Uncertainty Sets for Image Classifiers using Conformal Prediction

Team LezitNet

Bidit Sadhukhan

Srijan Rit

About the Paper

- ★ Authors: Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, Michael I. Jordan
- ★ Institution: University of California, Berkeley
- ★ Publication: ICLR 2021, The Ninth International Conference on Learning Representations

Contents

1 Introduction

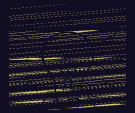
2 Why Uncertainty Quantification is necessary?

3 Algorithms of Platt's Scaling and Conformal Prediction

4 Introduction to APS and RAPS

5 Conclusion

Introduction



Imagine you are a doctor making a high-stakes medical decision based on diagnostic information from a computer vision classifier.

What would you want the classifier to output in order to make the best decision?

Do you think maximum likelihood diagnosis with an accompanying probability can always give the most essential piece of information?

Uncertainty Quantification

Which is a better response or prediction 

A. Classifier giving only the predicted class with maximum probability.

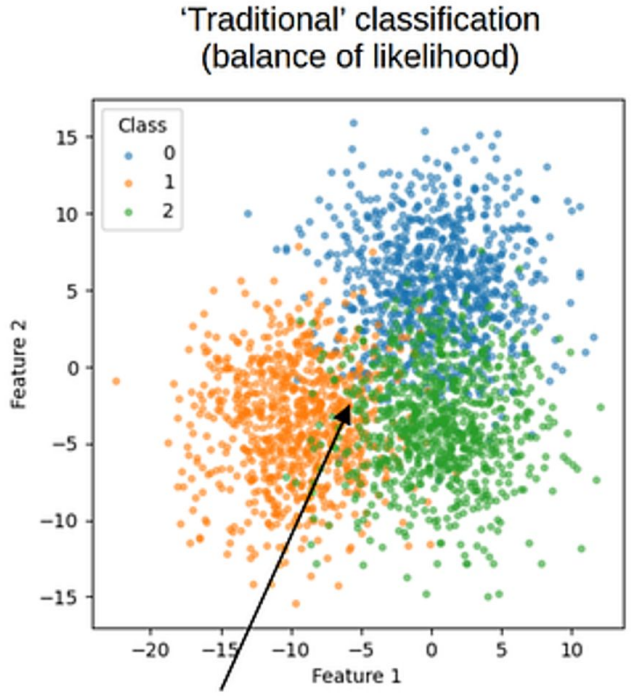
B. Classifier giving a set of predictions that probably covers the true diagnosis with a high probability (e.g., 90%).

Source: Conformal Prediction for Machine Learning Classification —From the Ground Up, Michael Allen

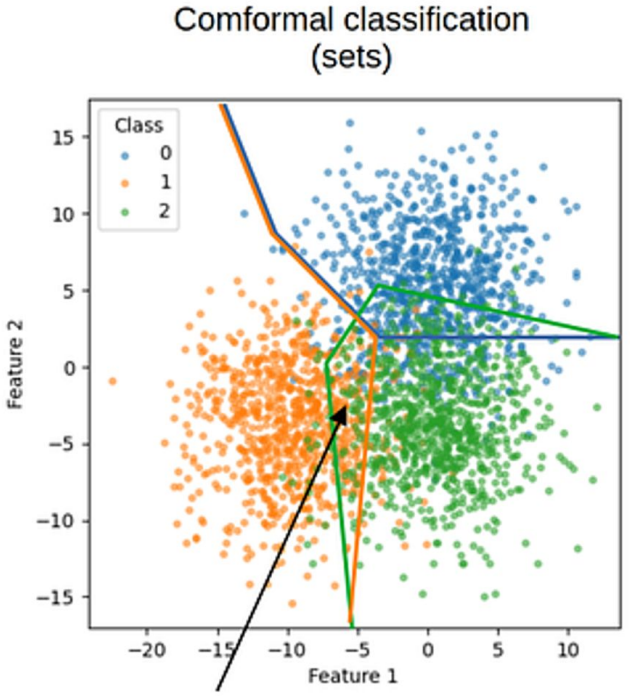
Uncertainty Quantification

Why should we care about uncertainty quantification 

- When we use model predictions to make decisions. How sure are we of those predictions? Is using just ‘most likely class’ good enough for the task we have?
- Instead of probabilities, we should focus on explaining the range of potential results and how confident the model is in each one.



Classification in overlap region based on highest probability: 'orange'



Classification in overlap region based on all relevant areas of coverage: {orange, green}

Source: https://christophmolnar.com/books/conformal-prediction/

Platt's Scaling

Platt scaling is a technique used to calibrate the output probabilities of a binary classifier, which means that it adjusts the probability scores so that they are more accurate and better represent the true likelihood of a particular class.

➢ Applications:

It is used in situations where the output probabilities of a classifier are not well-calibrated, meaning that they do not accurately reflect the true likelihood of a positive or negative outcome.

Platt's Scaling

Train a binary classifier on the training data

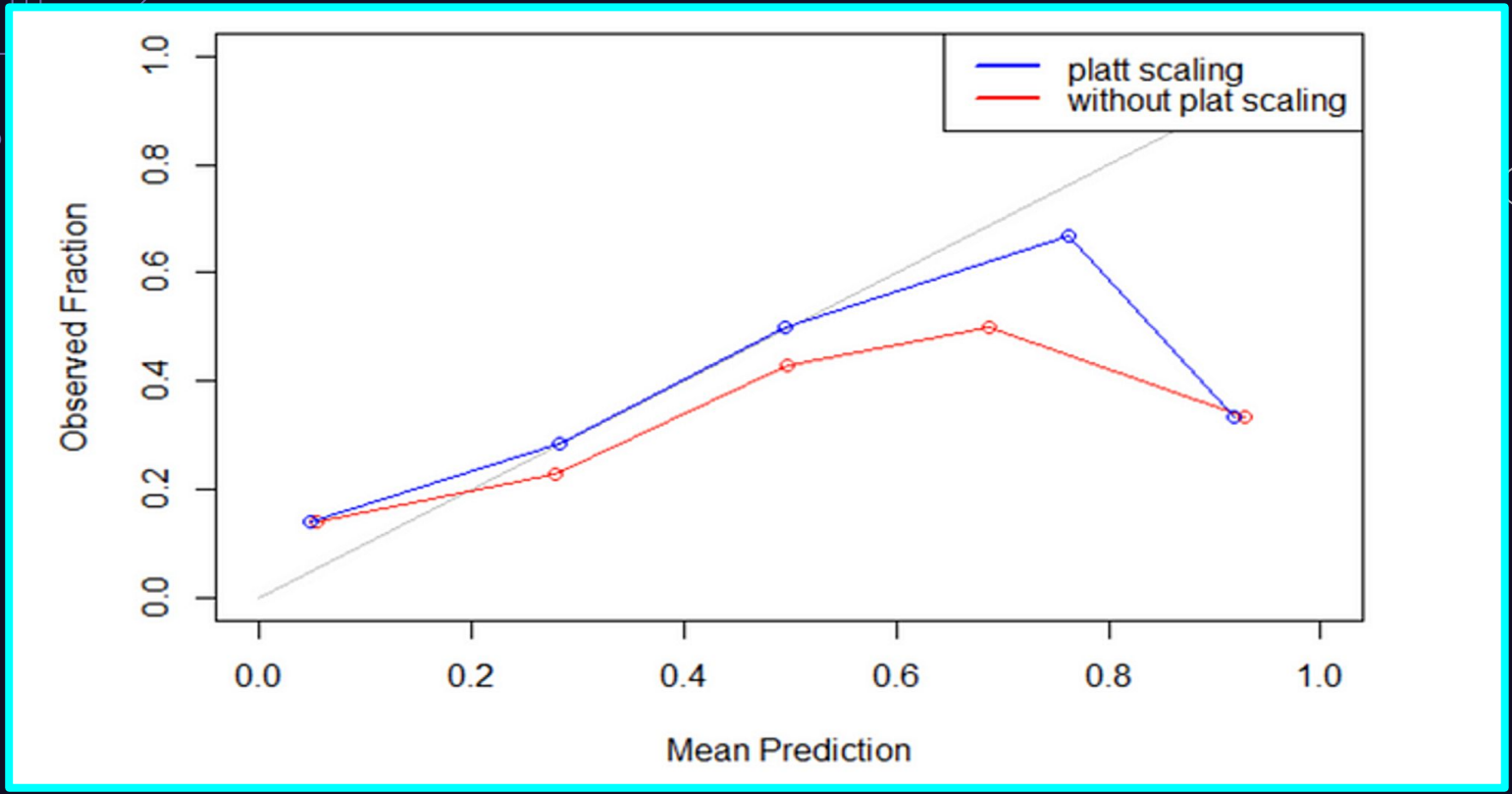
Generate output probabilities for the validation or test data using the binary classifier.

Evaluate the performance of the calibrated classifier on the validation or test data.

Use the logistic regression model to transform the output probabilities on the validation or test data to calibrated probabilities.

Fit a logistic regression model to the output probabilities on the calibration dataset, a subset of the training data that has been held out.

Source: Machine Learning Concept 68: Platt's Scaling, Chandra Prakash Bathula



Img Src: Analytics Vidhya.

Conformal Prediction

- Both a method of uncertainty quantification, and a method of classifying instances (which may be fine-tuned for classes or subgroups).
- Uncertainty is conveyed by classification being in sets of potential classes rather than single predictions.

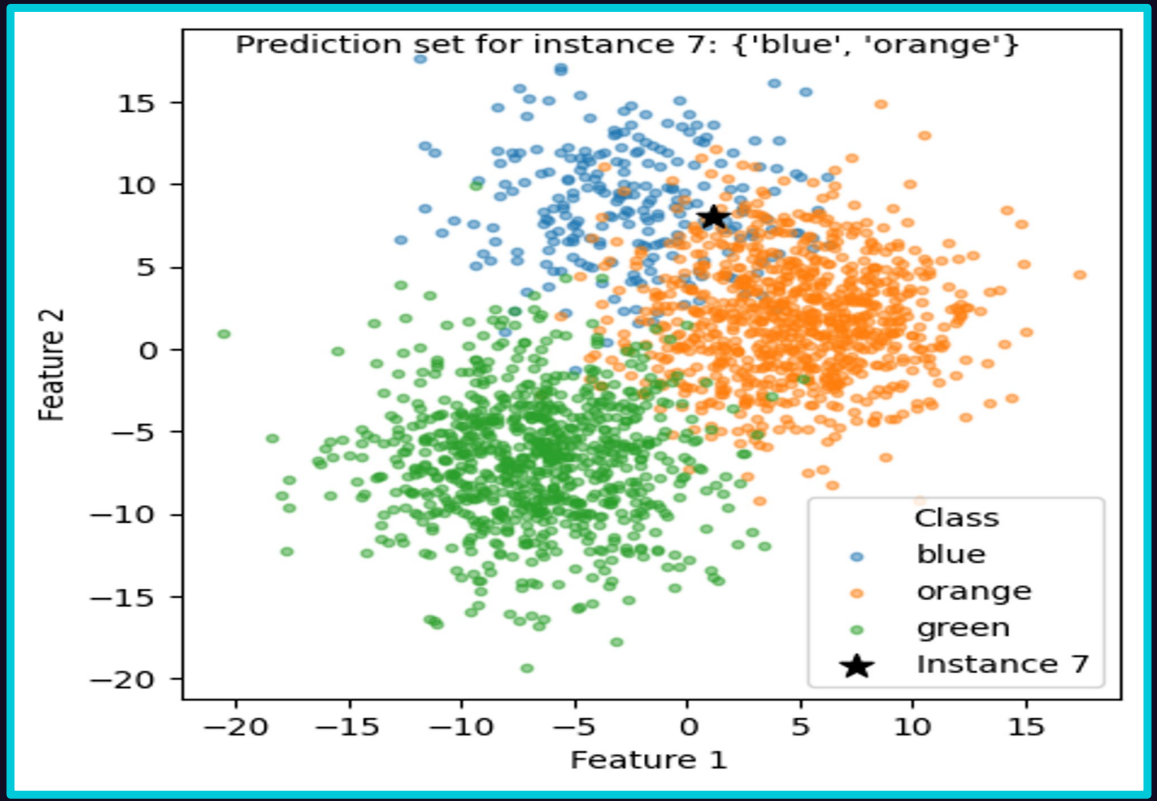
Conformal prediction specifies a *coverage*, which specifies the probability that the true outcome is covered by the prediction region.

Source: https://christophmolnar.com/books/conformal-prediction/

Conformal Prediction

- The interpretation of prediction regions in conformal prediction depends on the task.
- For classification we get prediction sets, while for regression we get prediction intervals.
- Instances on the borders between two classes were labelled with both classes rather than picking the class with highest probability.

Conformal Prediction



Conformal Prediction - Coverage

- Coverage is key to conformal prediction.
- In classification it is the normal region of data that a particular class inhabits.
- Coverage is equivalent to *sensitivity* or *recall*; it is the proportion of observed values that are identified in the classification sets.
- We can tighten or loosen the area of coverage by adjusting α ($coverage = 1 - \alpha$).

Algorithm 1 Naive Prediction Sets

Input: α, sorted scores s, associated permutation of classes I, boolean $rand$

procedure NAIVE($\alpha, s, I, rand$)
  $L \leftarrow 1$
  while $\sum_{i=1}^{L} s_i < 1 - \alpha$ **do** ▷ Stop if $1 - \alpha$ probability exceeded
    $L \leftarrow L + 1$
  if $rand$ **then** ▷ Break ties randomly (explained in Appendix B)
    $U \leftarrow \text{Unif}(0, 1)$
    $V \leftarrow (\sum_{i=1}^{L} s_i - (1 - \alpha))/s_L$
    if $U \leq V$ **then**
      $L \leftarrow L - 1$
  return $\{I_1, ..., I_L\}$

Output: The $1 - \alpha$ prediction set, $\{I_1, ..., I_L\}$

Adaptive Prediction Sets - Motivation

1. The probabilities output by CNNs are known to be incorrect, so the sets from naive do not achieve coverage.
2. Image classification models' tail probabilities are often badly miscalibrated, leading to large sets that do not faithfully articulate the uncertainty of the model.
3. Moreover, smaller sets that achieve the same coverage level can be generated with other methods.

Source: Nixon et.al(2019)

Adaptive Prediction Sets

The coverage problem is solved by picking a new threshold using holdout samples.

- For example, with α =10%, if choosing sets that contain 93% estimated probability achieves 90% coverage on the holdout, the 93% cutoff may be used instead.

The Adaptive Prediction Sets (APS) procedure provides coverage but still produces large sets.

Source: Romano et al. (2020)

Regularized Adaptive Prediction Sets

- To fix this, a regularization technique is introduced that tempers the influence of these noisy estimates, leading to smaller, more stable sets.
- The proposed algorithm is called Regularized Adaptive Prediction Sets (RAPS)
- RAPS is guaranteed to have better performance than choosing a fixed-size set.

Methodology

Our algorithm has three main ingredients.

1. A feature vector x, the base model computes class probabilities $\hat{\pi}_x \in R^k$, and we order the classes from most probable to least probable.
2. Adding a regularization term to promote small predictive sets.
3. Conformally calibrate the penalized prediction sets to guarantee coverage on future test points.

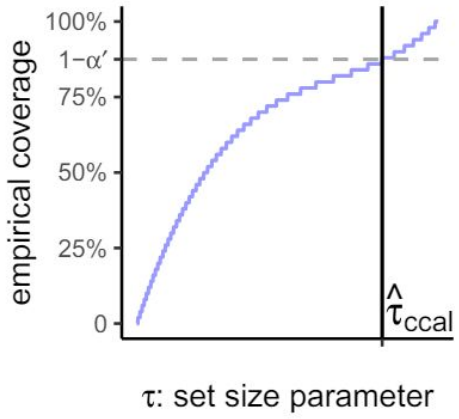
Methodology

- $\rho_x(y)$ = Total probability mass of the set of labels that are more likely than y.
- $o_x(y)$ = Ranking of y among the label based on the probabilities $\hat{\pi}$.
- τ = Tuning parameter that controls the size of the sets. It is the cumulative sum of the sorted, penalized classifier scores in RAPS

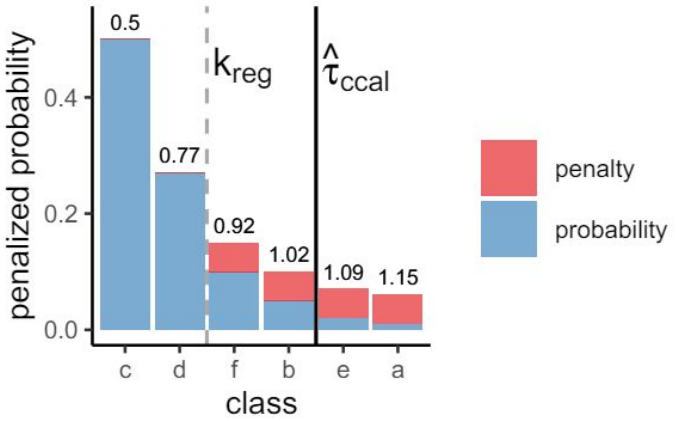
$$\mathcal{C}^*(x, u, \tau) := \left\{ y \ : \ \rho_x(y) + \hat{\pi}_x(y) \cdot u + \underbrace{\lambda \cdot (o_x(y) - k_{reg})^+}_{\text{regularization}} \leq \tau \right\}$$

Methodology

- First, the $\rho_x(y)$ term increases as y ranges from the most probable to least probable label, so our sets will prefer to include the y that are predicted to be the most probable.
- The second term, $\hat{\pi}_x(y) \cdot u$, is a randomized term to handle the fact that the value will jump discreetly with the inclusion of each new y.
- Lastly, the regularization promotes small set sizes.



(a) Conformal calibration



(b) A RAPS prediction set

Figure 3: **Visualizations of conformal calibration and RAPS sets.** In the left panel, the y-axis shows the empirical coverage on the conformal calibration set, and $1-\alpha' = \lceil (n+1)(1-\alpha) \rceil / n$. In the right panel, the printed numbers indicate the cumulative probability plus penalty mass. For the indicated value $\hat{\tau}_{\text{ccal}}$, the RAPS prediction set is {c, d, f, b}.

Algorithm 2 RAPS Conformal Calibration

Input: α; $s \in [0,1]^{n \times K}$, $I \in \{1, ..., K\}^{n \times K}$, and $y \in \{0, 1, ..., K\}^n$ corresponding respectively to the sorted scores, the associated permutation of indexes, and ground-truth labels for each of n examples in the calibration set; k_{reg}; λ; boolean $rand$

procedure RAPSC(α,s,I,y,λ)
  for $i \in \{1, \cdots, n\}$ **do**
    $L_i \leftarrow j$ such that $I_{i,j} = y_i$
    $E_i \leftarrow \Sigma_{j=1}^{L_i} s_{i,j} + \lambda(L_i - k_{reg})^+$
    if $rand$ **then**
      $U \sim \text{Unif}(0, 1)$
      $E_i \leftarrow E_i - U * s_{i,L_i}$
  $\hat{\tau}_{ccal} \leftarrow$ the $\lceil (1-\alpha)(1+n) \rceil$ largest value in $\{E_i\}_{i=1}^n$
  return $\hat{\tau}_{ccal}$

Output: The generalized quantile, $\hat{\tau}_{ccal}$ ▷ The value in Eq. (3)

RAPS

Algorithm 3 RAPS Prediction Sets

Input: α, sorted scores s and the associated permutation of classes I for a test-time example, $\hat{\tau}_{ccal}$ from Algorithm 2, k_{reg}, λ, boolean $rand$

procedure RAPS($\alpha, s, I, \hat{\tau}_{ccal}, k_{reg}, \lambda, rand$)

$L \leftarrow |\{j \in \mathcal{Y} : \Sigma_{i=1}^{j} s_i + \lambda(j - k_{reg})^+ \leq \hat{\tau}_{ccal}\}| + 1$

if $rand$ **then**

$U \leftarrow \text{Unif}(0, 1)$

$L \leftarrow L - \mathbb{I}\{(\Sigma_{i=1}^{L} s_i + \lambda(L - k_{reg})^+ - \hat{\tau}_{ccal})/(s_L + \lambda\mathbb{I}(L > k_{reg})) \leq U\}$

return $\mathcal{C} = \{I_1, ...I_L\}$ ▷ The L most likely classes

Output: The $1 - \alpha$ confidence set, $\mathcal{C}$ ▷ The set in Eq. (4)

RAPS

Why Regularize?

- The sets from APS are larger than necessary, because APS is sensitive to the noisy probability estimates far down the list of classes.
- The noise leads to a permutation problem of unlikely classes, where ordering of the classes with small probability estimates is determined mostly by random chance.

If 5% of the true classes from the calibration set are deep in the tail, APS will choose large 95% predictive sets.

The inclusion of the RAPS regularization causes the algorithm to avoid using the unreliable probabilities in the tail.

	Accuracy		Coverage				Size			
Model	Top-1	Top-5	Top K	Naive	APS	RAPS	Top K	Naive	APS	RAPS
ResNeXt101	0.678	0.874	0.900	0.888	0.899	0.899	7.48	43.0	50.8	**6.18**
ResNet152	0.67	0.876	0.899	0.896	0.900	0.900	7.18	25.8	27.2	**5.69**
ResNet101	0.657	0.859	0.901	0.894	0.900	0.898	9.21	28.7	30.7	**6.93**
ResNet50	0.634	0.847	0.898	0.894	0.899	0.900	10.3	30.3	32.3	**7.80**
ResNet18	0.572	0.802	0.902	0.895	0.900	0.900	17.5	35.3	37.4	**13.3**
DenseNet161	0.653	0.862	0.902	0.895	0.901	0.901	8.6	29.9	32.4	**6.93**
VGG16	0.588	0.817	0.902	0.897	0.900	0.899	15.1	31.9	32.8	**11.2**
Inception	0.573	0.797	0.900	0.893	0.900	0.899	21.8	145.0	155.0	**20.5**
ShuffleNet	0.559	0.781	0.899	0.892	0.900	0.899	26.0	66.2	71.7	**22.5**

Table 2: **Results on Imagenet-V2.** We report coverage and size of the optimal, randomized fixed sets, `naive`, `APS`, and `RAPS` sets for nine different Imagenet classifiers. The median-of-means for each column is reported over 100 different trials at the 10% level. See Section 3.2 for full details.

Conclusion

- RAPS is an upgraded form of APS, which has much better coverage than APS and Naive, while maintaining smaller set sizes, so it is better in quantifying the uncertainty.

- The algorithm is useful for multilabel classification and could be used to automatically screen for a large number of diseases and refer the patient to relevant specialists.

Conclusion

- RAPS is an upgraded form of APS, which has much better coverage than APS and Naive, while maintaining smaller set sizes, so it is better in quantifying the uncertainty.
- The algorithm is useful for multilabel classification and could be used to automatically screen for a large number of diseases and refer the patient to relevant specialists.

Thank You!

Code Demo