Uncertainty quantification using conformal inference

Community of Practice - Data Science

March 22, 2024

Agenda

Why uncertainty quantification?

Uncertainty Quantification (UQ) is essential in many situations:

- model predictions to make decisions
- design robust systems that can handle unexpected situations
- automated tasks with ML and need an indicator of when to intervene
- want to communicate the uncertainty associated with our predictions to business

Uncertainty in statistics[1]

Aleatoric uncertainty ('Data related')

Uncertainty that arises from the inherent randomness of an event, e.g. due to errors and noise in measurements.

Epistemic uncertainty ('Model related')

Uncertainty that arises from variability in real-world situations, unknown/latent data structures, errors during model training or errors in the model structure ('misspecification').

[1] C. Gruber et al. "Sources of Uncertainty in Machine Learning - A Statisticians' View". In: arXiv:2305.16703 (2023).

Prediction: measuring, collecting, cleaning data and training

- Model is trained on a random sample, making the model itself a **random variable**
- Some models are trained in a **non-deterministic way**, e.g. random weight initialization in neural nets, sampling mechanism(s) in ensemble methods
- Uncertainty increases in small samples ($\rightarrow$ *variance*!)
- Hyperparameter tuning, model selection, variable selection, all add uncertainty to the modeling process ('Epistemic')
- **Measurement errors** (label annotations, copying errors, faulty measurements, missing data etc.)

Uncertainty quantification in stats/machine learning

Heuristics of uncertainty quantification

Uncertainty heuristic

Numeric quantity to express the degree of uncertainty or confidence in a specific experimental outcome or measurement

Examples:

- Class probabilities (or functions of them, e.g. entropies)
- Bayesian posterior predictive distribution
- Bootstrapped predictive intervals (estimate $\mathrm{Var}(\widehat{f}(x))$)
- Quantile regression based predictive intervals

However: no theoretical guarantees to cover the true outcome!

Example: classification tasks

Model calibration

Calibration

Extent to which predicted probabilities reflect the true 'likelihood' of an event.

Example:

Model is 'well-calibrated' if an event that is assigned a 70% chance by the model, actually occured approximately 70% of the time.

- Especially crucial if model probabilities (thresholds etc.) are communicated to users (vs. ranks)

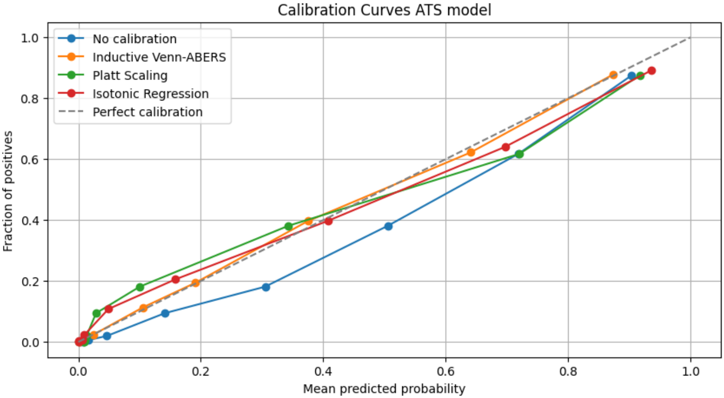
Model calibration - Counterexamples

- Gaussian naive Bayes: probabilities **often** close to 0 or 1 due to underlying assumptions about feature independence.
- Random forests: **seldom** return values close to 0 or 1 $\rightarrow$ average responses of multiple base learners.
- (Simple) Neural nets are **often well-calibrated**, but as architectures grew more and more complex over the years, modern nets are **frequently poorly calibrated**[2]

[2] C. Guo et al. "On Calibration of Modern Neural Networks". In: Proceedings of the 34 th International Conference on Machine Learning, Sydney, Austr 70 (2017).

Comparison of (re)calibration methods

Example: ATS model (Catboost)



Comparison of (re)calibration methods

Example: Catboost using ATS data

Method	Brier Score	Log Loss
No calibration	0.08674	0.2698
Inductive Venn-ABERS	0.08245	0.25787
Platt Scaling	0.08511	0.27265
Isotonic Regression	0.08328	0.26851

Table: Calibration methods comparison

Bayesian methods

Actually...

- Posterior inference guarantees theoretical coverage, too, however only given **all model assumptions** are fulfilled!
- Bayes theorem: $\pi(\theta|y) \propto \pi(\theta) \cdot f(y|\theta)$
- Parametric Bayes: Assumptions w.r.t. prior + DGP (aka likelihood structure)
- Prior sensitivity in small samples
- Can partially be mitigated using non-parametric priors (e.g. Dirichlet processes, Gaussian processes etc.) and/or semi-parametric data models (e.g. Bayesian Regression Trees etc.)

Bayesian prediction

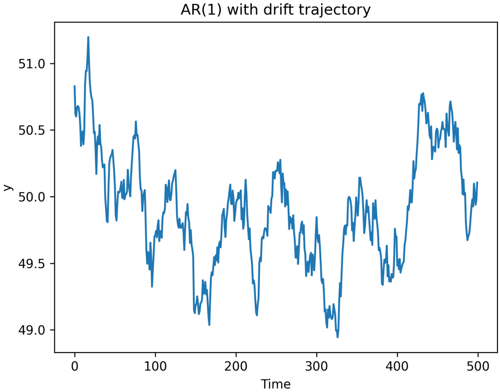
Posterior predictive density

$$f(\tilde{y}_i|\tilde{x}_i, \mathrm{X}) = \int_{\Theta} f(\tilde{y}_i|\theta, \tilde{x}_i) \cdot p(\theta|\mathrm{y}, \mathrm{X}) \cdot d\theta$$

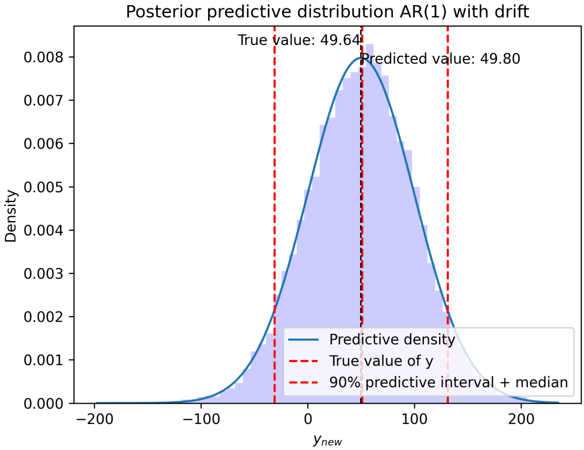
- 'Posterior weighted average' of sample density of new observation $\tilde{y}_i$
- More complete summary than predictive sets (shape, multimodality etc.)

Example: Time series process

$y_t = \mu + \phi y_{t-1} + \epsilon_t$, with $\epsilon_t \sim N(0, \sigma^2)$, $t = 1, ..., T$



Example (Cont.): Probabilistic prediction of y_{new}



Introduction to conformal prediction

Conformal prediction (CP): Set estimation

- Mainly **post-processing** approach to construct **predictive sets** $C_{n+1} : \chi \to \{\text{subsets of } \mathcal{Y}\}$, e.g. $\chi = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$, for any model prediction[3]
- **Distribution-free** and model agnostic approach
- Assuming **exchangeability** of training data with calibration data guarantees coverage of constructed predictive bands

[3] J. Lei and L. Wasserman. "Distribution-free prediction bands for non-parametric regression". In: Journal of the Royal Statistical Society: Series B: Statistical Methodology (2014), pp. 71–96.

~~Love~~ ~~Attention~~ Exchangeability is all you need

Definition

A sequence of random variables $X_1, X_2, X_3, \ldots$ is exchangeable if for any finite permutation σ of the indices $1, 2, 3, \ldots$ the joint probability distribution of the permuted sequence

$$X_{\sigma(1)}, X_{\sigma(2)}, X_{\sigma(3)}, \ldots$$

is the same as the joint probability distribution of the original sequence.

- Weaker assumption than 'independent and identically distributed'

Conformal prediction - Some theory

Given a (fixed) model $\pi_y(x) \approx p(y|x)$, a set of exchangeable calibration examples, $(x_i, y_i), \dots, (x_n, y_n)$ and a test example, x_{n+1}, construct confidence set $C(x_{n+1}) \subseteq [K]$ of labels that contains the true labels y_{n+1} with high probability:[4]

$$p(y_{n+1} \in C(x_{n+1})) \geq 1 - \alpha \qquad \text{(conformal coverage guarantee)}$$

- Coverage guarantee is marginal (vs. conditional) across examples and calibration sets
- $\alpha \in [0, 1]$ is a user-specified confidence level independent of data distribution and model

[4] V. Vovk, A. Gammerman, and C. Saunders. "Machine-learning applications of algorithmic randomness". In: International Conference on Machine Learning (1999), pp. 444–453.

Instructions for Conformal Prediction

For a general input x and output y (not necessarily discrete):

1. Identify a heuristic notion of uncertainty using the pre-trained model.
2. Define the non-conformal score function $s(x, y) \in \mathbb{R}$ (Larger scores encode worse agreement between x and y)
3. Compute $\hat{q}$ as the $\lceil (n+1) \cdot (1-\alpha) \rceil$ quantile of the calibration scores $s(x_1, y_1), \ldots, s(x_n, y_n)$
4. Use this quantile to form the prediction sets for new examples: $C(x_{n+1}) = \{y : s(x_{n+1}, y) \leq \hat{q}\}$

Examples: (Non-) conformity score functions

Continuous/Ordered response ('Regression'):

- $s(x_i, y_i) = \frac{|y_i - \hat{f}(x_i)|}{\sigma(x_i)}$ ('heteroscedastic/adaptive')
- $s(x_i, y_i) = p(y_i | x_i, y_{(-i)}, x_{(-i)})$ ('Conformal Bayes')

Discrete response ('Classification'):

- $s_i = 1 - \hat{f}(x_i)_{y_i}$ ('high if softmax output of true class low')

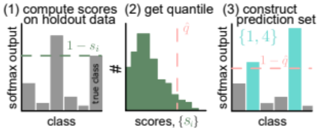
Some common approaches

- Split conformal prediction (CP)
- Full CP
- Adaptive (i.e. per observation) CP
- Mondrian (class-conditional) CP
- Conformal predictive distribution

Idea for classification problems

- CP works by making 'hypotheses' as to value of label y of test object X_{test}
- Hypothesis to test: hypothetical example ('candidate'), (X_{test}, y_{hyp}), was drawn i.i.d from the same distribution as the training examples.
- Compute p-value of this hypothesis
- Reject those hypotheses whose p-value is less than the significance level ϵ
- **The labels of the hypotheses we could not reject constitute the prediction set**

Illustration of CP for classification[5]




```python
# 1: get conformal scores. n = calib_Y.shape[0]
cal_smx = model(calib_X).softmax(dim=1).numpy()
cal_scores = 1-cal_smx[np.arange(n),cal_labels]
# 2: get adjusted quantile
q_level = np.ceil((n+1)*(1-alpha))/n
qhat = np.quantile(cal_scores, q_level, method='higher')
val_smx = model(val_X).softmax(dim=1).numpy()
prediction_sets = val_smx >= (1-qhat) # 3: form prediction sets
```

[5] A. N. Angelopoulos and S. Bates. "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification". In: arXiv:2107.07511 (2021).

Idea for regression problems

- We observe both $X_i \in \mathcal{X}$ and $Y_i \in \mathbb{R}$, $i = 1, \ldots, n$, and want a prediction set for Y_{n+1} based on X_{n+1}.
- Suppose that $\hat{f}_n(x)$ is any point predictor, trained on (X_i, Y_i), $i = 1, \ldots, n$
- Define (absolute) residuals made on training set, $R_i = |Y_i - \hat{f}_n(X_i)|$, $i = 1, \ldots, n$ let $\hat{q}_n = \lceil (1-\alpha)(n+1) \rceil$ smallest of $R_1, \ldots, R_n$
- Define the prediction set to be $\hat{C}_n(x) = \{y : |y - \hat{f}_n(x)| \leq \hat{q}_n\}$, or $\hat{C}_n(x) = [\hat{f}_n(x) - \hat{q}_n, \hat{f}_n(x) + \hat{q}_n]$

Hands-on demo

https://github.developer.allianz.io/CDO-AAC/uncertainty_quantification

Summary

- Model agnostic method for uncertainty quantification
- Quickly growing research field with applications beyond supervised learning
- MAPIE package: `https://github.com/scikit-learn-contrib/MAPIE`
- Collection of articles, code etc.: `https://github.com/valeman/awesome-conformal-prediction`
- Venn-ABERS calibration for binary and multiclass classification `https://github.com/ip200/venn-abers`

References

[1] A. N. Angelopoulos and S. Bates. "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification". In: arXiv:2107.07511 (2021).

[2] C. Gruber et al. "Sources of Uncertainty in Machine Learning - A Statisticians' View". In: arXiv:2305.16703 (2023).

[3] C. Guo et al. "On Calibration of Modern Neural Networks". In: Proceedings of the 34 th International Conference on Machine Learni
70 (2017).

[4] J. Lei and L. Wasserman. "Distribution-free prediction bands for non-parametric regression". In: Journal of the Royal Statistical Society: Series B: Statistical Methodo
(2014), pp. 71–96.

[5] V. Vovk, A. Gammerman, and C. Saunders. "Machine-learning