(Inductive) Conformal prediction:

Basics and recent advances for multi-variate regression

Work by *Sébastien Destercke*, Soundouss Messoudi and Sylvain Rousseau

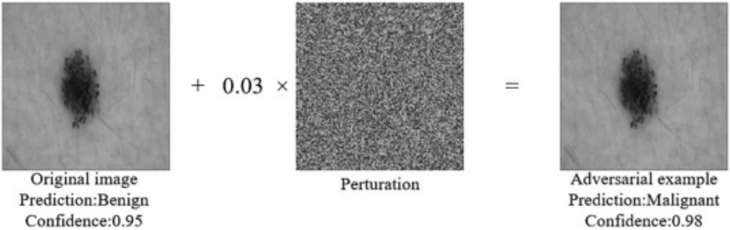
within Soundouss thesis.

CNRS, Heudiasyc

WG R&A

Why conformal prediction?

- Generic classifier trying to predict Y from observing X does not provide strong statistical guarantee



- Such guarantee can be seen as requiring predictions $\hat{Y}$ to contain the observed truth y with a given (coverage) probability, i.e., $1 - \epsilon$

$$P(y \in \hat{Y}) > 1 - \epsilon$$

  with ϵ a specified error rate.
- Conformal prediction allows one to have it with weak assumptions

Transuctive vs inductive conformal prediction

CP started as a transductive, online learning setting:

- Observe the (exchangeable[1]) sequence

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

- Observe x_{n+1}, predict the possible y_{n+1}.

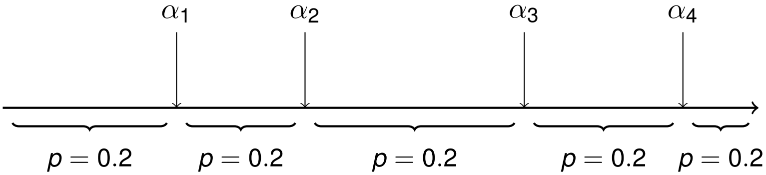
Then was proposed in an **inductive setting**, where

- you keep some calibration data $\mathcal{D}_{cal}$
- you learn a model $h : X \to Y$ from a (disjoint) training set $\mathcal{D}_{tr}$
- you predict new observations using those.

[1] Future inferences do not depend on the order of observation

(Very) basic ideas of inductive conformal prediction

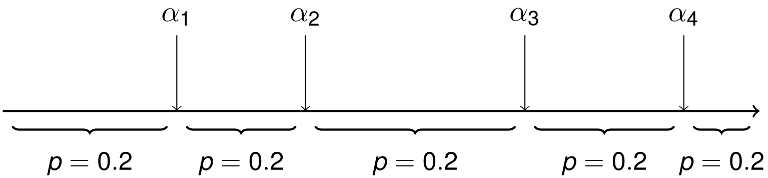
- Calibrating observations $(x_i, y_i) \in \mathcal{D}_{cal}$ issued from distribution **Q**
- For each (x_i, y_i), we associate a conformity score α_i depending on (x_i, y_i) and $h(x_i)$ (e.g., model score given to y_i)
- The lower α_i, the better.
- Assume we have 4 calibrating observations with $\alpha_1 < \ldots < \alpha_4$



- The probability of a next item score falling into a bin is $1/|\mathcal{D}_{cal}|$

Predicting a new item: classification

- We observe x. Completing it with possible class y gives α_y
- Given

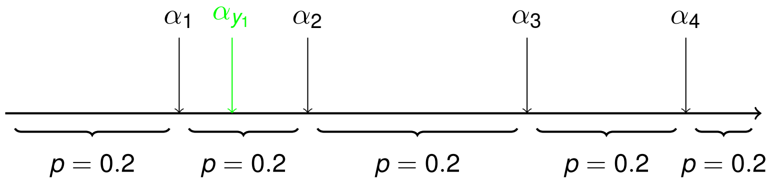


  We have $P(\alpha_y \leq \alpha_4) = 0.8$.
- Retaining all classes $y \in Y$ with $\alpha_y \leq \alpha_i$ as $\hat{Y}$ will ensure

$$P(y \in \hat{Y}) = i/(n+1)$$

Predicting a new item: classification

- We observe x. Completing it with possible class y gives α_y
- Given



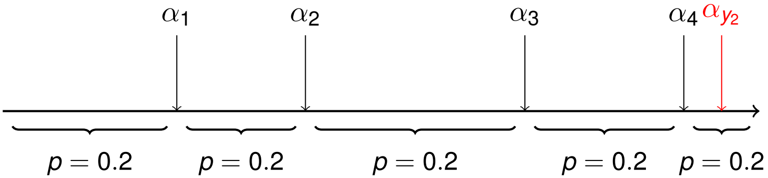
We have $P(\alpha_y \leq \alpha_4) = 0.8$.

- Retaining all classes $y \in Y$ with $\alpha_y \leq \alpha_i$ as $\hat{Y}$ will ensure

$$P(y \in \hat{Y}) = i/(n+1)$$

Predicting a new item: classification

- We observe x. Completing it with possible class y gives α_y
- Given



  We have $P(\alpha_y \leq \alpha_4) = 0.8$.
- Retaining all classes $y \in Y$ with $\alpha_y \leq \alpha_i$ as $\hat{Y}$ will ensure

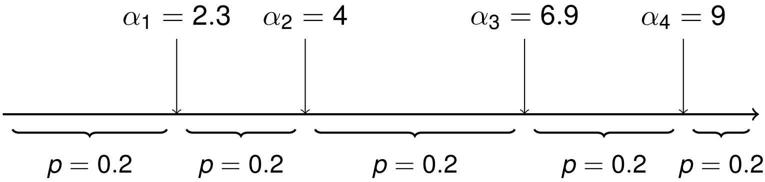
$$P(y \in \hat{Y}) = i/(n+1)$$

Predicting a new item: regression

- We observe x, but Y is now a continuous variable
- Conformity score based on h=distance, for example

$$\alpha_i = |y_i - h(x_i)|$$

- Given



We sill have $P(\alpha_y \leq \alpha_4) = 0.8$.

- Meaning that $P(|y - h(x_i)| \leq \alpha_j) = j/(n+1)$. Take all values within

$$[h(x_i) - \alpha_j, h(x_i) + \alpha_j]$$

Predicting a new item: regression continued

- if $\alpha_i = |y_i - h(x_i)|$ then predicted interval

$$[h(x_i) - \alpha_j, h(x_i) + \alpha_j]$$

  is of constant length
- solution: use of normalised conformal scores

$$\alpha_i = \frac{|y_i - h(x_i)|}{\sigma_i}$$

  where σ_i is an estimation of the local error
- In this case, intervals

$$[h(x_i) - \alpha_j \sigma_i, h(x_i) + \alpha_j \sigma_i]$$

  depend on the local behaviour

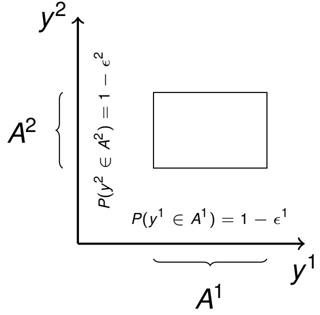
Multi-variate regression and conformal prediction

- The input X is unchanged
- We now observe multi-variate outputs $y \in \mathbb{R}^m$

how can we adapt conformal approaches to have multivariate prediction regions with guaranteed error rates?

A first idea

- Fixing marginal error rates ϵ^j for each dimension j, and apply previous recipes dimension-wise
- How to relate it to the global error?



- We only have

$$\max(\epsilon^1 + \epsilon^2 - 1, 0) \leq 1 - \epsilon^g = P(y \in A^1 \cap A^2) \leq \min(1 - \epsilon^1, 1 - \epsilon^2)$$

Copulas to the help

- Finding the relation between ϵ^g and ϵ^i
- Taking the product

$$P(y \in A^1 \cap A^2) = P(y^1 \in A^1)P(y^2 \in A^2)$$

$\leftrightarrow$ Bonferroni multi-test correction
$\rightarrow$ leads to poorly calibrated results

- One idea: as $P(|y^k - h^k(x_i)| \leq \alpha_j)$ can be seen as a cumulative distribution F^k $\rightarrow$ use tools combining such cumulative distributions $\rightarrow$

Copula

Just a little bit of details

- Given uniform r.v. U^k
- A function $C : [0,1]^m \to [0,1]$
- A copula C describes their joint distribution

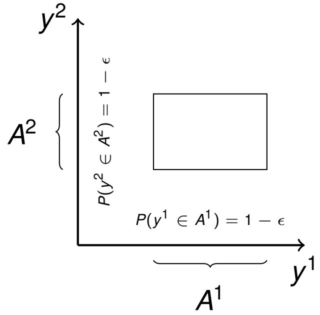
$$C(u_1, u_2, \ldots, u_d) = P(U_1 \leq u_1, U_2 \leq u_2, \ldots, U_d \leq u_d)$$

- See $C(u_1, u_2, \ldots, u_d)$ as giving ϵ^g in function of $\epsilon^1, \ldots, \epsilon^m$
- Learn it from calibrating data.

In practice

- In general, we assume $\epsilon^1 = \ldots = \epsilon^m$
- Find the value ϵ such that

$$\epsilon^g = C(\underbrace{\epsilon, \ldots, \epsilon}_{m \text{ times}})$$



A second idea

Copula idea work well in practice, yet

- The framework is essentially a combination of univariate inferences ($\rightarrow$ not "truly" multivariate)
- It does not capture potential dependencies depending on a covariate structure (hyper-cubes are axis-aligned)

$\rightarrow$ directly use a multi-variate conformal score

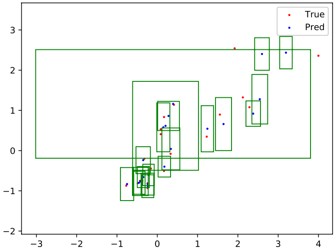
Proposed score

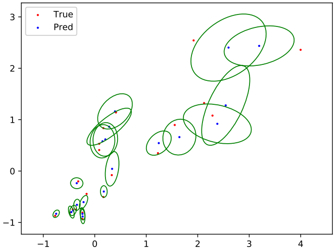
- We propose the following score given (x_i, y_i):

$$\alpha_i \propto (y_i - h(x^i)\Sigma_i(y_i - h(x^i))$$

  where Σ_i is a local covariance matrix.
- We essentially find a local ellipsoid region with guaranteed coverage
- Up to now, we estimate it by taking a regularized matrix estimated from neighbours

An illustration of the results





Some concluding remarks

- Easy framework to derive robust predictions
- Can differentiate to some extent ambiguity vs lack of information

But still a lot to do

- Ensure conditional coverage $P(y \in \hat{Y}|x) > 1 - \epsilon$
- deal with non-i.i.d./exchangeable cases (transfer learning, time series, ...)

References I