A comparative study of conformal prediction methods for valid uncertainty quantification in machine learning

Nicolas Dewolf

April 25, 2024

Introduction

Motivation

- Uncertainty is a fundamental notion.

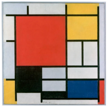
Motivation

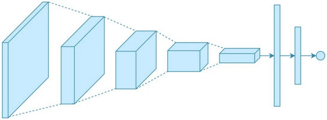
- Uncertainty is a fundamental notion.
- Sadly, it has became a secondary notion

Motivation

- Uncertainty is a fundamental notion.
- Sadly, it has became a secondary notion
- Conformal prediction tries to fix this issue.

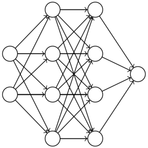
Predictions





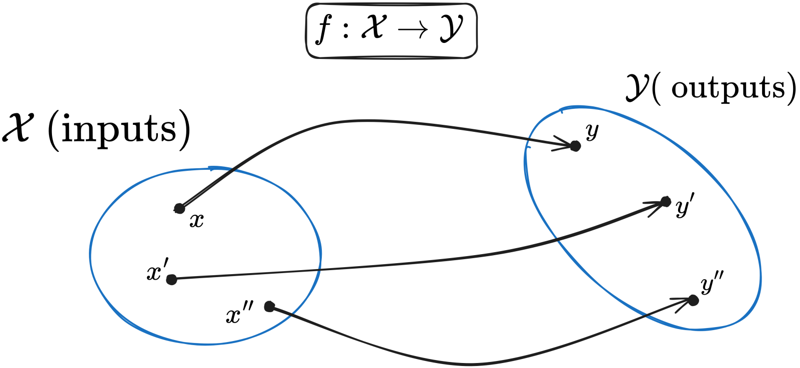
Mondriaan

Team 1	Team 2	Weather	Time
A	B	Rainy	Morning
A	D	Sunny	Morning
B	C	Sunny	Morning
D	C	Rainy	Evening
C	A	Foggy	Evening
A	C	Sunny	Evening



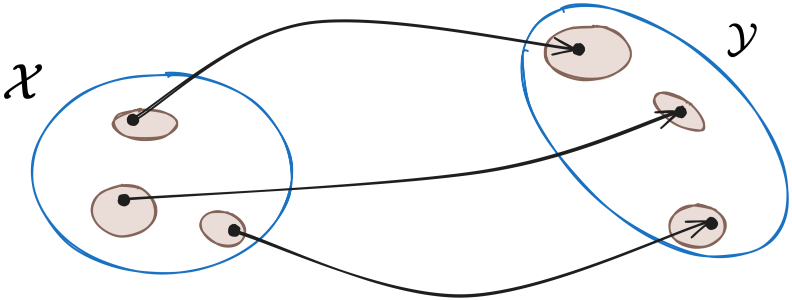
Points
1
2
0
1
0
4

Predictions



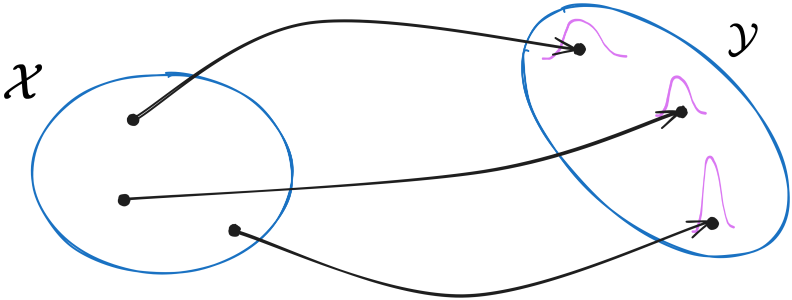
Measurement noise

$$f : \mathcal{X} \to \mathcal{Y}$$

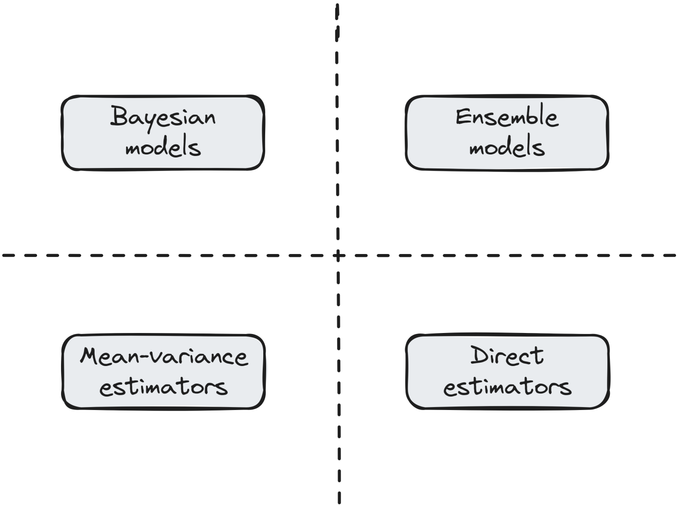


Stochasticity

$$f : \mathcal{X} \to \mathbb{P}(\mathcal{Y})$$



Models


Bayesian models
Ensemble models
Mean-variance estimators
Direct estimators

Confidence predictors

Modelling probability distributions might be too hard.

Confidence predictors

Modelling probability distributions might be too hard.

Confidence predictor

A (set-valued) function from feature tuples to (sets of) possible responses.

Confidence predictors

Modelling probability distributions might be too hard.

Confidence predictor

A (set-valued) function from feature tuples to (sets of) possible responses.

This is similar to confidence intervals in statistics.

Example: Classification

$$\longrightarrow \qquad P\Big(y \in \{2, 3, 9\}\Big) \geq 90\%$$

Example: Classification

$$\longrightarrow \quad P\Big(y \in \{2, 3, 9\}\Big) \geq 90\%$$

$$\longrightarrow \quad P\Big(y \in \{8\}\Big) \geq 90\%$$

Overview

1. Marginal validity

Overview

1. Marginal validity
2. Conditional validity

Overview

1. Marginal validity
2. Conditional validity
3. Clusterwise validity

Overview

1. Marginal validity
2. Conditional validity
3. Clusterwise validity
4. Future perspectives

Marginal Validity

Problems

Important limitations to standard techniques that make them unappealing (to ML practitioners):

- model limitations (e.g. linearity)

Problems

Important limitations to standard techniques that make them unappealing (to ML practitioners):

- model limitations (e.g. linearity),
- data assumptions (e.g. normality)

Problems

Important limitations to standard techniques that make them unappealing (to ML practitioners):

- model limitations (e.g. linearity),
- data assumptions (e.g. normality), and
- computational inefficiency (e.g. Bayesian inference).

Solution

Conformal prediction tries to overcome all of these issues:

- no model constraints

Solution

Conformal prediction tries to overcome all of these issues:

- no model constraints,
- weak data assumptions

Solution

Conformal prediction tries to overcome all of these issues:

- no model constraints,
- weak data assumptions,
- efficient implementations exist

Solution

Conformal prediction tries to overcome all of these issues:

- no model constraints,
- weak data assumptions,
- efficient implementations exist, and
- can incorporate other methodologies (e.g. online learning).

Exchangeability

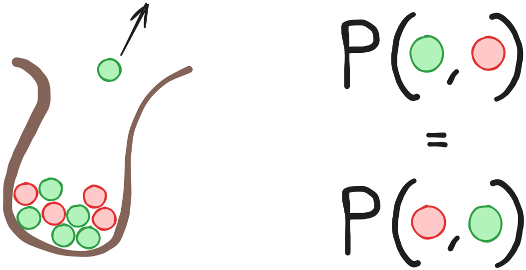
Exchangeability

If the probability of observing a data sequence is independent of its order, it is said to be **exchangeable**.

Exchangeability

Exchangeability

If the probability of observing a data sequence is independent of its order, it is said to be **exchangeable**.



Exchangeability

- Irrelevant order implies that the ranks of the data points are uniformly distributed

Exchangeability

- Irrelevant order implies that the ranks of the data points are uniformly distributed
- This is the working horse of my dissertation!

Nonconformity

Nonconformity measure

A function $A : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that assigns a *(nonconformity) score* to every data point.

Nonconformity

Nonconformity measure

A function $A : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that assigns a *(nonconformity) score* to every data point.

x	$\rho(x)$	y	$A(x,y)$
0.5	1	2.5	1.5
2.5	5	3	2
1	2	10	8

Example: Regression

Given a regression model $\rho : \mathcal{X} \to \mathbb{R}$, some typical nonconformity measures are:

Example: Regression

Given a regression model $\rho : \mathcal{X} \to \mathbb{R}$, some typical nonconformity measures are:

- Standard (residual) score:

$$A_{\text{res}}(x, y) := |\rho(x) - y|\,,$$

Example: Regression

Given a regression model $\rho : \mathcal{X} \to \mathbb{R}$, some typical nonconformity measures are:

- Standard (residual) score:

$$A_{\mathsf{res}}(x, y) := |\rho(x) - y|,$$

- Normalized (residual) score:

$$A^{\sigma}_{\mathsf{res}}(x, y) := \frac{|\rho(x) - y|}{\sigma(x)},$$

  where $\sigma : \mathcal{X} \to \mathbb{R}^+$ is an uncertainty estimate such as the standard deviation.

Conformal prediction

The (*inductive*) conformal prediction algorithm has the following simple workflow:

Conformal prediction

The (*inductive*) conformal prediction algorithm has the following simple workflow:

1. Choose a *calibration set* $\{(x_i, y_i)\}_{i \leq n}$, a *nonconformity measure* $A : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and a *significance level* $\alpha \in [0, 1]$.

Conformal prediction

The (*inductive*) conformal prediction algorithm has the following simple workflow:

1. Choose a *calibration set* $\{(x_i, y_i)\}_{i \leq n}$, a *nonconformity measure* $A : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and a *significance level* $\alpha \in [0, 1]$.
2. Calculate the score $a_i := A(x_i, y_i)$ for every calibration point.

Conformal prediction

The (*inductive*) conformal prediction algorithm has the following simple workflow:

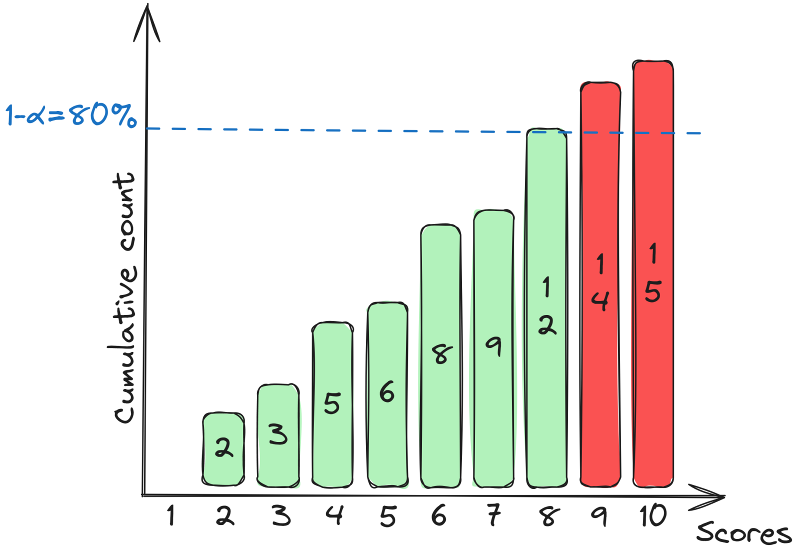
1. Choose a *calibration set* $\{(x_i, y_i)\}_{i \leq n}$, a *nonconformity measure* $A : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and a *significance level* $\alpha \in [0, 1]$.
2. Calculate the score $a_i := A(x_i, y_i)$ for every calibration point.
3. Determine the *critical score* $a^* := q_{(1-\alpha)(1+1/n)}\big(\{a_i\}_{i \leq n}\big)$.

Conformal prediction

The (*inductive*) conformal prediction algorithm has the following simple workflow:

1. Choose a *calibration set* $\{(x_i, y_i)\}_{i \leq n}$, a *nonconformity measure* $A : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and a *significance level* $\alpha \in [0, 1]$.
2. Calculate the score $a_i := A(x_i, y_i)$ for every calibration point.
3. Determine the *critical score* $a^* := q_{(1-\alpha)(1+1/n)}\big(\{a_i\}_{i \leq n}\big)$.
4. For a new x, include all y such that $A(x, y) \leq a^*$.

Conformal prediction



Statistical guarantees

Theorem (Conservative validity)

If the data is exchangeable, the conformal predictor is *(conservatively) valid*:

$$\mathrm{Prob}\big(Y \in \Gamma^{\alpha}(X)\big) \geq 1 - \alpha\,.$$

Statistical guarantees

Theorem (Conservative validity)

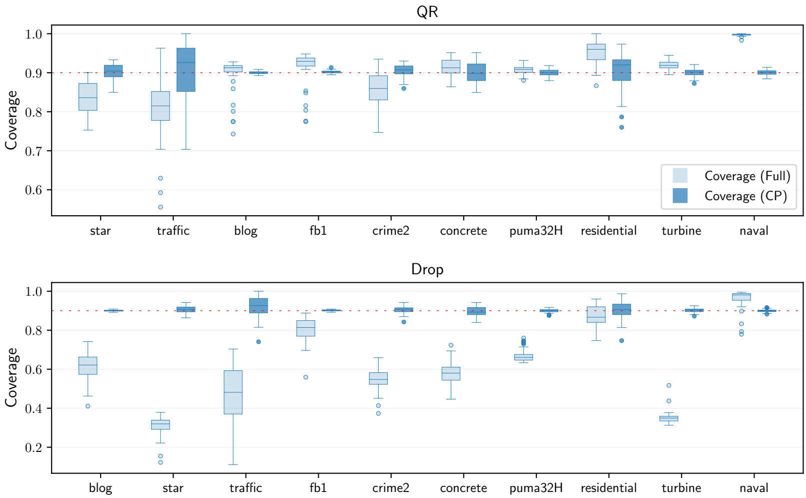
If the data is exchangeable, the conformal predictor is *(conservatively) valid*:

$$\mathrm{Prob}\big(Y \in \Gamma^{\alpha}(X)\big) \geq 1 - \alpha\,.$$

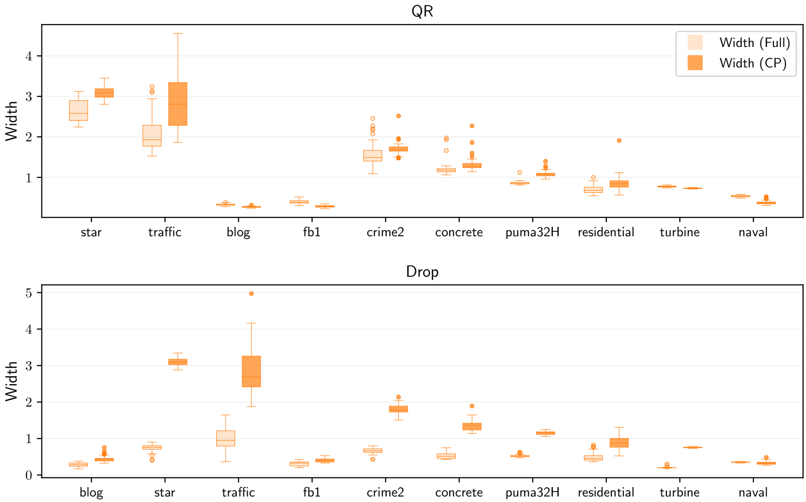
Theorem (Strict validity)

If the nonconformity scores are also distinct, the conformal predictor is *strictly valid*:

$$\mathrm{Prob}\big(Y \in \Gamma^{\alpha}(X)\big) = 1 - \alpha\,.$$

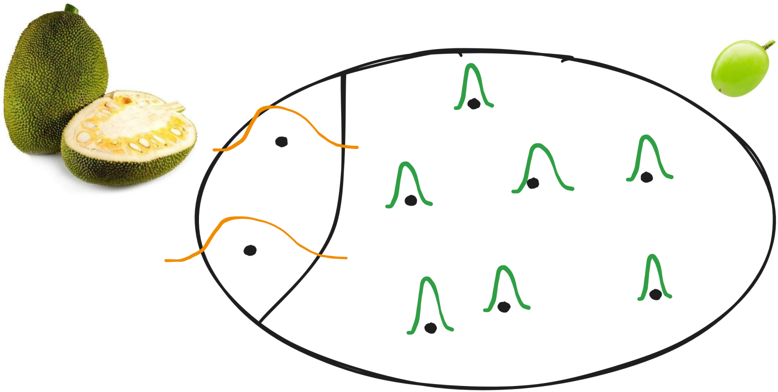

QR
Coverage
1.0
0.9
0.8
0.7
0.6
Coverage (Full)
Coverage (CP)
star
traffic
blog
fb1
crime2
concrete
puma32H
residential
turbine
naval
Drop
Coverage
1.0
0.8
0.6
0.4
0.2
blog
star
traffic
fb1
crime2
concrete
puma32H
residential
turbine
naval

Experiments

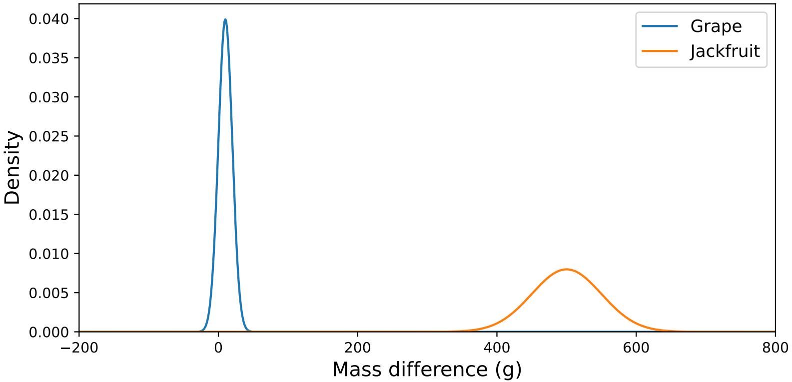


Conditional Validity

Cheating



Example: Weight prediction

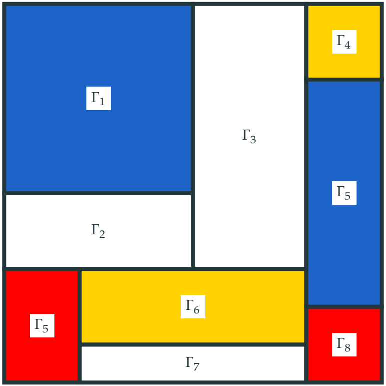


Mondriaan

Given a partition of the instance space

$$\kappa : \mathcal{X} \times \mathcal{Y} \rightarrow \{0, 1, \dots, n\},$$

we construct a model for each subgroup



Alternative

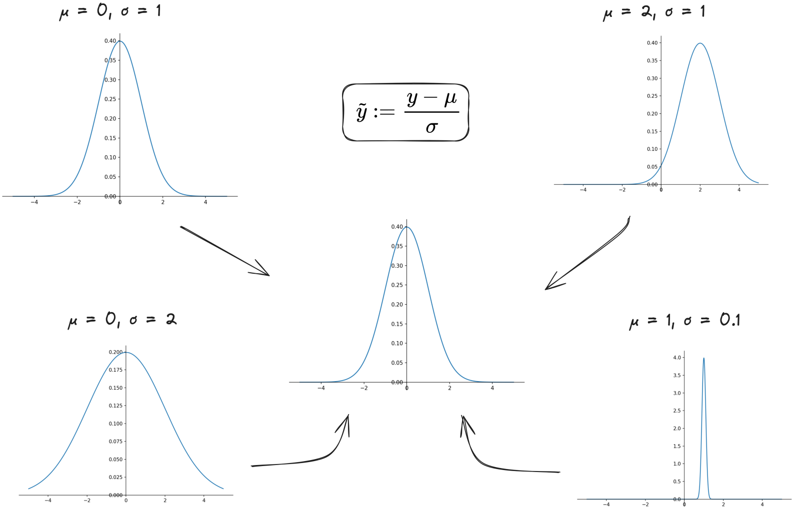
Can we approximate conditional validity with a single conformal predictor?

Pivots

Pivot

If the distribution of $f(X_\theta)$, with $X_\theta \sim P_\theta$, is independent of the parameter $\theta \in \Theta$, the function f is said to be **pivotal** for the family of distributions $\{P_\theta\}_{\theta \in \Theta}$.

Standardization



Contribution (Pivotal measure)

If the nonconformity measure is pivotal with respect to the classwise distributions, the conformal predictor is conditionally valid.

Independence

Contribution (Pivotal measure)

If the nonconformity measure is pivotal with respect to the classwise distributions, the conformal predictor is conditionally valid.

Intuition: We can combine data sets if they come from the same distribution.

Normalization

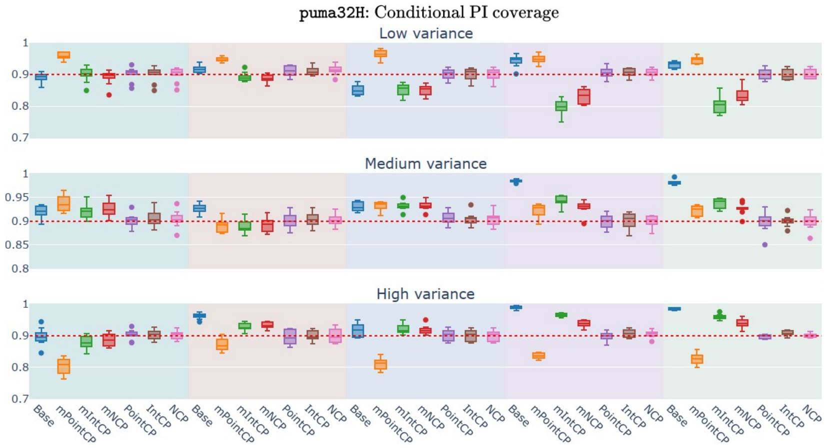
Contribution (Parametric form)

If the conditional distribution is of the form

$$f(y \mid x) = \frac{1}{\sigma(x)} g\left(\frac{y - \mu(x)}{\sigma(x)}\right),$$

the nonconformity measure A^{σ}_{res} gives a conditionally valid conformal predictor.

Experiments



Clusterwise Validity

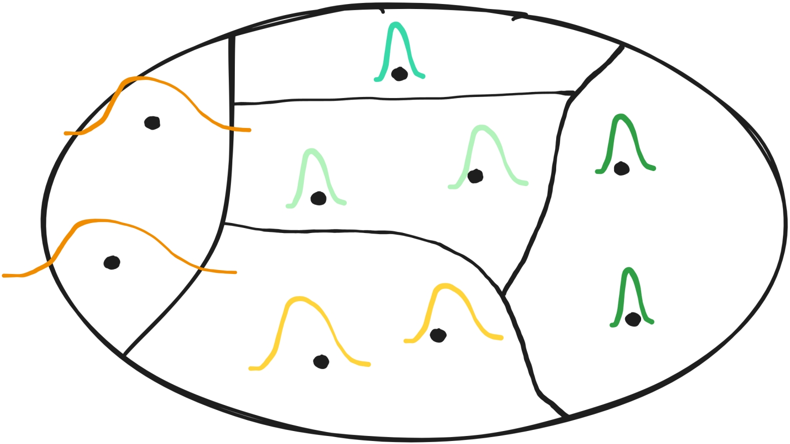
Remaining issue

- Mondrian approach: strong guarantees, but data required per class.

Remaining issue

- Mondrian approach: strong guarantees, but data required per class.
- non-Mondrian approach: guarantees (in pivotal scenario) but data required for correct models.

Clusterwise validity



Clusterwise validity

Theorem (Clusterwise validity)

The deviation from conditional validity is bounded by the *statistical diameter* of the cluster ω:

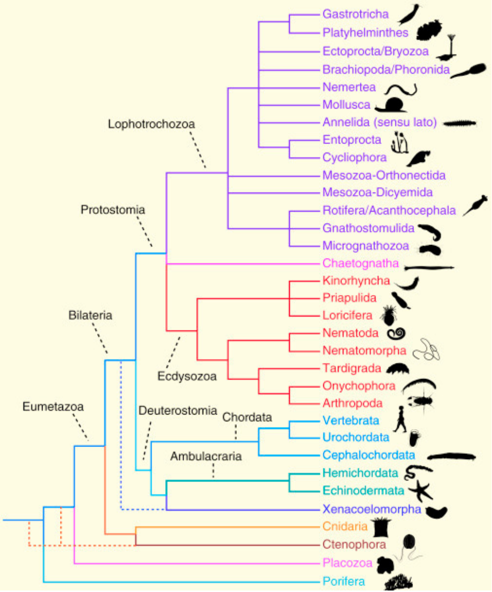
$$\text{Prob}\big(Y \in \Gamma^{\alpha}(X) \mid \kappa(X,Y) = c, c \in \omega\big) \geq 1 - \alpha - \max_{c' \in \omega} d(c, c').$$

Lipschitz

Contribution (Lipschitz continuity)

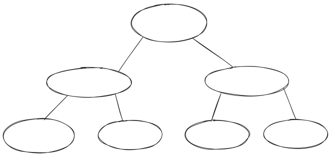
If the conditional distributions $P_{Y|X}$ depend smoothly on X, the clusterwise validity result remains valid.

Hierarchies

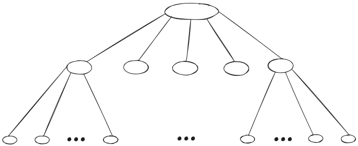


Extreme classification

Hierarchies can be too coarse!



vs.



Conclusion

- Conformal prediction is versatile and easy to use.

Conclusion

- Conformal prediction is versatile and easy to use.
- Conditional validity is important and can be achieved (approximately).

Conclusion

- Conformal prediction is versatile and easy to use.
- Conditional validity is important and can be achieved (approximately).
- Interpretation and usefulness of results is not always straightforward.

Future perspectives

Interesting possibilities:

- extreme classification

Future perspectives

Interesting possibilities:

- extreme classification,
- multivariate problems

Future perspectives

Interesting possibilities:

- extreme classification,
- multivariate problems, and
- time series.