Adaptive Prediction Sets with Class Conditional Coverage

Alex Derhacobian, John Guibas, Linden Li, Bharath Namboothiry
SUNets: `alexder`, `jguibas`, `lindenli`, `brn`
Category: Theory

Abstract

While modern machine learning models can perform classification tasks with high accuracy, deploying them in high risk settings requires methods to measure their uncertainty. Prior methods of uncertainty quantification output a prediction set that contains the true label with a user-specified probability, but these guarantees only hold on average over the entire dataset rather than for each class. Here, we present CCAPS[1], an algorithm which can convert a black-box classifier to output a prediction set of labels with a finite-sample coverage guarantee, conditioned on the input being *from any class*. CCAPS works on any dataset, can wrap around any model, and is flexible enough to allow users to choose a subset of classes on which they want the coverage guarantees to hold. Through experiments with classical machine learning classifiers and modern deep neural networks on MNIST, CIFAR-10 and ImageNet, we demonstrate the empirical effectiveness of our theoretical guarantees.

1 Introduction

Machine learning classifiers are increasingly being used for a variety of high-stakes applications [1, 2, 3]. Suppose a doctor is working at a hospital with a new influx of patients experiencing respiratory symptoms. To handle the large number of patients, they are interested in accelerating diagnosis by utilizing a deep neural network which can predict whether a patient has COVID, the FLU, or the COMMON COLD from a lung scan. They would like to rely on the predictions of the model, but have no way of measuring the uncertainty of this diagnosis.

Existing methods such as [4, 5] predict a set of labels that contains the right diagnosis with a user-specified probability. However in this setting, these techniques might not be as helpful. Consider the case where a medical image classifier is trained on a dataset where 95% of lung scans are of patients that are COVID-negative. Since most respiratory illnesses are likely mild cases of the FLU or the COMMON COLD, the model could output the resulting prediction set {FLU, COMMON COLD} and satisfy the coverage guarantee.

What would help a doctor the most is not coverage across the whole dataset but *class-conditional coverage*, or the guarantee that the true label is contained in the prediction set conditioned on *any class*, such as COVID. Formally speaking, suppose we have a calibration set of n data samples $\{(X_i, Y_i)\}_{i=1}^n$ with features X_i and discrete classes Y_i. Our goal is to construct a set-valued function $\hat{C}_{n,\alpha} : \mathcal{X} \to 2^{\mathcal{Y}}$ for an unseen data sample (X_{n+1}, Y_{n+1}) which achieves *class-conditional coverage*. In other words,

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1}) \mid Y_{n+1} = y\right] \geq 1 - \alpha. \tag{1}$$

In this paper, we introduce CCAPS (Class-Conditional Adaptive Prediction Sets), an algorithm which can convert a black-box classifier to output a predictive set of labels formally guaranteed to satisfy

[1] Our code is available on GitHub: `https://github.com/jtguibas/ccaps`.

(1) given a calibration set. CCAPS works regardless of how big the calibration set is (*finite-sample guarantee*), works on any dataset (*distribution-free*), can wrap around any model (*model-agnostic*), and is flexible enough to allow users to choose a subset of classes on which they want the coverage guarantees to hold.

In summary our contributions are as follows: 1) we introduce CCAPS, a conformal prediction algorithm to guarantee class-conditional coverage; 2) we formally justify the theoretical guarantees of CCAPS; and 3) we demonstrate the empirical effectiveness of CCAPS through a comprehensive set of experiments.

2 Related work

CCAPS is largely inspired by the conformal prediction literature. Conformal prediction was introduced in [6] as a non-parametric approach to constructing prediction intervals with probabilistic guarantees for a variety of black-box learning algorithms [7]. It has since been adapted to a variety of tasks in machine learning: [8] applied the result to low-dimensional non-parametric regression, [9] then utilized conformal methods in high-dimensional regression, and [10] recently applied it to image segmentation with expected loss controlled to a user-specified level.

There has been significant work done into improving set-valued classifiers built on conformal methods. While [11] argues that achieving conditional coverage on a given example is impossible, [4, 12] attempt to achieve approximately conditional coverage. [13] proposed the use of density estimators per class to construct prediction sets that have desirable coverage guarantees and are adaptive to each class. [14, 15] applied these techniques to modern neural networks and deep learning by utilizing the outputs of the softmax function for calibration. [16, 17] generalized the conformal procedure to the multi-class context, while [16] also introduced the desiderata of reducing the expected prediction set cardinality of a set-valued classifier, and presented an algorithm to control user-specified error levels conditioned on each class. Follow up work in [5] introduced a regularization term to penalize large set sizes. Our approach will utilize split-conformal methods as in [9, 18] that splits a given dataset for conformal prediction on most black-box learning algorithms.

CCAPS builds mostly off of the work of Adaptive Prediction Sets by Romano et. al which guarantee marginal coverage and perform favorably in terms of approximate conditional coverage compared to alternative methods [4]. To our knowledge, no one has yet adapted the techniques introduced in [4] to the class conditional setting.

3 Method

3.1 Adaptive Prediction Sets

We will introduce a general technique for developing prediction sets with weaker coverage guarantees than class-conditional coverage. In particular, we follow the presentation in [4]. Suppose we have n data samples $\{(X_i, Y_i)\}_{i=1}^n$ with features $X_i \in \mathcal{X}$ and discrete classes $Y_i \in \mathcal{Y} = \{1, ..., K\}$. The samples are drawn exchangeably from a joint distribution P_{XY}. The goal of this section is to construct a set-valued function $\hat{\mathcal{C}}_{n,\alpha} : \mathcal{X} \to 2^{\mathcal{Y}}$ for an unseen data sample (X_{n+1}, Y_{n+1}) which achieves *marginal coverage*. In other words,

$$\mathbb{P}\left[Y_{n+1} \in \hat{\mathcal{C}}_{n,\alpha}(X_{n+1})\right] \geq 1 - \alpha. \tag{2}$$

Given a black-box learning algorithm, we can train a model π and compute $\hat{\pi}_y(x)$ which estimates $\pi_y(x) = \mathbb{P}[Y_i = y | X_i = x]$ for each $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Let k^* denote the class with the k-th largest probability mass. Then, for any threshold $\tau \in [0, 1]$, [4] defines the conditional quantile function as follows:

$$Q(x; \pi, \tau) = \min\{k \in \{1, ..., K\} : \pi_{1^*}(x) + ... + \pi_{k^*}(x) \geq \tau\}. \tag{3}$$

If we have access to $\pi_y(x)$, then the smallest prediction set for a confidence level α can be produced by the set $\{1^*, ..., Q(x; \pi_y(x), 1 - \alpha)^*\}$. [4] defines the following prediction set generating function, which we will denote as S:

$$S(x; \pi, \tau) = \begin{cases} \{1^*, ..., (Q(x; \pi, \tau) - 1)^*\} & \text{with probability } T(x; \pi, \tau) \\ \{1^*, ..., Q(x; \pi, \tau)^*\} & \text{otherwise} \end{cases} \tag{4}$$

where

$$T(x;\pi,\tau) = \frac{\sum_{k=1}^{Q(x;\pi,\tau)(x)} \pi_{k^*}(x) - \tau}{\pi_{Q(x;\pi,\tau)^*}(x)} \tag{5}$$

$T(x;\pi,\tau)$ can be thought of as the tie-breaking probability to produce strictly smaller sets while maintaining coverage. We will now define the inverse quantile conformity score function E,

$$E(x,y;\pi) = \min\{\tau \in [0,1] \,:\, y \in S(x;\pi,\tau)\} \tag{6}$$

We will now discuss how we can use these functions to develop prediction sets which achieve the guarantees described in (2). Given a black-box model $\hat{\pi}$, for each data sample in $\{(X_i,Y_i)\}_{i=1}^{n}$, we will compute $E_i = E(X_i,Y_i;\hat{\pi})$, which is the threshold value for the true label to be contained in the set generated by S. Then, we will compute the $1-\alpha$ quantile of $\{E_i\}_{i=1}^{n}$ denoted $\hat{\tau}$. We can then create prediction sets for a new data point (X_{n+1},Y_{n+1}) with a model via:

$$\hat{C}_{n,\alpha}^{SC}(X_{n+1}) = S(X_{n+1};\hat{\pi},\hat{\tau}) \tag{7}$$

For more details on this procedure, please refer to [4].

3.2 Our Method

The procedure described above guarantees marginal coverage over the entire dataset. However, recall that we are interested in creating prediction sets that guarantee *class-conditional* coverage. Formally speaking, for each each class $y \in \mathcal{Y}$ we would like to guarantee:

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_{n,\alpha}^{sc}(X_{n+1}) \mid Y_{n+1} = y\right] \geq 1-\alpha. \tag{8}$$

To create prediction sets which satisfy this guarantee, we adapt the split conformal calibration procedure from Algorithm 2 in [9]. We repeat the calibration step K times, once for each class. In particular, we partition the calibration set by class and compute the threshold $\hat{\tau}_k$ as done in Section 3.1. This threshold guarantees marginal coverage on its respective calibration set and because the calibration set contains examples exclusively from one class, this threshold now satisfies the property of being *class-conditional*.

Now we have a set of thresholds $\{\hat{\tau}_1,...,\hat{\tau}_K\}$ which guarantee marginal coverage conditioned on their respective class. Since we do not have access to the true label at test time, we must use the worst-case threshold $\hat{\tau}_{\max} = \max_k\{\hat{\tau}_1,...,\hat{\tau}_K\}$ to guarantee class conditional coverage. Since $\hat{\tau}_{\max} \geq \hat{\tau}_k$ for $k = \{1,\ldots,K\}$, the threshold $\hat{\tau}_{\max}$ guarantees class conditional coverage. For technical details, please refer to the pseudocode for CCAPS below.

Algorithm 1 CCAPS (Class Conditional Adaptive Prediction Sets)

Input: dataset $\{(X_i,Y_i)\}_{i=1}^{n}$, new data point X_{n+1}, confidence level α, number of classes K, exclusion threshold $\gamma \in [0, K-1]$
Split the training data into 2 subsets, $\mathcal{I}_{\text{train}}, \mathcal{I}_{\text{cal}}$
Split $\mathcal{I}_{\text{cal}}$ into K disjoint subsets $\mathcal{I}_{\text{cal}}^{(1)},\ldots,\mathcal{I}_{\text{cal}}^{(K)}$, where $\mathcal{I}_{\text{cal}}^{(k)} = \{(X_i,Y_i) \in \mathcal{I}_{\text{cal}} \mid Y_i = k\}$
$\hat{\pi} \leftarrow$ black-box learning model trained on $\mathcal{I}_{\text{train}}$
for $k \in 1,\ldots,K$ **do**
    $E_i \leftarrow E(X_i,Y_i;\hat{\pi}),\ \forall\ (X_i,Y_i) \in \mathcal{I}^{(k)}$ where E is defined in (6)
    $\hat{\tau}_k \leftarrow$ the $\lceil(1-\alpha)(1+|\mathcal{I}_{\text{cal}}^{(k)}|)\rceil$-th largest value in $\{E_i\}_{i\in\mathcal{I}_{\text{cal}}^{(k)}}$
end for
$\{\hat{\tau}_k\} \leftarrow \{\hat{\tau}_k\}$ excluding top γ largest $\hat{\tau}_k$
$\hat{\tau}_{\max} \leftarrow \max_k\{\hat{\tau}_k\}$
$\hat{\mathcal{C}}_{n,\alpha}^{SC}(X_{n+1}) \leftarrow S(X_{n+1};\hat{\pi},\hat{\tau}_{\max})$
Output: The $1-\alpha$ class-conditional prediction set $\hat{\mathcal{C}}_{n,\alpha}^{SC}(X_{n+1})$ for unobserved label Y_{n+1}

Class-conditional coverage is a much stronger guarantee than marginal coverage. As a result, this can lead to much larger set sizes. Hence, one might wish to condition on a specific subset of classes rather than all classes. To address this, we propose a new hyperparameter γ, the *exclusion threshold*. We can exclude the top γ classes with the largest $\hat{\tau}$ values. This will decrease our average set size by excluding the γ "hardest" classes.

3.3 Theoretical Guarantees

Theorem 1 (CCAPS Coverage Guarantee). *Let the data be denoted $\{(X_i, Y_i)\}_{i=1}^{n+1}$. If the data is exchangeable and the black box learning algorithm utilized to train the model in Algorithm 1 is invariant to permutations of its input samples, the output of Algorithm 1 satisfies the following for each class $k \in \{1, \dots, K\}$ (up to the exclusion threshold γ):*

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_{n,\alpha}^{sc}(X_{n+1}) \mid Y_{n+1} = k\right] \geq 1 - \alpha$$

A full proof is provided in the Appendix.

4 Experiments

Hyperparameters			Coverage		Min. CC		Set Size	
Dataset	# Calib	Method	APS	CCAPS	APS	CCAPS	APS	CCAPS
MNIST (C=10)	10K	RFC	0.90	0.96	0.79	**0.90**	1.49	2.12
MNIST (C=10)	10K	MLP	0.90	0.92	0.87	**0.89**	0.89	0.98
CIFAR10 (C=10)	10K	RFC	0.90	0.93	0.85	**0.90**	5.43	6.27
CIFAR10 (C=10)	10K	MLP	0.90	0.94	0.85	**0.90**	4.94	6.68
ImageNet (C=1K)	25K	RN-152	0.90	0.99	0.52	**0.91**	6.34	296.
CIFAR-10 (C=10)	5K	MLP	0.91	0.93	0.86	**0.90**	5.28	5.5
CIFAR-10 (C=10)	2.5K	MLP	0.90	0.94	0.85	**0.90**	4.26	5.00
CIFAR-10 (C=10)	1K	MLP	0.90	0.92	0.80	**0.89**	4.33	4.99
CIFAR10-LT (C=10)	10K	RFC	0.90	0.99	0.29	**0.90**	5.31	9.65

Table 1: Results are displayed for a random forest classifier and a one-layer multi-layer perceptron on the two datasets CIFAR-10 and MNIST trained on 10000 examples as well as ImageNet where $\alpha = 0.1$. Results are averaged over 50 trials.

We empirically demonstrate the performance of CCAPS through a variety of datasets, learning algorithms, and parameters in **Table 1**. In terms of datasets, we test on MNIST [19], CIFAR-10 [20], ImageNet [21], CIFAR-10 [20] with one of the classes reduced by five times. In terms of learning algorithms, we test the usage of Random Forests, Neural Networks [22], and ResNet152 [23, 24]. In the following sections, we go into more detail about our experiments.

4.1 Experiment 1: Demonstrating Class Conditional Coverage

Refer to **Table 1** for results. We report the marginal coverage, the minimum conditional coverage over all classes (denoted Min. CC), and the average set size. Compared to APS, we are able to achieve class conditional coverage across all classes (hence the minimum coverage across all the classes is 0.9). In challenging settings like ImageNet across multiple classes, CCAPS is still able to guarantee class conditional coverage while APS does not. Set sizes for CCAPS show only small increases when compared to APS on most datasets, which indicates that we have to make only a small sacrifice to invoke our coverage guarantee. The set size difference is dramatic only for ImageNet, but we are still able to rule > 70% of possible classes while still preserving class conditional coverage.

In practice, a practitioner may only the need the coverage guarantee to hold for any subset of the classes: we demonstrate in the next section that our algorithm is flexible enough to accomodate for this and exhibit its nice properties.

4.2 Experiment 2: Finite-sample Guarantees

We demonstrate finite-sample guarantees by showing that we achieve class conditional coverage for calibration sets of varying sizes. In our experiments, we ran APS and CCAPS on CIFAR-10 with 5000, 2500, and 1000 calibration points. In all cases, CCAPS achieves class conditional coverage for all classes. See **Table 1** for results.

4.3 Experiment 3: Long-tailed Datasets

Class conditional coverage is largely motivated by the class imbalanced setting. As a result, we purposely modified CIFAR-10 to have one severely rare class. In particular, we reduce the first class in CIFAR-10 by five times. Results are shown in **Table 1**.

5 Conclusion

In conclusion, we have introduced CCAPS, an algorithm which can convert a black-box classifier to output a predictive set of labels formally guaranteed to satisfy *class-conditional coverage* given a calibration set. CCAPS works regardless of how big the calibration set is (*finite-sample guarantee*), works on any dataset (*distribution-free*), can wrap around any model (*model-agnostic*), and is flexible enough to allow users to choose a subset of classes on which they want the coverage guarantees to hold. We empirically demnostrate the effectiveness of CCAPS through a comprehensive set of experiments.

In future work, we would like to explore the idea of *approximate* class-conditional coverage. In other words, can we strike a middle ground between marginal coverage and class-conditional coverage? We recognize the fact that achieving class-conditional coverage results in larger prediction set sizes, so we would like to see if we can loosen the guarantee and get significantly smaller prediction sets.

6 Appendix

Proof of Theorem 1. By our construction of the prediction set in (6), we know that:

$$Y_{n+1} \in \hat{\mathcal{C}}^{SC}_{n,\alpha}(X_{n+1})$$

if and only if

$$\min\{\tau \in [0,1] : Y_{n+1} \in S(X_{n+1}; \hat{\pi}, \tau)\} \leq \hat{\tau}_{\max}$$

or, equivalently, if and only if

$$E_{n+1} \leq \hat{\tau}_{\max}$$

We want to show of course that the probability of this event, conditioned on Y_{n+1} being any class k, is at least $1-\alpha$. Let the label Y_{n+1} be any label k in out set of classes $\{1,\dots,K\}$. Recall that $\hat{\tau}_k$ is indeed the $\lceil (1-\alpha)(1+|\mathcal{I}^{(k)}_{\text{cal}}|)\rceil$-th largest value in $\{E_i\}_{i\in\mathcal{I}^{(k)}_{\text{cal}}}$. Since all the conformity scores E_{n+1} and $\{E_i\}_{i\in\mathcal{I}^{(k)}_{\text{cal}}}$ are exchangeable, we know that the probability E_{n+1} being at most α is at least $1-\alpha$. So, we can write:

$$\mathbb{P}\left[E_{n+1} \leq \hat{\tau}_k \mid Y_{n+1} = k\right] \geq 1-\alpha$$

Moreover, from Algorithm 1, recall that we set $\hat{\tau}_k \leq \hat{\tau}_{\max}$. If $E_{n+1} \leq \hat{\tau}_k$, then $E_{n+1} \leq \hat{\tau}_{\max}$, which implies k is in the prediction set. Hence,

$$E_{n+1} \leq \hat{\tau}_k \implies k \in \hat{\mathcal{C}}^{SC}_{n,\alpha}(X_{n+1})$$

With this, we can conclude that

$$\mathbb{P}\left[Y_{n+1} \in \hat{\mathcal{C}}^{SC}_{n,\alpha}(X_{n+1}) \middle| Y_{n+1} = k\right] \geq 1-\alpha$$

which is the desired class conditional coverage guarantee. It trivially follows that this proof can be generalized for any exclusion threshold γ. □

7 Acknowledgements

We would like to especially thank Anastasios Angelopoulos and Stephen Bates for early advice on the project. Furthermore, we would like to thank Yao Liu for his invaluable feedback on the writeup and the idea as our project mentor.

8 Contributions

Alex created and ran the experiments, John and Linden did the writeup, and Bharath developed the proofs for our formal results. All authors contributed equally to the development of the algorithm.

References

[1] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, May 2019.

[2] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604, Sep 2018.

[3] Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1):4–21, 2017.

[4] Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. 2020.

[5] Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction. 2021.

[6] Vladimir Vovk, A. Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

[7] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. 2007.

[8] Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

[9] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression, 2017.

[10] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets, 2021.

[11] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference, 2020.

[12] Maxime Cauchois, Suyash Gupta, and John Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction, 2020.

[13] Yotam Hechtlinger, Barnabás Póczos, and Larry Wasserman. Cautious deep learning, 2019.

[14] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017.

[15] Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. Measuring calibration in deep learning, 2020.

[16] Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, Jun 2018.

[17] Leying Guan and Rob Tibshirani. Prediction and outlier detection in classification problems, 2019.

[18] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Machine Learning: ECML 2002*, pages 345–356, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[19] Y. LECUN. The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*.

[20] Antonio Torralba, Rob Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.