Tutorial on Conformal Predictive Distributions

Paolo Toccaceli

Centre for Reliable Machine Learning
Royal Holloway, University of London

Contents of the tutorial

- **Conformal Predictive Distributions** (Vovk et al., 2017) are a novel approach to estimating the probability distribution of a continuous variable that depends on a number of features.
- CPDs probabilities correspond to long-term relative frequencies (within statistical fluctuation) under minimal assumptions.
  - CPDs require only that the data be generated independently by an unknown but fixed distribution.
  - No assumption on the type of distribution
  - No need of a prior
- Outline of the tutorial
  - Motivation, context
  - Predictive Distribution, Conformal Predictive Distribution
  - KRRPM
  - Evaluation of PD and real-life examples

The roots of prior-free predictive distributions

- Suppose data $\mathcal{D}$ is generated by a known distribution with unknown parameter θ, which we want to estimate.
- Bayesian statistics can provide a distribution for the parameter
  - Estimating a probability distribution comes naturally to Bayesian methods
    $$p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$$
  - but you have to specify a prior $p(\theta)$.
- Frequentist statistics provides Confidence Intervals
  - Given $\alpha \in [0, 1]$ coverage probability, we compute an interval
    $$L(\mathcal{D}, \alpha), U(\mathcal{D}, \alpha)$$
    that contains the actual θ a fraction α of the time.
- $p(\theta|\mathcal{D})$ does not make sense in a strict frequentist framework.
- But frequentists recognized that predictive distributions would be useful!

Prior-free posterior distributions

- "The Holy Grail of parametric statistics" (Efron, 2010)
- Early attempts to arrive at prior-free posterior distributions can be traced back to Fisher's fiducial approach in the 1930s.
  - Not completely formalized; controversial; referred to as "Fisher's biggest blunder".
- There is now a resurgence of interest in the topic.
- Conformal Predictive Distributions are part of this trend.

Predictive Distributions: our setting

- Unrestricted randomness, i.i.d. data
  - $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ generated independently by fixed, but unknown $P(X, Y)$
- Intuitively, we seek an $F(y, x)$ that has the properties of a Cumulative Distribution Function in y.

$$F(y, x, \mathcal{D}) = P\{Y \leq y | X = x)$$

As long as $(x_i, y_i) \sim P(X, Y)$, the intervals $(-\infty, y)$ contain y_i with relative frequency $F(y, x_i)$

A property of CDF

- Let $F_X()$ be the (Cumulative) Distribution Function of the Random Variable X.

$$F_X(x) = P\{X \leq x\}$$

- The RV that is obtained by evaluating the F_X on the RV X is uniformly distributed.

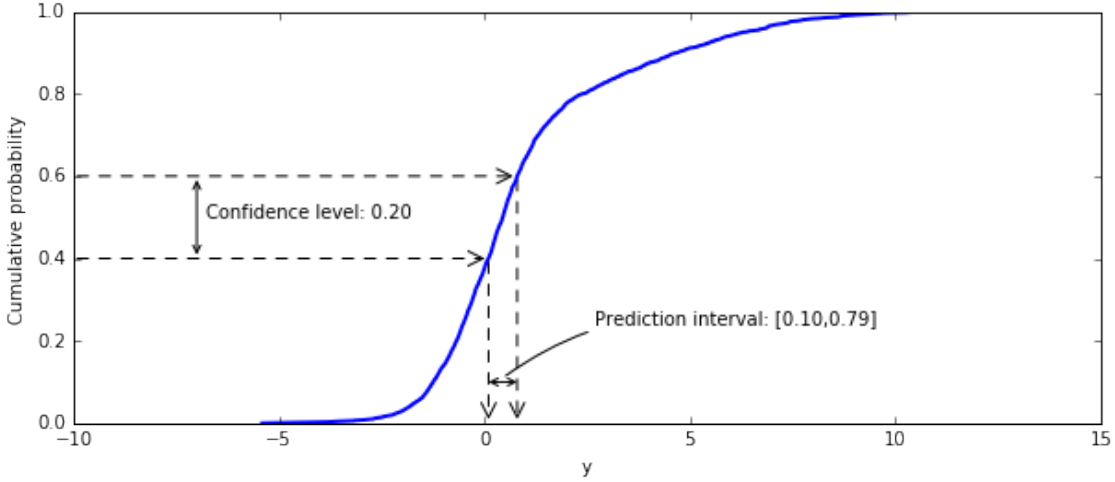
$$F_X(X) \sim U(0,1)$$

Predictive Distributions

- Suppose we have observations $z_1, \ldots, z_n$ from a set $\mathbf{Z} = \mathbf{X} \times \mathbb{R}$ and a test object $x_{n+1} \in \mathbf{X}$
- Let's call Predictive Distribution a function $Q(z_1, \ldots, z_n, (x_{n+1}, y))$ that
  - for any choice of training sequence $z_1, \ldots, z_n$ and any choice of test object x_{n+1} has the following properties of a CDF:
    - $Q(z_1, \ldots, z_n, (x_{n+1}, y))$ is monotonically increasing in y
    - $\lim_{y \to -\infty} Q(z_1, \ldots, z_n, (x_{n+1}, y)) = 0$
    - $\lim_{y \to +\infty} Q(z_1, \ldots, z_n, (x_{n+1}, y)) = 1$.
  - for any joint probability distribution P on $\mathbf{Z}$,
    - $Q(z_1, \ldots, z_n, z_{n+1}) \sim U$ when $(z_1, \ldots, z_{n+1}) \sim P^{n+1}$.
- This definition omits some technicalities to keep things simple.

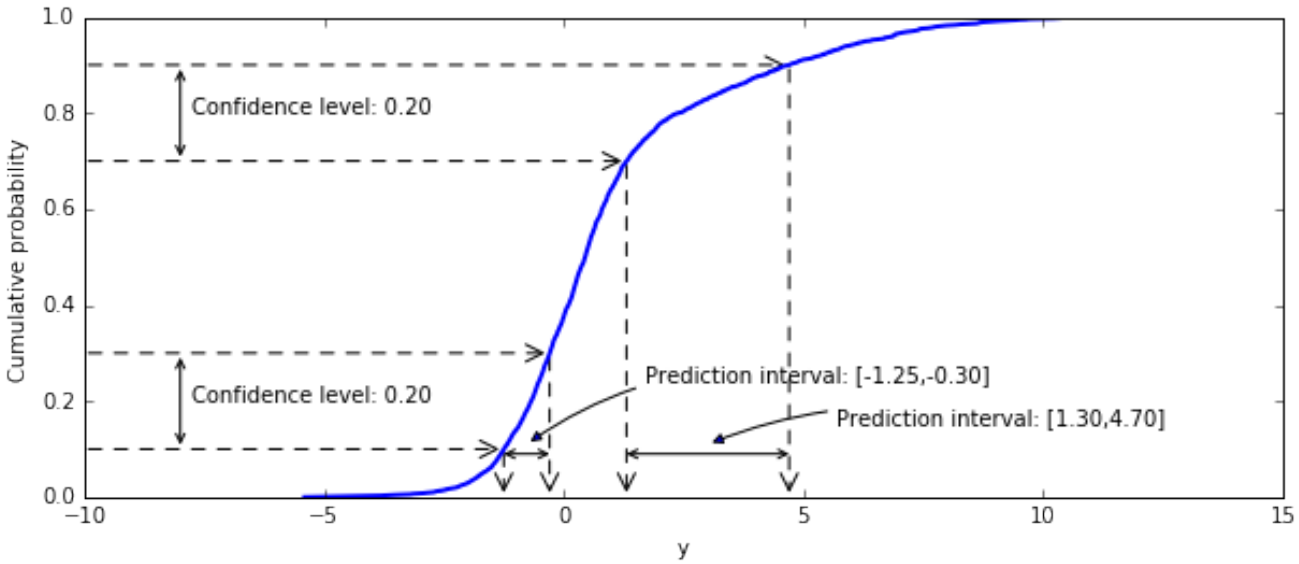
Predictive Distributions

- The "uniformity" property confers the ability to produce prediction intervals with **guaranteed coverage**
- Guaranteed coverage is the key property of predictive distributions
  - We can choose a confidence level α and we can read, off the predictive distribution, intervals of y in which the actual value falls with rate α (barring statistical fluctuation).



Predictive Distribution Interpretation

- Note that one can choose different prediction intervals for a given confidence level.



- One can choose the narrowest
  - i.e. where the slope of the predictive distribution is largest
- Or one around the median (previous slide)

An aside: Why CDF?

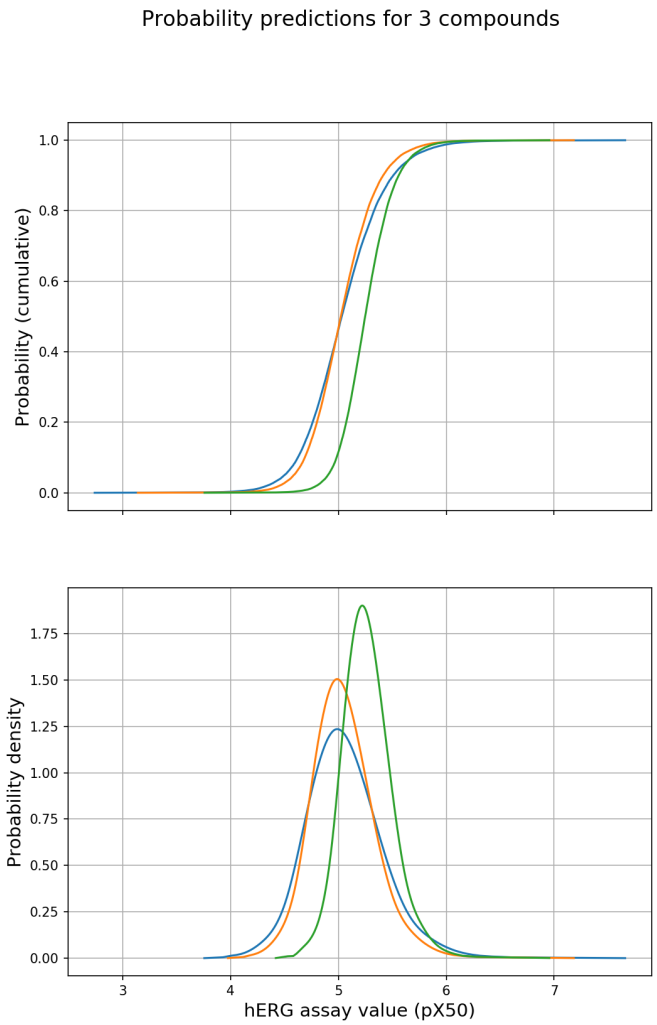
- The ECDF is guaranteed to converge to CDF for i.i.d. data (Glivenko-Cantelli theorem)
- The ECDF converges fast to the CDF (Dvoretzky–Kiefer–Wolfowitz inequality)
- To obtain a PDF you have to solve an ill-posed (unstable) problem:

$$\int h(x-t)f(t)dt = F(t) \qquad \text{where } h(x) \text{ is the step function}$$

  - stable: a small variation in the right-side $F(x)$ results in a small change in the solution $f(x)$

An aside: Why CDF?

- Pros
  - Computing an ECDF from data is straightforward.
  - You can read probabilities easily off the chart.
- Cons
  - Perceptually challenging to evaluate density
  - You can't find the mode immediately.



Conformal Predictive Distributions

- It is possible to obtain Predictive Distributions by using a variant of Conformal Prediction for Regression.
- We define as conformity measure a function

$$A(z_1, \ldots, z_{n+1})$$

invariant w.r.t. permutations of the first n arguments.

- Given (x_{n+1}, y), we compute conformity scores α_i^y as:

$$\begin{aligned} \alpha_i^y &:= A(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n, (x_{n+1}, y), z_i), \qquad i = 1, \ldots, n, \\ \alpha_{n+1}^y &:= A(z_1, \ldots, z_n, (x_{n+1}, y)). \end{aligned}$$

- Subject to some conditions on $A()$, the predictive distribution is then

$$Q(z_1, \ldots, z_n, (x_{n+1}, y)) := \frac{1}{n+1}\left(\left|\left\{i = 1, \ldots, n+1 \mid \alpha_i^y < \alpha_{n+1}^y\right\}\right|\right)$$

Requirements on the conformity measure

- Q() can be viewed as the estimate of probability under i.i.d. of drawing an observation with a smaller value of the CM than the hypothetical observation (x_{n+1}, y).
  It is as if we were testing the null hypothesis of i.i.d. using α as test statistic and computing the p-value Q().

- In contrast to CP, not all functions with the specified invariance property are conformity measures that result in valid predictive distributions.

- For the resulting Q() to have the properties of CDF, the conformity measure must be such that:
  - $\alpha^y_{n+1} - \alpha^y_i$ is a monotonically increasing function of $y \in \mathbb{R}$
  - $\lim_{y \to \pm\infty}(\alpha^y_{n+1} - \alpha^y_i) = \pm\infty$

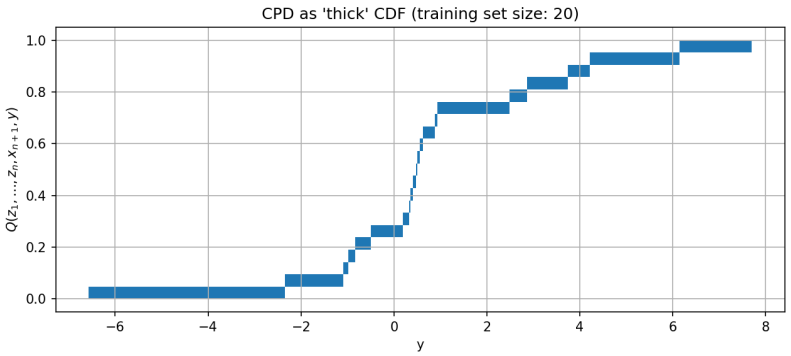
- A simple example is the $y - \hat{y}_{n+1}$ where $\hat{y}_{n+1}$ is the estimate obtained with K nearest neighbours regression.

Some details

- The definition was simplified for the sake of clarity.
- The properly defined CPD has a randomness element

$$Q(z_1, \ldots, z_n, (x_{n+1}, y)) := \frac{1}{n+1}\left(\left|\left\{i = 1, \ldots, n+1 \mid \alpha_i^y < \alpha_{n+1}^y\right\}\right|\right) + \frac{\tau}{n+1}\left(\left|\left\{i = 1, \ldots, n+1 \mid \alpha_i^y = \alpha_{n+1}^y\right\}\right|\right)$$

  where $\tau \sim U(0,1)$
- Informally, it's a "thick" distribution function, but the thickness is $\frac{1}{n+1}$, so it matters little as soon as you have a reasonably sized training set



Kernel Ridge Regression Predictive Machine

- Let's use Kernel Ridge Regression

$$\hat{y}_{n+1} := k'(K + aI)^{-1} Y$$

where $k_i = \mathcal{K}(x_i, x_{n+1})$, $K_{i,j} := \mathcal{K}(x_i, x_j)$, $i, j = 1, \ldots, n$.
Unfortunately, $y - \hat{y}_{n+1}$ is not a proper conformity measure. It fails to produce a strictly increasing function in y for high-leverage objects.

- Another possibility is to include the test object with the hypothetical label in the training set of the KRR.

$$\hat{\bar{y}}_{n+1} := \bar{k}'(\bar{K} + aI)^{-1} \bar{Y}$$

where $\bar{k}_i = \mathcal{K}(x_i, x_{n+1})$, $\bar{K}_{i,j} := \mathcal{K}(x_i, x_j)$, $\bar{y}_i = y_i$, $i, j = 1, \ldots, n+1$.
This too fails to guarantee a strictly increasing function in y.

Kernel Ridge Regression Predictive Machine

- A conformity measure is obtained as:

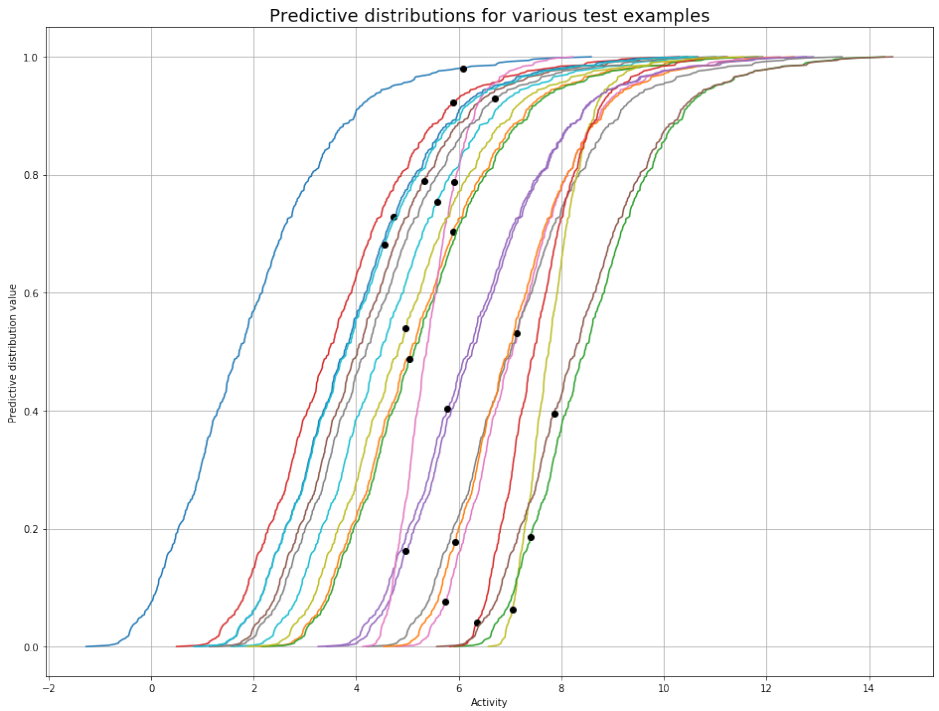
$$\frac{y_{n+1} - \hat{\bar{y}}_{n+1}}{\sqrt{1 - \bar{h}_{n+1}}}$$

  where $\bar{h}_{n+1}$ is the element $\bar{H}_{n+1,n+1}$ of the hat matrix $\bar{H} := (\bar{K} + aI)^{-1}\bar{K}$

- There exists an explicit form of the resulting KRRPM.
  It can be implemented in a way that avoids recomputing from scratch the hat matrix for every test object.

Application of KRRPM to Chemoinformatics

- Data set from a study on an enzyme
  - 1368 compounds
  - 68 features (PhysChem properties)
- Training set: 1000 observations, randomly sampled
- KRRPM using Laplace kernel



Desirable properties of PD

- **Validity**
  The predicted probabilities correspond to the long-term relative frequencies
- **Specificity** (a.k.a. Sharpness)
  The intervals for a given probability are as narrow as possible
- You can always make a valid predictive distribution: just output for all test objects the same PD, the ECDF of the label. But this would be a terrible forecast as it would have no specificity.
- CPDs guarantee validity under i.i.d.
  One can concentrate on improving specificity

Evaluating Predictive Distributions

- For probabilistic predictions, there are some established "scores"
  - Brier loss: $\frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2$
  - Log loss: $\frac{1}{n}\sum_{i=1}^{n}(o_i \log p_i + (1 - o_i)\log(1 - p_i))$
- These apply to tasks in which predictions must assign probabilities to a set of mutually exclusive **discrete** outcomes.
- They can be used on predictive distributions but they do not evaluate the PDs in their entirety.
- There are metrics and diagnostic tools for PDs
  - PIT, CRPS
  - Validity plot, Interval boxplots

Assessing validity

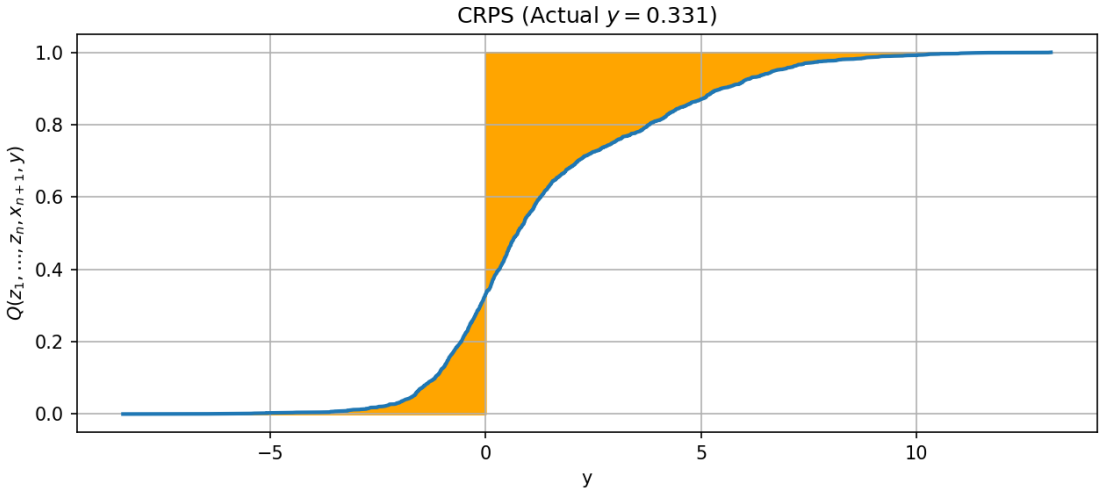
- Probability Integral Transform (PIT): evaluate $F_i()$ on the actual label y_i
- If the predictions $F_i(y)$ are ideal, $F_i(y_i)$ are variates from a $U(0,1)$ distribution.
- The PIT can be used to check validity.
- The histogram of the PIT should be as flat as possible.
- Perhaps better, the ECDF of the PIT should be as close as possible to the (0,0)-(1,1) diagonal
  - Kolmogorov-Smirnov statistic and K-S test

Continuous Ranked Probability Score (CRPS)

- The quadratic measure of discrepancy between the forecast CDF $F(y, x)$ and the "ideal forecast CDF" given the scalar observation y

$$\mathrm{CRPS}(F, x, y) = \int_{\mathbb{R}} [F(t, x) - \mathbb{I}(t \geq y)]^2 \, dt$$

where $\mathbb{I}()$ is the indicator function.

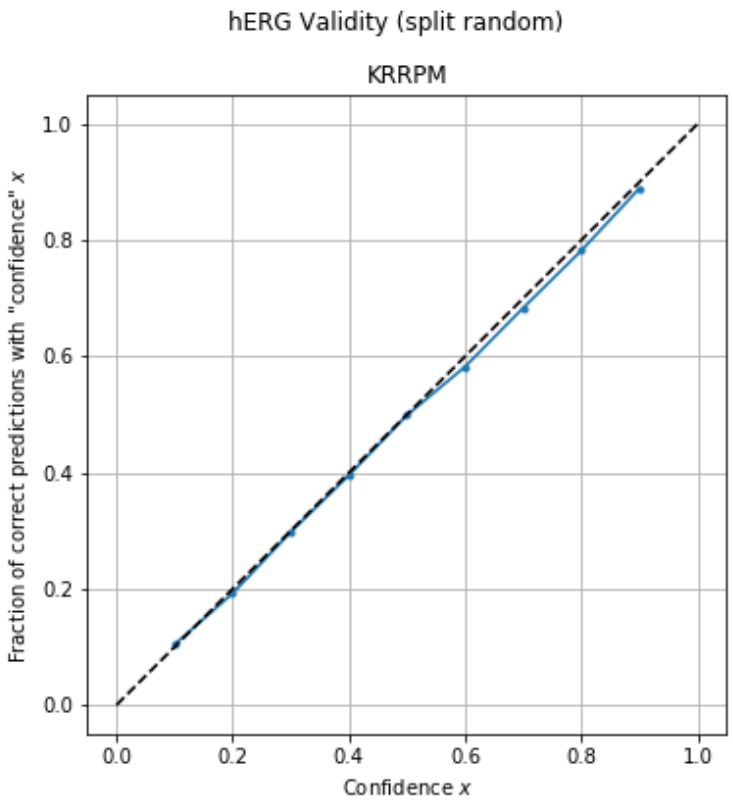


- For a number of predictions, one takes the average:

$$\overline{\mathrm{CRPS}}(F) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{CRPS}(F, x_i, y_i)$$

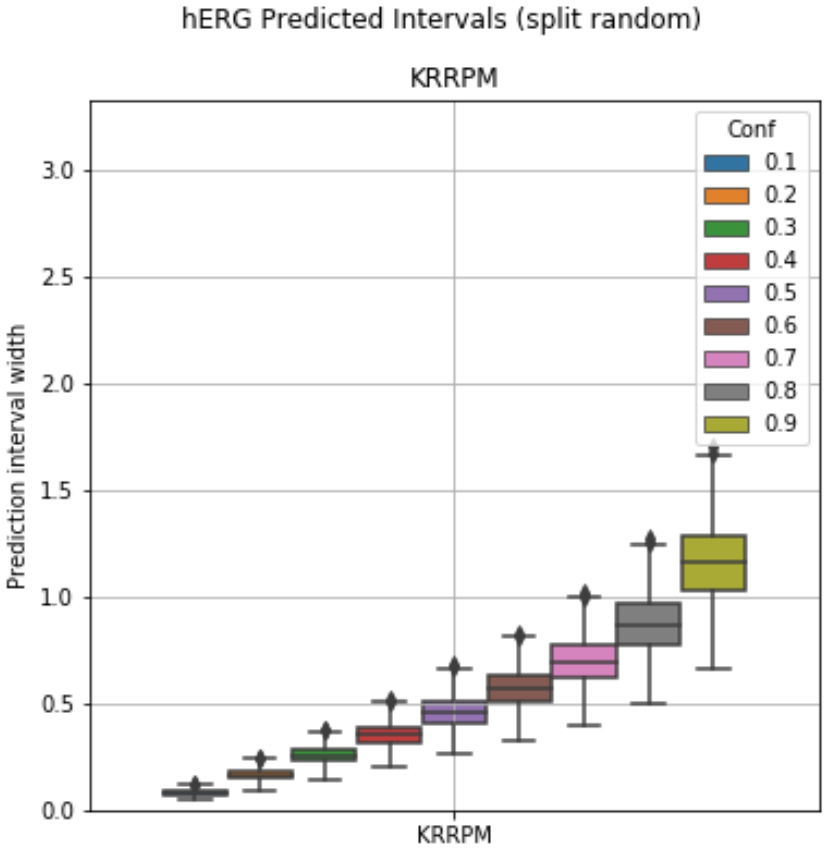
Practical diagnostic tools: Validity plot

- Validity plot: actual coverage vs. confidence
- For all the confidence values of interest (e.g. $0.1, 0.2, \ldots, 0.9$)
  - compute the intervals for the objects in the validation set
  - compute the relative frequency of "interval contains actual label"
- The relative frequency should be close to the confidence

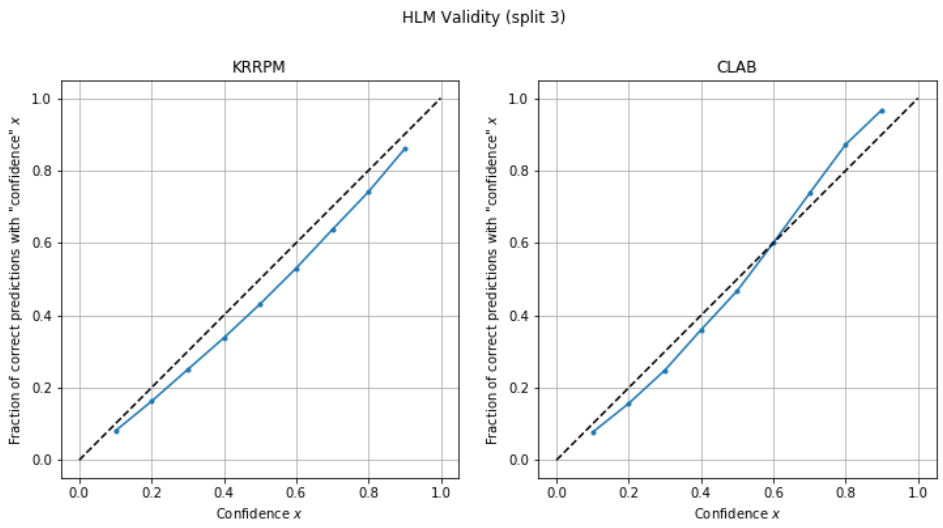


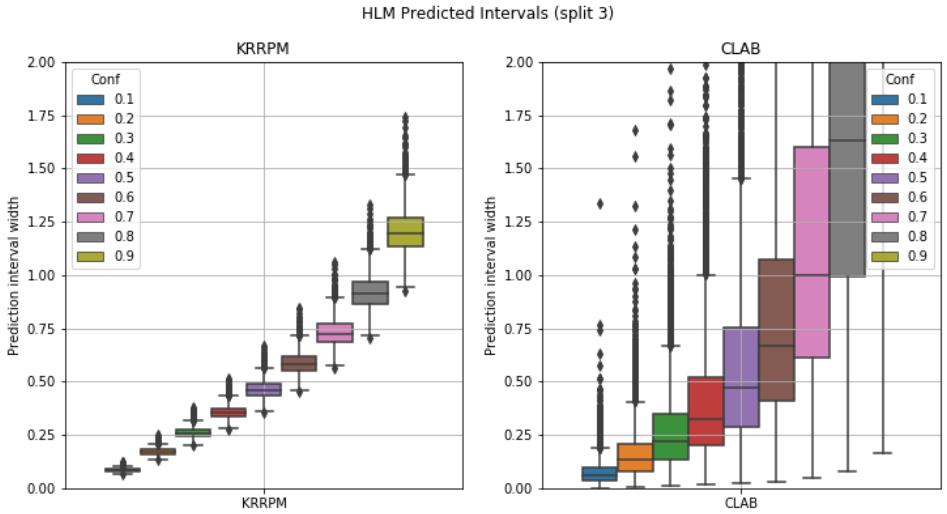
Practical diagnostic tools: Interval boxplots

- Descriptive statistics of the intervals
- The narrower the intervals, the more useful the predictions.



A real-life comparison





Observations

- The validity guarantee rests on the i.i.d. assumption.
  - $\{(x_1, y_1), \ldots, (x_n, y_n)\} \sim P^n(X, Y)$, where $(x_i, y_i) \sim P(X, Y)$
  - It is a minimal assumption of regularity
  - But it can be violated very easily in practice!
    - $P(X, Y) = P(X)\, P(Y|X)$
    - $P(X)$ varies: covariate shift
    - $P(Y|X)$ varies: in high-dimensional spaces, one visits just a small portion
- You should not assess validity and efficiency separately.
  - There is a trade-off between the two

Summary

- Conformal Predictive Distributions offer a prior-free, non-parametric way to estimate the distribution of random variable Y for an object x, based on previous data $(x_1, y_1), \ldots, (x_n, y_n)$.
- CPDs have a validity guarantee under a minimal assumption of test and training data being i.i.d.
  - CPD require only that the data be generated independently by an unknown but fixed distribution.
- KRRPM uses the flexible and regularized method of Kernel Ridge Regression to generate Conformal Predictive Distributions.

Acknowledgements

- Some material was produced as part of research funded via the AstraZeneca grant "Automated Chemical Synthesis" under the guidance of Prof. Alexander Gammerman and Dr. Claus Bendtsen

References

V. Vovk, J. Shen, V. Manokhin, M. Xie.
Nonparametric predictive distributions based on conformal prediction
Proceedings of Machine Learning Research, 60:82-102, 2017.

V. Vovk, I. Nouretdinov, V. Manokhin, A. Gammerman.
Conformal Predictive Distribution with Kernels
Braverman Readings in Machine Learning, Springer, 2018.

N. Hjort, T. Schweder.
Confidence distributions and related themes
Journal of Statistical Planning and Inference, 195(2018)1-13.

J. Shen, R. Liu, M. Xie.
Prediction with Confidence–A general framework for predictive inference
Journal of Statistical Planning and Inference, 195(2018)126-140.