be made abou
a vertification score P defined by

$$P=\frac{1}{n}\sum_{j=1}^{r}\sum_{i=1}^{n}(f_{ij}-E_{ij})^2 \qquad (2$$

where E_{ij} takes the value 1 or 0 according to whethe
the event occurred in class j or not. Before discussin

Does calibration matter?

- Calibration matters in decisions involving risk
  - Should a bank approve a loan / mortgage?
  - Does X-ray / MRI scan show potentially life threatening condition?
  - Is there pedestrian on the road (self-driving cars)?
- Relative degree of confidence
  - Model I have higher confidence than Model II
- When might calibration be **not** necessary?
  - Relative ranking
  - Most likely prediction

Calibrated classification – an illustration

Model 1

Portfolio	Value	Default Probability	Expected Loss
Portfolio A	100	0.03	3
Portfolio B	250	0.01	2.5

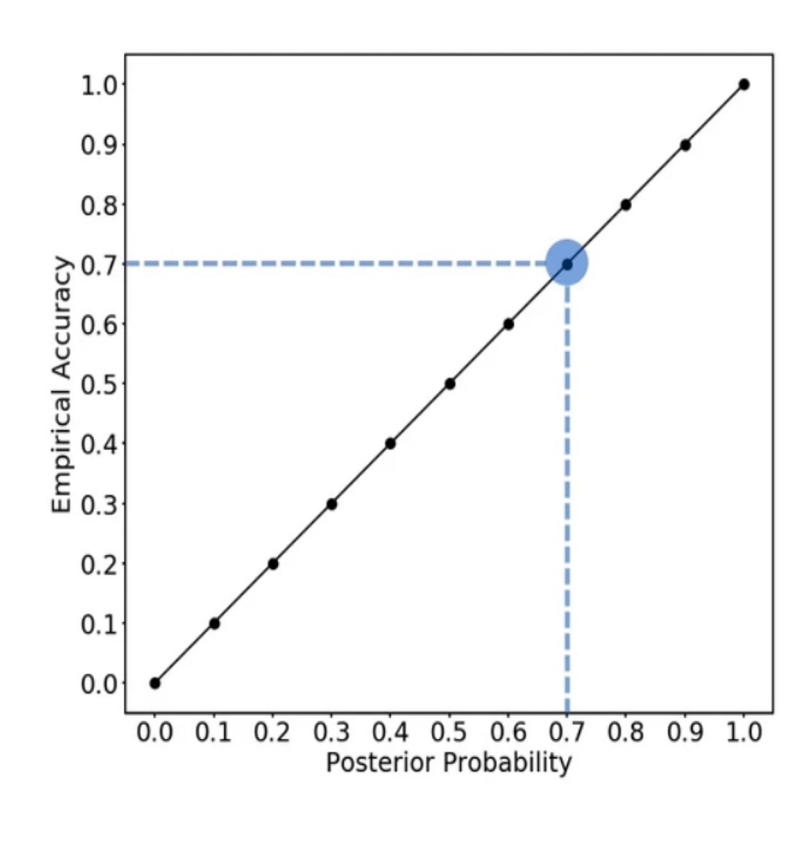
Model 2

Portfolio	Value	Default Probability	Expected Loss
Portfolio A	100	0.025	2.5
Portfolio B	250	0.015	3.75

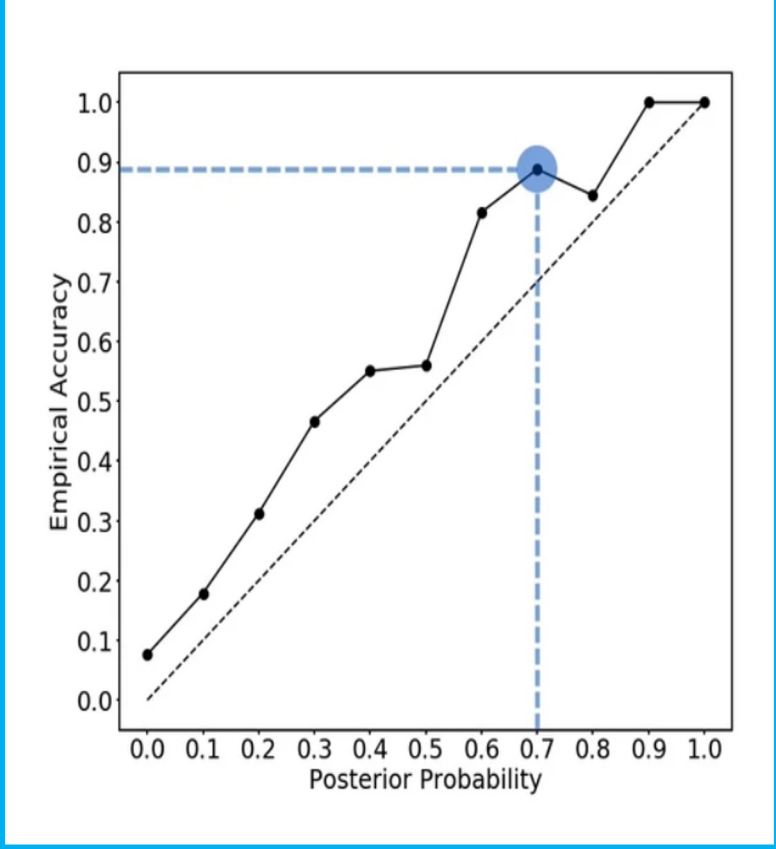
Model 1 – portfolio B is safer in terms of expected loss.
Model 2 – portfolio A is safer in terms of expected loss.
Which model will you trust to make decision whether to invest into Portfolio A or Portfolio B?

Calibration – the reliability diagram

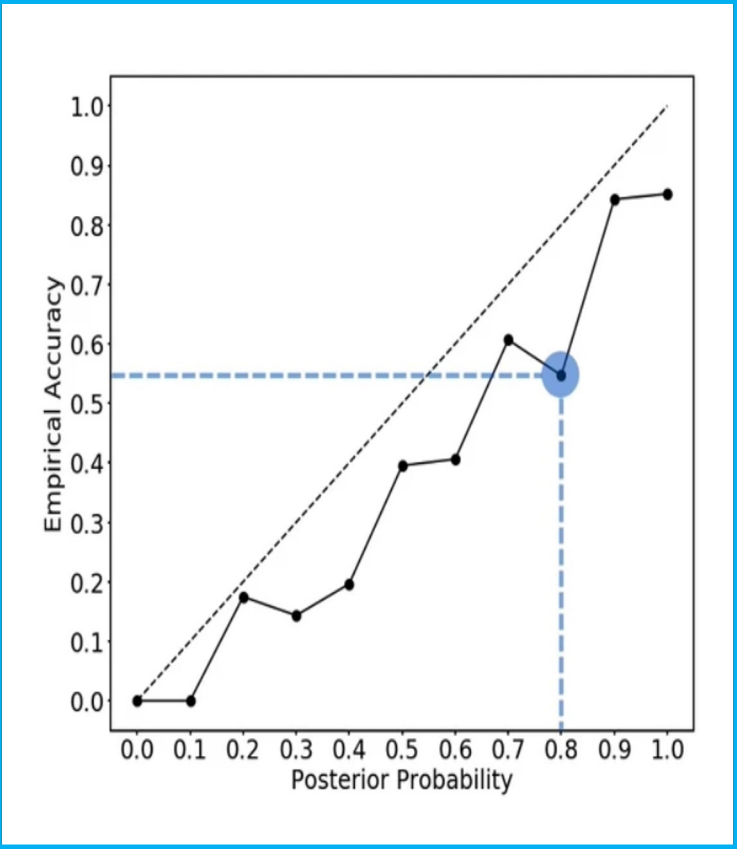
Calibrated



Under-confident



Over - confident



[Images from Desai & Durrett, 20]

Are machine learning models well-calibrated?

- The answer is NO, as shown both in classical and more recent papers most of the models are not well calibrated
- SVM needs additional calibration (Platt, 2001)
- Until recently it was thought that classical (shallow) neural networks are well-calibrated (Caruana & Mizil, 2005), this was debunked in later studies (Johansson, Gabrielson, 2019)
- More recent research – deep learning is miscalibrated (Guo, 2017; Mukhoti, 2020)
- Modern research -> conclusions in Caruana & Mizil, 2005 revisited. Traditional neural networks are mis-calibrated as well (Johansson, Gabrielson, 2019)
- Lack of calibration results in significant issues especially for critical applications such as health, finance, self-driving cars, pharma
- Examples of lack of calibration in other industries that are less exposed to regulation – advertising markets (e.g. Meta generates 95% of profits from advertising – ranking models rely on calibrated predictions!)

How to calibrate classifiers?

Platt's scaling (Platt, 2001)

Histogram binning (Zadrozny & Elkan, 2001)

• Isotonic regression (Zadrozny & Elkan, 2002)

• Nested dichotomies (Leathart, 2019)

Beta calibration (Kull, 2017)

• Scaling-binning (Kumar, 2019)

• Probability calibration trees (Leathart, 2018)

Bayesian Binning into Quantiles (Naeni, 2015

Ensemble of Near Isotonic Regression (Naeni, 2016)

• Temperature Scaling (Guo, 2017)

Entropy penalty (Pereyra, 2017)

Maximum Mean Calibration Error (Kumar, 2018)

Label smoothing (Mueller, 2019)

Focal loss (Mukhoti, 2020)

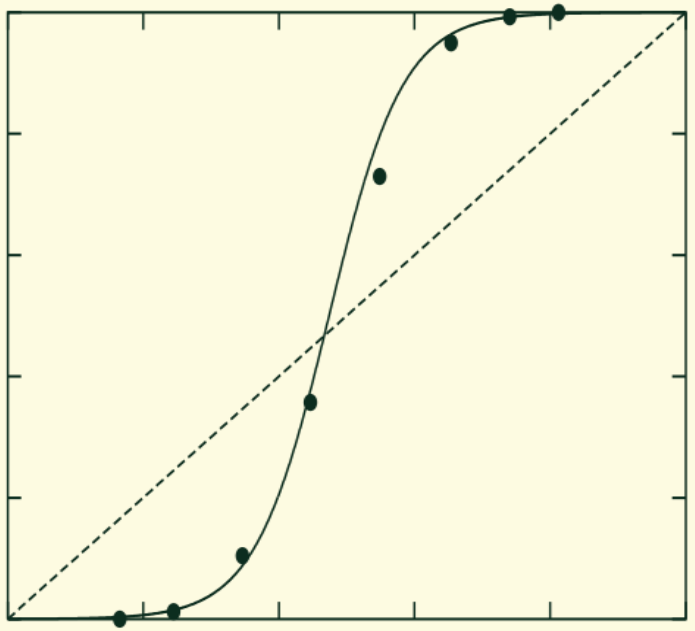
For overview of these methods see Machine Learning for Probabilistic Prediction

Platt's scaling

- Platt observed that the relationship between classification scores produced by the SVM and empirical probabilities tended to be of the sigmoid nature that can be described by the parametrized sigmoid function

- Parameter estimation via maximum likelihood on a new dataset (x_i,t_i) where $t_i = (y_i+1)/2$

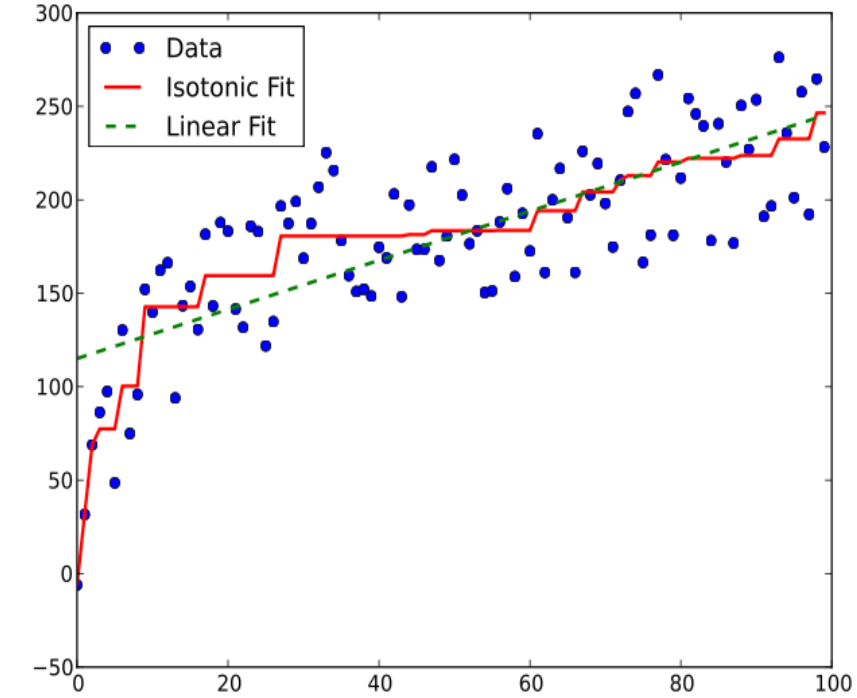
- Logistic calibration can be derived by assuming that the scores within both classes are normally distributed with the same variance (very restrictive assumption)
- In practice – Platt's scaling was designed specifically for SVM, other ML models don't produce sigmoid shapes calibration scores



$$P(y=1|x) = \frac{1}{1+\exp(Ax+B)}$$

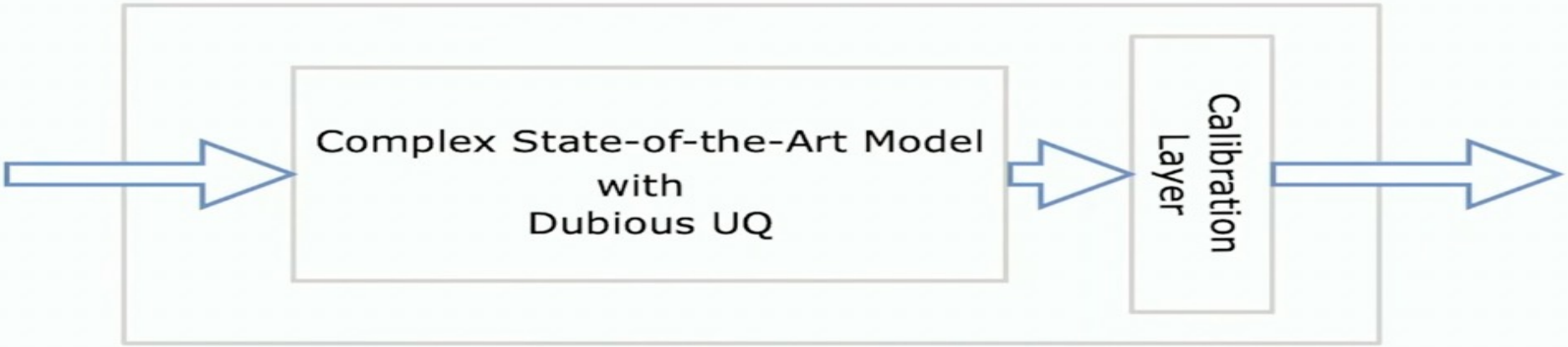
Isotonic Regression

- Isotonic regression (Zadrozny & Elkan, 2002)
- Increasing (isotonic) function between classification scores s and probabilities p
- For fitting isotonic regression, Zadrozny and Elkan use pair-adjacent-violators (PAV) algorithm, this non- parametric algorithm solves fitting problem in linear time O(N) by computing stepwise-constant isotonic function using general mean-squared error as an error metric
- Advantages of IR –non-parametric approach, does not rely on specific shape of calibration scores but tends to overfit especially in smaller datasets
- IR relies on perfect ranking by the underlying model (in practice having ROC AUC of 1 is never possible unless synthetic dataset)



Conformal Prediction

Uncertainty Quantification via Conformal Methods



- Appropriate for black-box deep-learning predictors, with no need for separate variance estimation
- Appropriate for complex Bayesian models, where the assumptions are hard to justify
- Essential for decision-making, risk control, and assembling modules

Source: Michael Jordan on Conformal Prediction

Conformal Prediction

- Powerful machine learning framework with origins in Kolmogorov's complexity
- Converts point predictors into probabilistic predictors (calibration layer around any black box)
- Unlike other appoaches, CP does not require the black box to be analyzed or even retrained
- Outputs well-calibrated probabilities as standard
- No need for prior probabilities (unlike in Bayesian learning)
- The only assumption is that of i.i.d (exchangeability to be precise, which is a weaker assumption than i.i.d)

Conformal Prediction

- Robust mathematical guarantees of validity of predictions (lack of bias)
- Is underlying model agnostic, one can use any statistical, machine or deep learning model as an underlying predictor
- Conformal Prediction guarantees validity of predictions in final samples of any size
- The framework has been around for some time, first book published in 2005 early papers date to late 1990s
- Exponential growth during the last 2-3 years 🚀🚀🚀🚀🚀

Why Conformal Prediction?

- One of the most influential machine learning researchers - Professor Michael I. Jordan: '**Conformal Prediction ideas are THE answer to UQ (uncertainty quantification), I think it's the best I have seen - its simple, generalisable etc.**' (ICML 2021 UQ workshop).

- Professors - Larry Wasserman (Carnegie Mellon) '**So the beauty of the conformal thing is how simple it is to do it and how general it is. So I think you know ideas that catch on, general ideas that are pretty general and easy to implement that you can picture yourself using in real applications are the reason that people using conformal prediction.**'

Source: Awesome Conformal Prediction

Conformal Prediction

Background: Conformal Prediction

(**Vovk, Gammerman, Shafer** et. al. ~1999-present)
(**Lei, Wasserman**, et. al. ~2015-present)

Conformal prediction: prediction sets that cover 90% (say) of future points

- data: $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \ldots, n+1$
- prediction set: $T: \mathcal{X} \to 2^{\mathcal{Y}}$

Theorem (informal). The sets from conformal prediction satisfy

$$90\% \leq P(Y_{n+1} \in T(X_{n+1})) \leq 90\% + 1/n$$

\+ finite-sample coverage
\+ any predictive model
\+ any distribution

\– sometimes wrong error notion
\– conservative for high-d outputs

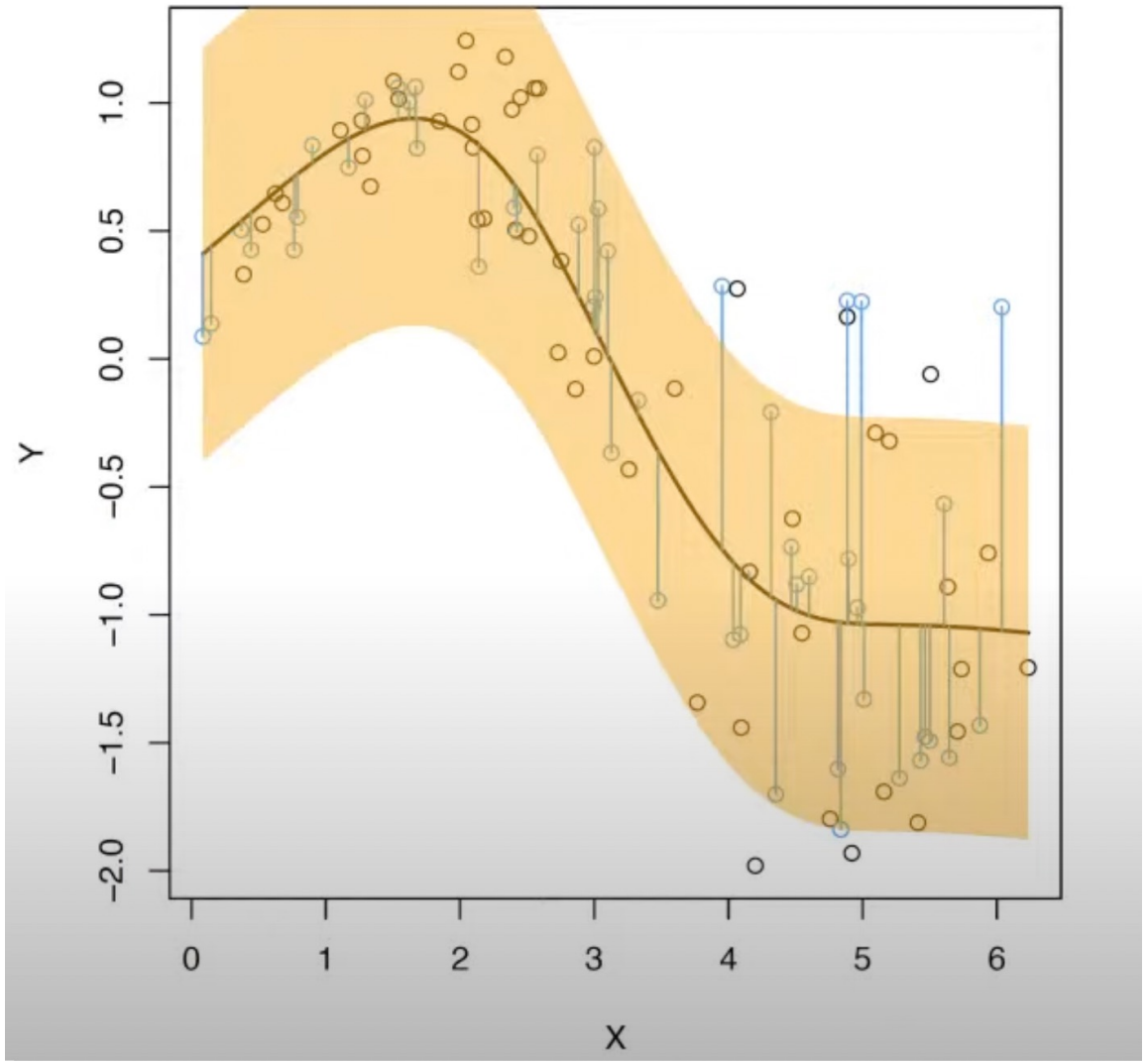
Source: Michael Jordan on Conformal Prediction

How does Conformal Prediction work (split conformal, regression task)

Step 1	Step 2	Step 3	Step 4	Step 5	Step 6
Define a noncomformity function. For example in regression task this could be as simple as y – y_hat	Train the model using proper training set (data set split into proper training / calibration / test set)	Use trained model from Step 2 to produce point predictions for (i) calibration set (ii) test set	Measure the nonconformity of labeled objects on the calibration set using the chosen non-conformity measure	Use the distribution of errors from Step 4 to obtain certaine error quantile (say 0.95%)	For any new test objects object xn+1 we use error quantile from Step 5 to produce Prediction Interval around point predictions

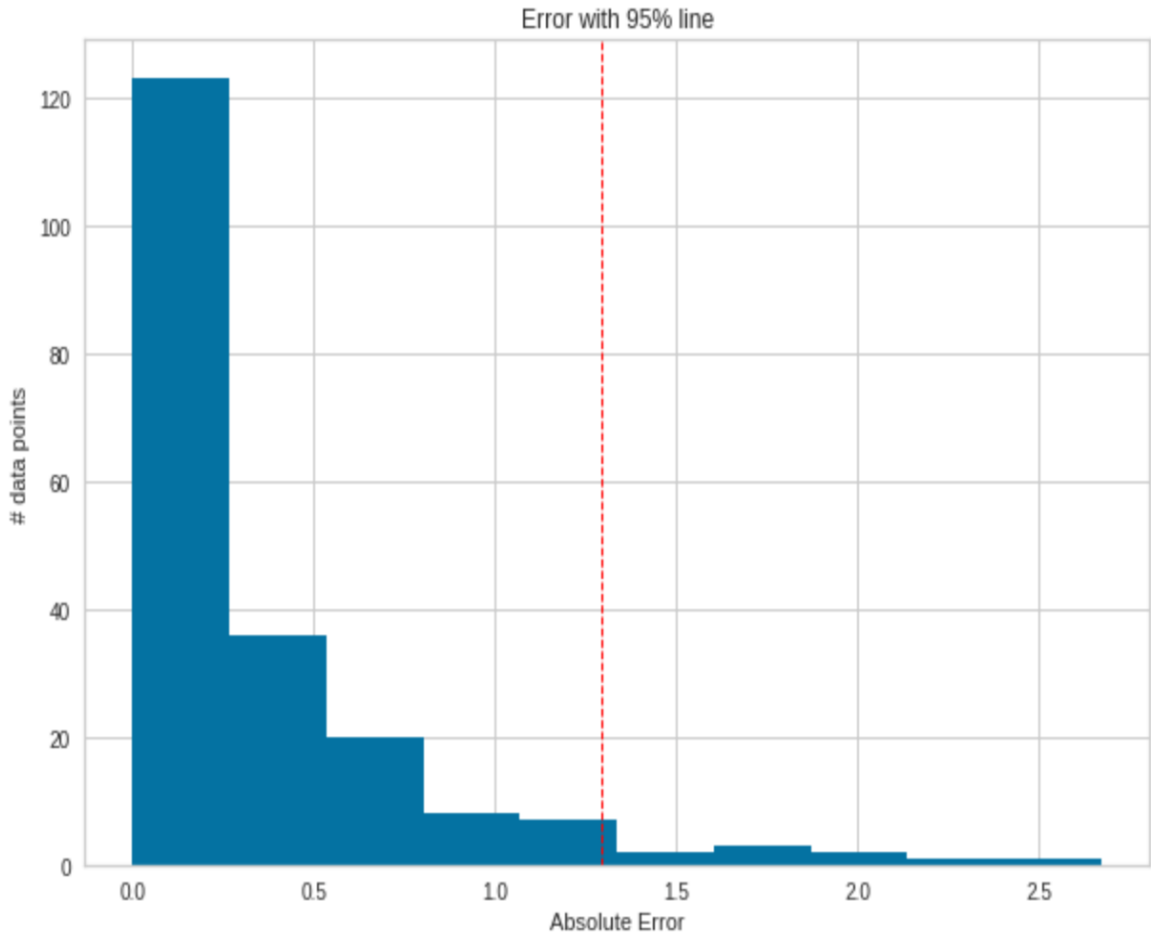
How does Conformal Prediction work (split conformal, regression task)?

- Steps 2 & 3 – use trained model to produce point predictions for the calibration set
- Main idea: look at holdout (aka calibration) set
- Draw yellow band with certain percentile
- Future test point is as likely to be in the predictive and as calibration set points due to exchangeability assumption



Obtaining Prediction Intervals

- Steps 4 & 5: compute Prediction Intervals using the distribution of errors from the calibration set
- One can chose any quantile, in this case 95% quantile is chosen to form 95% Prediction Intervals
- Calibrated by default, as standard in Conformal Prediction such 95% PI will contain ~95% of points on any unseen test set regardless of the underlying prediction model, the data distribution and the dataset size



Distribution of prediction error on the calibration set

Obtaining Prediction Intervals

- Step 6: Use point predictions generated by the ML trained on the proper training set in Steps 2-3 to obtain point predictions on the test dataset

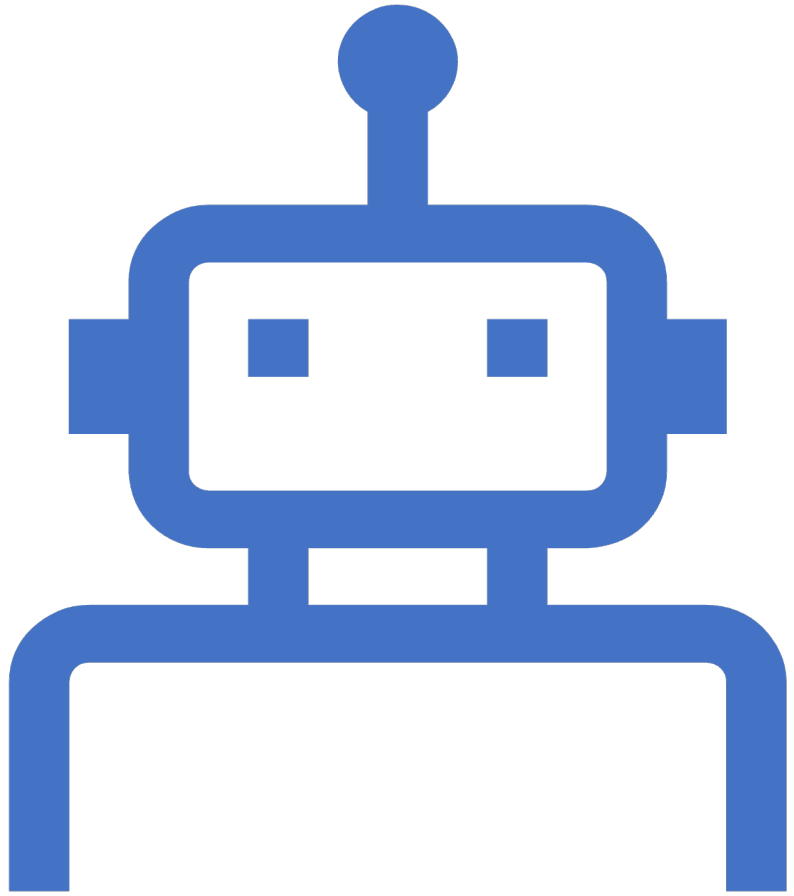
- Use Prediction Intervals calculated in Steps 4-5 using selected confidence level and add/subtract half of the calculated length of PIs to each point prediction on the test set

	actual	predicted	lower_interval	upper_interval
0	0.762	0.757060	-0.535779	2.049899
1	1.732	2.373121	1.080282	3.665960
2	1.125	2.293931	1.001092	3.586769
3	1.370	1.948320	0.655481	3.241159
4	1.856	2.179780	0.886941	3.472619
...	...	...	...	...
408	1.073	1.211820	-0.081019	2.504659
409	0.517	0.959280	-0.333559	2.252119
410	2.316	1.870050	0.577211	3.162889
411	0.738	0.959650	-0.333189	2.252489
412	2.639	2.824910	1.532071	4.117749

Making probabilistic predictions

Machine Learning classification

- Most models at best produce classification scores, **not class probabilities**
- In Scikit-learn naming class scores ‘predict_proba’ is misleading for users
- Calibrated probabilities are key for correct decision making
- Meanwhile most of machine learning / deep learning models don’t produce well calibrated probabilities
- Model calibration evaluation is key



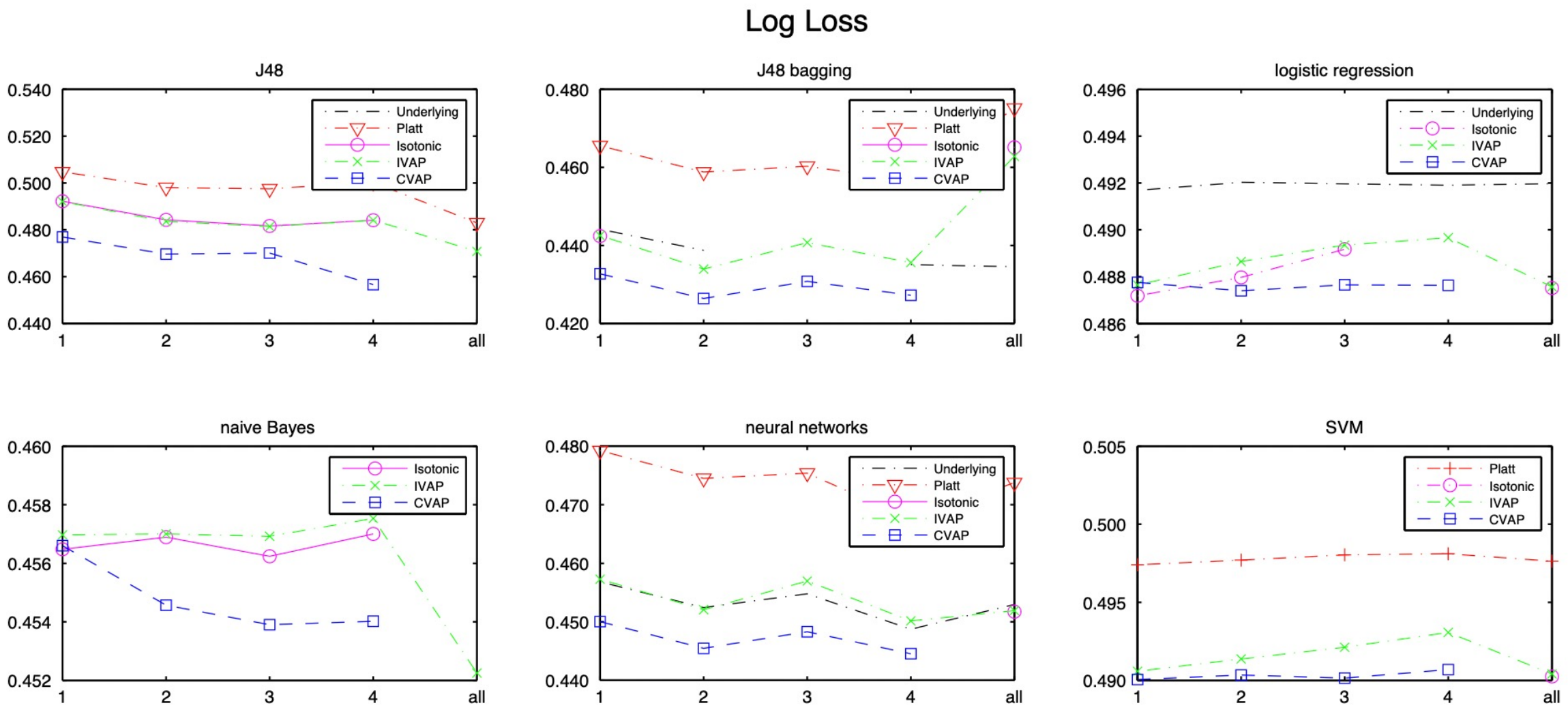
Python’s «predict_proba» Doesn’t Actually Predict Probabilities (and How to Fix It)

Conformal Prediction for Classification (Venn ABERS)

- Many machine learning algorithms for classification are *scoring classifiers*: they output a prediction score $s(x)$ and the prediction is obtained by comparing the score to a threshold
- One could apply a function g to $s(x)$ to calibrate the scores so that $g(s(x))$ can be used as predicted probability. Assumption - $g()$ to be a non-decreasing function
- Isotonic regression is known to overfit, basic idea – fit isotonic regression twice using both of possible labels for class 1 (0 and 1)!
- Venn-ABERS predictors prevent overfitting and inherit the in-built validity guarantee of Conformal Prediction
- Cost – multi-probability output, but the benefit is that this also provides an indication of the reliability of the probability estimates (indication of the uncertainty on the probability estimate itself) [p = p_1 / (1 – p_0 + p_1)
- No assumptions on function shape (compare to Platt's Scaling that uses a sigmoid as calibrating function)

Paper talk: Large-Scale Probabilistic Prediction With and Without Validity Guarantees - Prof. Vladimir Vovk

Conformal Prediction for Classification (Venn ABERS)



Multi-class probabilistic classification using inductive and cross Venn–Abers predictors

- Venn-ABERS predictors have beer successfully extended to multi-class
- The basic idea is to reduce multiclas classifiers to the binary classification cas and then combine the results form th pairwise classification to obtain multiclas probabilities

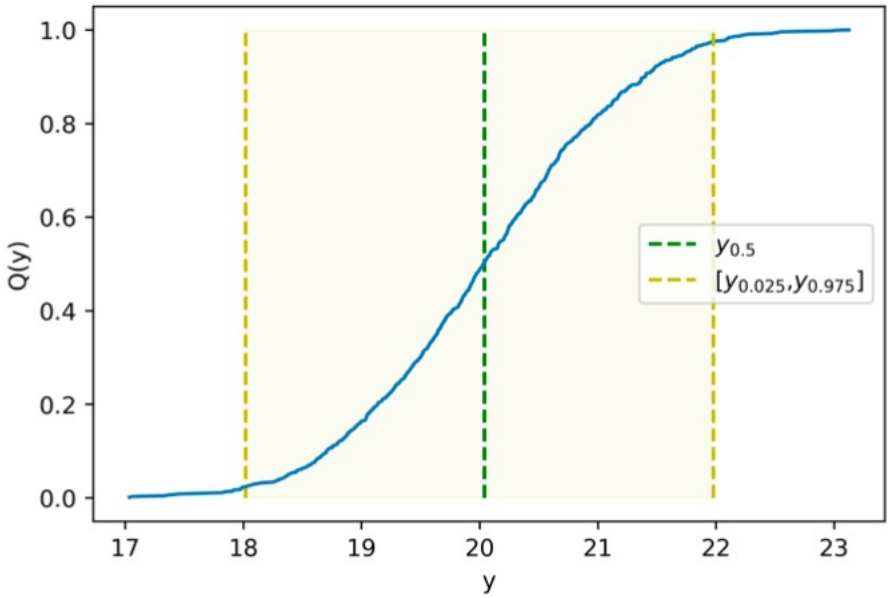
Github : Multi-class-probabilistic-classification

Table 3.1: The Log loss for the waverfont data set

	uncalibrated	sigmoid	isotonic	ivap	cvap
Naive Bayes	0.5096	0.2756	0.3189	0.2279	0.2267
KNN	0.4759	0.2686	0.2655	0.3644	0.3506
Support Vector Machine	0.1881	0.1927	0.2536	0.1985	0.1914
logistic regression	0.1963	0.1987	0.2576	0.2052	0.2037
neural network	0.4172	0.3601	0.3333	0.4036	0.3925
Random Forest	0.2491	0.2002	0.2192	0.2066	0.2054
LightGBM	0.2377	0.2534	0.2331	0.2199	0.2080
XGBoost	0.2657	0.2510	0.2319	0.2189	0.2103
CatBoost	0.1948	0.2308	0.2246	0.2158	0.2013
Ada Boost	0.5808	0.2447	0.3360	0.2498	0.2278

Conformal Prediction for Regression (Conformal Predictive Distributions)

Conformal predictive systems for regression output conformal predictive distributions (cumulative distribution functions)



[Tutorial on Conformal Predictive Distributions]

[Github: Conformal regressors and predictive systems]

Conformal Predictive Distributions

A *split conformal predictive system* can be constructed as follows:

1. randomly divide the training data into two disjoint subsets; the proper training set and the calibration set
2. train the underlying model h using the proper training set
3. calculate scores $\alpha_1, \ldots, \alpha_q$ for the calibration set, where

$$\alpha_i = \frac{y_i - h(\boldsymbol{x}_i)}{\sigma_i}$$

4. let $\alpha_{(1)}, \ldots, \alpha_{(q)}$ be the scores sorted in ascending order
5. for each test object $\boldsymbol{x}$ with difficulty σ:
   let $C_{(i)} = h(\boldsymbol{x}) + \alpha_{(i)}\sigma$ for $i \in \{1, \ldots, q\}$
   let $C_{(0)} = -\infty$ and $C_{(q+1)} = \infty$
   output the conformal predictive distribution:

$$Q(y) = \begin{cases} \frac{n+\tau}{q+1} & \text{if } y \in (C_{(n)}, C_{(n+1)}) \text{ for } n \in \{0, \ldots, q\} \\ \frac{n'-1+(n''-n'+2)\tau}{q+1} & \text{if } y = C_{(n)} \text{ for } n \in \{1, \ldots, q\} \end{cases}$$

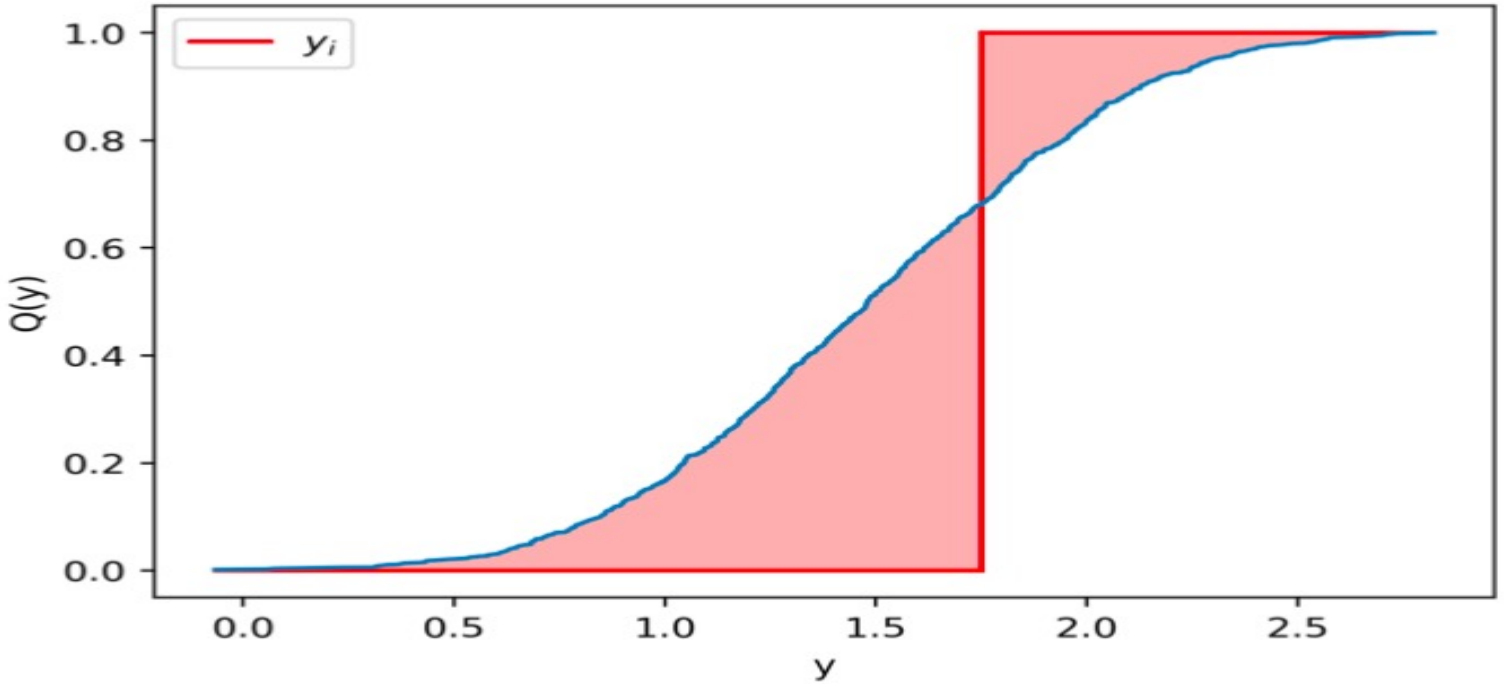
Source: Conformal Regressors and Predictive Systems – a Gentle Introduction

Conformal Prediction for Regression (Conformal Predictive Distributions)

- Conformal Predictive Distributions come (as standard in Conformal Prediction) with a validity guarantee - the output of the conformal predictive distributions (the p-values) for the correct target values are distributed uniformly on [0, 1]
- Having calibrated probabilistic prediction (CDF) allows optimal decision making and robust risk management and control
- The validity can be investigated by testing whether the p-values for a test set are distributed uniformly on [0, 1], e.g., using the Kolmogorov-Smirnov test
- When using the conformal predictive distributions for calibration, the predictive performance, e.g., as measured by mean-squared error, can be compared to the original underlying model
- Continuous ranked probability score (CRPS) is another option, which uses the full conformal predictive distributions

Conformal Prediction for Regression (Conformal Predictive Distributions)

$$CRPS(Q, y_i) = \int_{-\infty}^{\infty} (Q(y) - \mathbb{1}(y \geq y_i))^2 dy$$



Resources

- Awesome Conformal Prediction by Valery Manokhin, 2021
- Machine Learning for Probabilistic Prediction (PhD thesis, Valery Manokhin)
- Multi-class probabilistic classification using inductive and cross Venn–Abers predictors (Valery Manokhin, 2017)
- A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. Anastasios N. Angelopoulos, Stephen Bates. ArXiv preprint. 2021.
- Algorithmic Learning in a Random World. Vladimir Vovk, Alexander Gammerman, Glenn Shafer. Springer. 2005.
- Symposium on Conformal and Probabilistic Prediction with Applications (COPA conference 2012-2022).
- Workshop on Distribution-Free Uncertainty Quantification (DFUQ @ ICML 2021-2022).
- Cross-conformal predictive distributions (Vovk et. al. 2018)
- Nonparametric predictive distributions based on conformal prediction (Vovk et. al. 2017)