Università di Pisa

Academic year 2022/2023

Conformal Prediction

Statistics for Data Science

Chiara De Nigris
586013

Emanuele Sabatini
637756

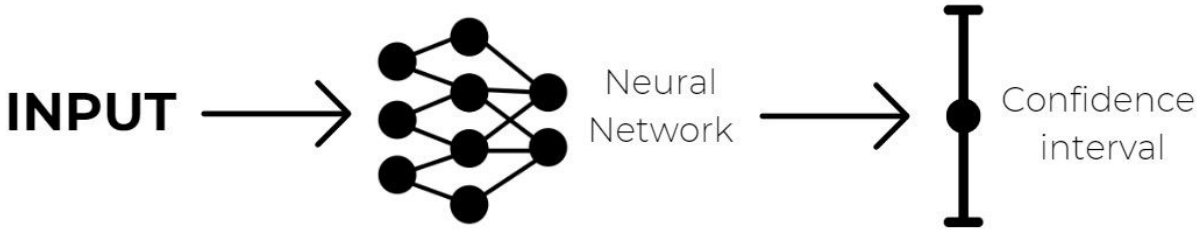
Michele Papucci
544376

Introduction

Conformal prediction is a paradigm for creating statistically finite and rigorous **confidence intervals** for model and data agnostic predictions, only assuming exchangeability of the data.

These sets are **guaranteed** to encompass the **ground truth** with a probability specified by the user.

Domain: **high-risk settings** which demand uncertainty quantification to avoid consequential model failures (e.g. medical diagnostics).



Conformal Prediction - Objectives

Conformal prediction takes **any heuristic notion of uncertainty** from **any model** and converts that notion into a rigorous one.

How does it work in practice?

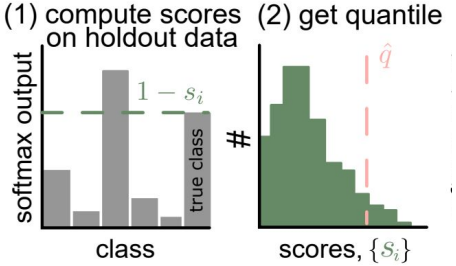
1. We identify a heuristic notion of uncertainty in the model (e.g., Softmax Scores);
2. We define a **score function** $s(X, Y) \in \mathbb{R}$ (Where a higher s value, means that Y isn't a good label for X) $s_1 = s(X_1, Y_1), \ldots, s_n = (X_n, Y_n)$
3. We compute $\hat{q}$ as $\frac{\lceil (n+1)(1-a) \rceil}{n}$ quantile of the calibration scores $\mathcal{T}(x) = \{y : s(X, y) \leq \hat{q}\}$
4. We use $\hat{q}$ to create prediction set for unseen inputs

Calculating $\hat{q}$

$\hat{q}$ is the threshold value that we use to select which of the scores produced by the model for each label is high enough to be included in the prediction set.
The value is dependent on a parameter α which represent the probability of error we can accept from the conformal prediction. It's the probability that the **ground truth label isn't included in the prediction set**.
To calculate $\hat{q}$ given α and a calibration set, we:

1. Input the calibration set into the model and obtain the scores for the **ground truth label**;
2. We calculate $\hat{q}$ as: $\frac{\lceil (n+1)(1-a) \rceil}{n}$

$\hat{q}$ is a value that, in the calibration set, each ground truth label score has a probability of 1 - α to be greater of.



Coverage Guarantee and Adaptiveness

The **conformal calibration coverage guarantee** theorem states that given $\hat{q}$ and $\mathcal{T}$ defined as we described, is proven that the **marginal coverage** holds:

$$P(Y_{n+1} \in \mathcal{T}(\mathcal{X}_{n+1})) \geq 1 - a$$

What this mean is that even for unseen evidences we have a probability of having the **ground truth label** in the prediction set greater or equal to 1 - α.

Adaptiveness is a key feature of Conformal Prediction, and is defined as the ability of the procedure to **generate larger prediction sets for harder inputs and smaller prediction sets for easier inputs**. Is also formalized by asking for the **conditional coverage**:

$$P(Y \in \mathcal{T}(\mathcal{X})|\mathcal{X}) \geq 1 - a$$

Conditional coverage is a stronger property than marginal coverage and is **not guaranteed** and it states that for any subset of the population the coverage should be greater than 1 - α.

Conformal Prediction - Image Classification

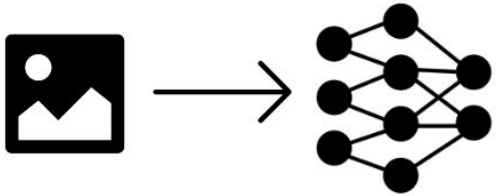
We applied the technique to an **image classification problem**. Given:

- a calibration set, $(X_1, Y_1), \ldots, (X_n, Y_n)$ where X are **Imagenet images** and Y classes;
- a fitted predicted model $\hat{f}$ (a neural network classifier) that outputs softmax scores for each class: $\hat{f}(x) \in [0,1]^K$ with K as the number of classes;

We then proceed to:

1. Create a calibration set containing 2500 images that were not used in training $\hat{f}$.
2. Choose an α value of 0,1. Which thanks thanks to the **conformal calibration coverage guarantee** gives us a probability of 0,9 of having the ground truth label in each prediction set.
3. We define our score function as: $s(X) = 1 - X$.
4. Calculate $\hat{q}$ on the **scores** obtained by passing the classification model outputs to .
5. Use $\hat{q}$ to do a series of prediction from a set of images not used for training and not used for the calibration set.
6. Run a series of validation test to see empirically how the coverage and adaptability performed in a real scenario.

Conformal Prediction - Image Classification



{class1, class2}

{ fox squirrel 0.99 }

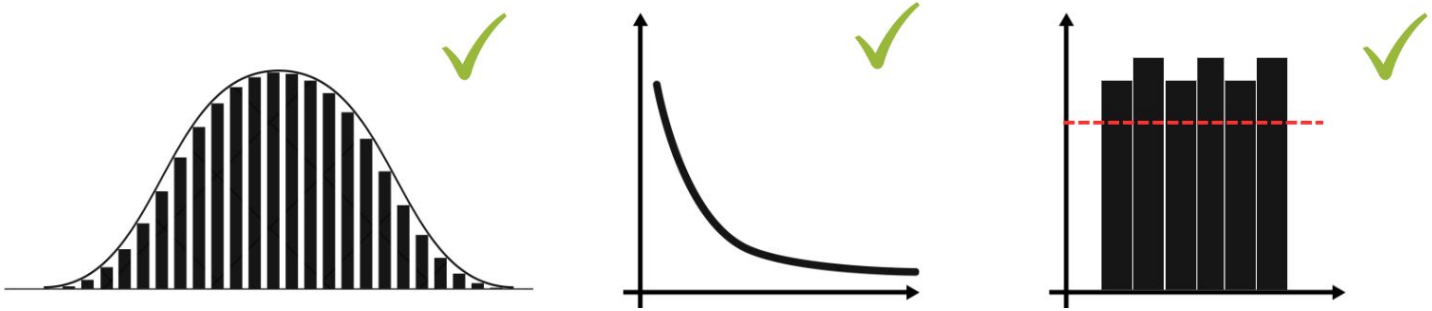
{ fox squirrel 0.82, gray fox, 0.03, bucket, 0.02, rain barrel 0.02 }

{ marmot 0.30, fox squirrel 0.22, mink 0.18, weasel 0.16, beaver 0.03, polecat 0.01 }

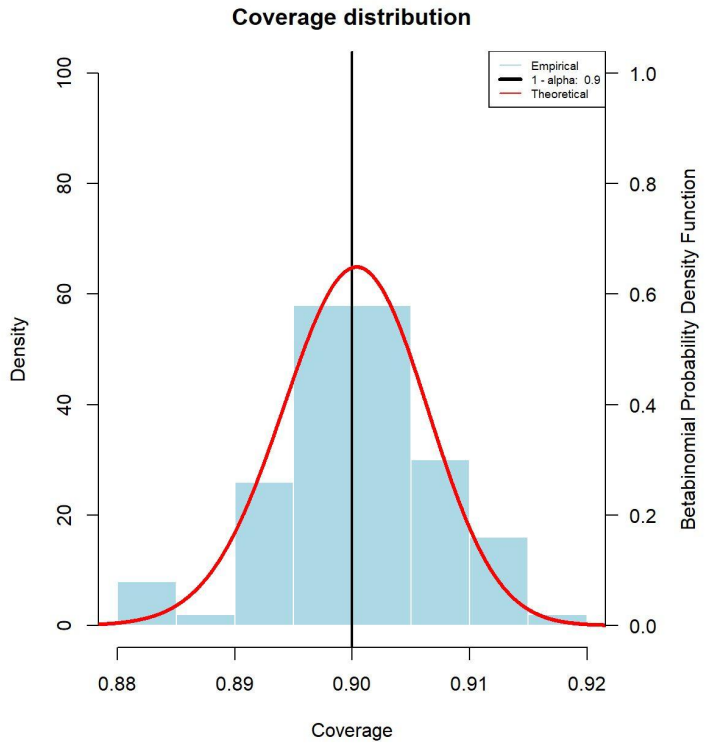
Evaluation

Lastly, we evaluated prediction sets checking three main aspects:

1. Coverage
2. Set size
3. Adaptiveness



Evaluation - Coverage

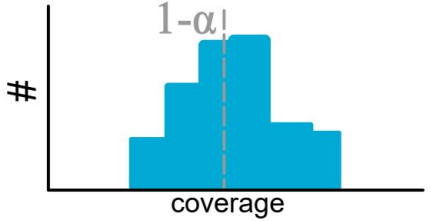


First step has been to assess whether the conformal procedure has the **correct coverage**. This can be accomplished by running the procedure R times, randomly splitting the data into a calibration and validation datasets, and then calculating the empirical average coverage:

$$C_j = \frac{1}{n'} \sum_{i=1}^{n'} \mathbb{1}\{Y_i^{(val)} \in \mathcal{T}(X_i^{(val)})\}$$

for $j = 1, \ldots, R$

Where n' is the size of validation set. A histogram of these R coverage values should look be centered at 1 – α.

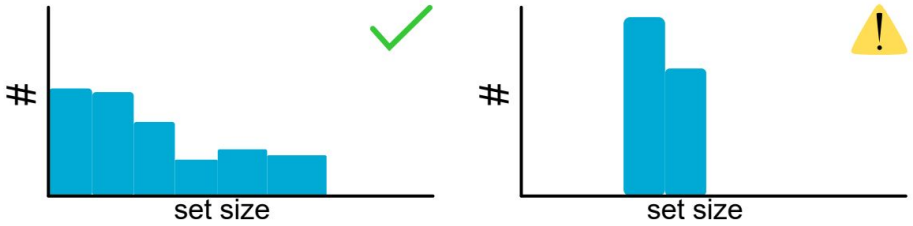


Evaluation - Set Size(I)

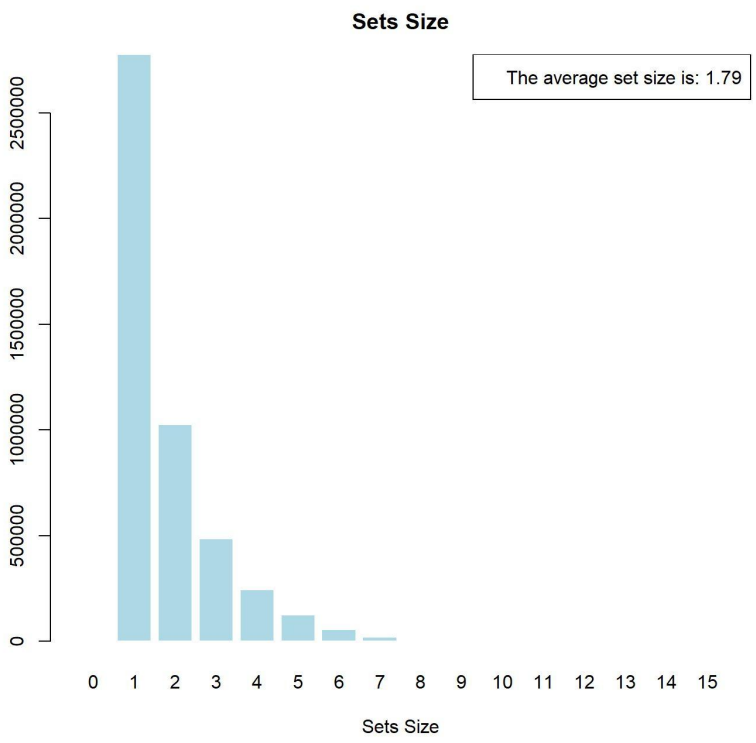
Randomically splitting data, we built prediction sets and evaluated their **dimension** plotting histograms of set sizes.

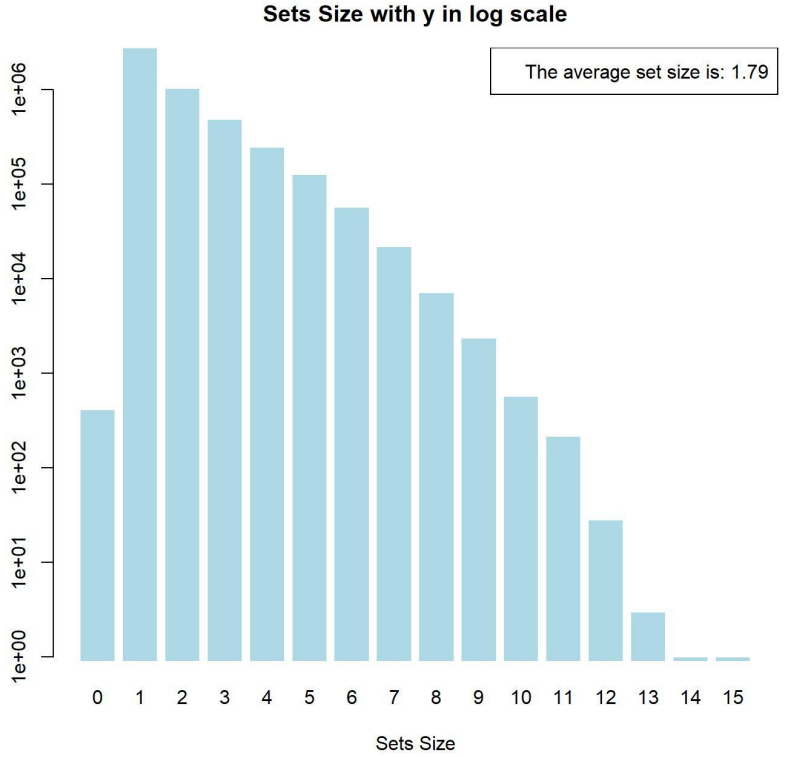
A large average set size indicates the conformal procedure is not very precise, indicating a possible problem with the score or underlying model.

The spread of the set sizes shows whether the **prediction sets properly adapts to the difficulty of examples**.



Evaluation - Set Size(II)

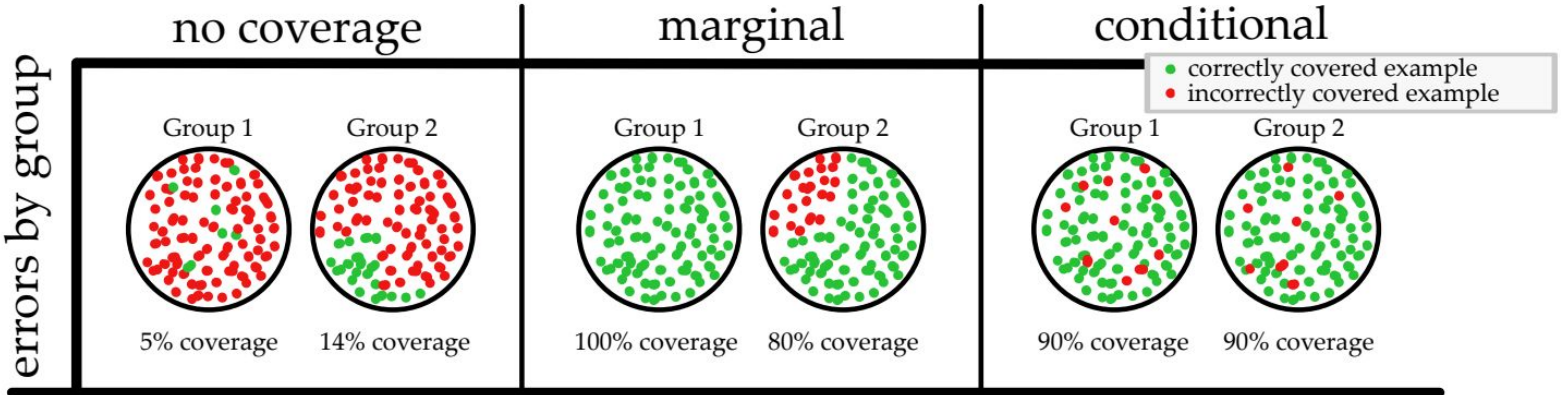




Evaluation - Adaptiveness(I)

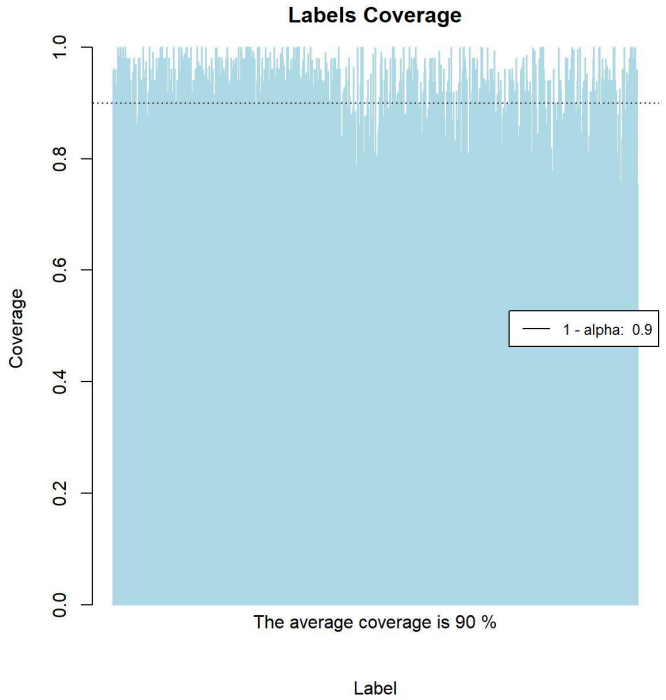
Adaptiveness: we want the procedure to return larger sets for harder inputs and smaller sets for easier inputs.

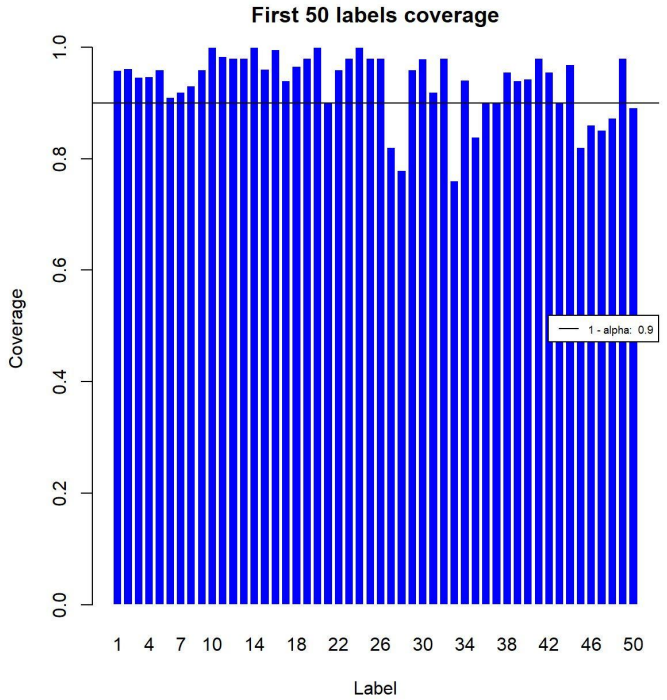
Adaptiveness is typically formalized by asking for the **conditional coverage** property which is not ensured by conformal prediction, where only **marginal coverage** is guaranteed to be achieved.



Evaluation - Adaptiveness(II)

In order to evaluate marginal coverage, we plot average coverage for each label.





Conformal Prediction - Image Classification

The prediction set is: ['jay']

Our prediction set is: [**jay**]

The prediction set is: ['impala', 'gazelle']

Our prediction set is: [hartebeest, **impala**, gazelle]

Conclusions - Key points

- ➔ **Conformal prediction** is a rigorous method to create confidence intervals for classification and regression task;
- ➔ The method is **model-** and **data-agnostic**;
- ➔ Conformal predictions **guarantees** that on average **the ground truth will be present in the prediction set** with a probability specified by the user;
- ➔ The prediction set is **informative**, its size implies how hard was classifying the inputs for the model.