Conformal Prediction for Ensembles

Improving Efficiency via Score-Based Aggregation

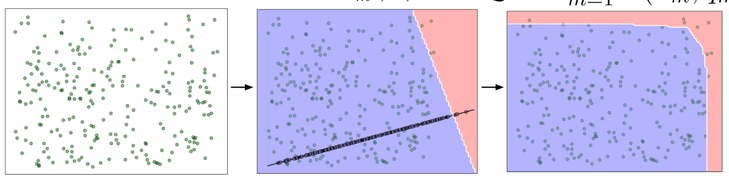
Eduardo Ochoa Rivera*, Yash Patel*, Ambuj Tewari

Department of Statistics, University of Michigan

Overview

Methods for conformal aggregation have been proposed for ensembled prediction, where the prediction regions of individual models are merged to retain coverage guarantees while minimizing conservatism. Merging the prediction *regions* directly, however, can miss out on opportunities to further reduce conservatism by exploiting structures present in the conformal *scores*. We, therefore, propose a novel framework that extends the standard scalar formulation of a score function to a multivariate score that produces more efficient prediction regions.

$\mathcal{S} = \{(s_1(x_i), s_2(x_i))\}_{i=1}^{N_C}$ $\Pi_{u_m}(\mathcal{S})$ $\widehat{\mathcal{Q}} = \cap_{m=1}^{M} H(u_m, \widehat{q}_m)$



Setup

Suppose we have K predictors $f_1(x), ..., f_K(x)$ and corresponding scores $s_1(x, y), ..., s_K(x, y)$ are defined. Naively, one may define a map $g : \mathbb{R}^K \to \mathbb{R}$, e.g., $g(s) = \sum_{k=1}^{K} s_k$ and conformalize this aggregated score. However, using a *fixed* g fails to adapt to any disparities in uncertainties present across predictors.

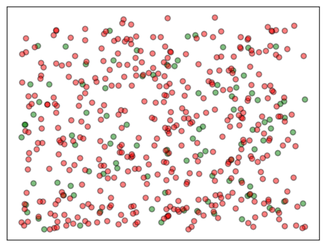
Importantly, we assume the score functions are non-negative, i.e., $s_k : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, which is typically the case as the score serves as a generalization of the residual.

Intuition

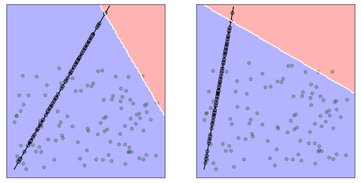
Our method seeks to directly generalize the approach of split conformal, by "ordering" the collection of multivariate calibration scores and taking the $1 - \alpha$ score under such an ordering.

This multivariate "ordering" is established as a pre-ordering $\lesssim$ over $\mathbb{R}^K$. Roughly speaking, an "acceptance region," so called as it serves as the criterion used to ultimately decide which y are accepted into the prediction region, is defined as $\widehat{\mathcal{Q}} := \{s \mid s \lesssim \widehat{q}\}$, where $\widehat{q}$ is the $1 - \alpha$ empirical quantile of $\mathcal{S}_C$ under $\lesssim$.

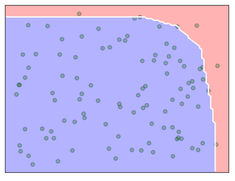
Method



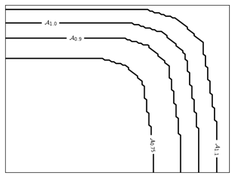
(a) Calibration scores are split into $\mathcal{S}_C^{(1)}$ (green) and $\mathcal{S}_C^{(2)}$ (red).



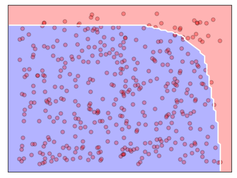
(b) Pre-ordering points are projected onto multiple directions and directional quantiles define half-spaces.



(c) Intersection of half-spaces yields the multivariate quantile envelope.



(d) The envelope induces nested sets $\mathcal{A}_t$ and a partial order on $\mathbb{R}^K$.



(e) Using $\mathcal{S}_C^{(2)}$, the envelope is adjusted to obtain the final scaling factor.

Algorithm 1 CSA

procedure CSA
  Inputs: Scores $s_1, \ldots, s_K$, Calibration $\mathcal{D}_\mathcal{C}$, Mis-coverage α
  $[\beta_{\text{lo}}, \beta_{\text{hi}}] \leftarrow [\alpha/M, \alpha]$, $\widehat{\mathcal{Q}} \leftarrow \emptyset$
  $\sigma \sim \text{Unif}(\text{Permutations of } \{1, \ldots, N_C\})$
  Split $\left\{(s_k(x_{\sigma(i)}, y_{\sigma(i)}))_{k=1}^{K}\right\}_{i=1}^{N_C}$ into $\mathcal{S}_\mathcal{C}^{(1)}$ and $\mathcal{S}_\mathcal{C}^{(2)}$
  $\{u_m\}_{m=1}^{M}$ with $u_m \leftarrow \text{UnifHypersphere}(K)$
  while $\left|\mathcal{S}_\mathcal{C}^{(1)} \cap \widehat{\mathcal{Q}}\right| / N_{\mathcal{C}_1} \notin 1 - \alpha \pm \epsilon$ **do**
    $\beta \leftarrow (\beta_{\text{lo}} + \beta_{\text{hi}})/2$
    $\left\{\widetilde{q}_m \leftarrow (1 - \beta) \text{ quantile of } \{u_m^\top s_i\}_{s_i \in \mathcal{S}_\mathcal{C}^{(1)}}\right\}_{m=1}^{M}$
    $\widehat{\mathcal{Q}} \leftarrow \bigcap_{m=1}^{M} H(u_m, \widetilde{q}_m)$
    If $\left|\mathcal{S}_\mathcal{C}^{(1)} \cap \widehat{\mathcal{Q}}\right| / N_{\mathcal{C}_1} > 1 - \alpha$ **then** $\beta_{\text{lo}} \leftarrow \beta$ **else** $\beta_{\text{hi}} \leftarrow \beta$
  end while
  $\widehat{t} \leftarrow (1 - \alpha) \text{ quantile of } \left\{\max_{m \in [M]} \left(u_m^\top s_i / \widetilde{q}_m\right)\right\}_{s_i \in \mathcal{S}_\mathcal{C}^{(2)}}$
  Return $\{(u_m, \widehat{t}\,\widetilde{q}_m)\}_{m=1}^{M}$
end procedure

Theorem

Suppose $(X_1, Y_1), \ldots, (X_{N_C}, Y_{N_C}), (X', Y')$ *are exchangeable, where* $\mathcal{D}_C := \{(X_i, Y_i)\}_{i=1}^{N_C}$. *Assume further that* K *non-negative maps* $s_k : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ *have been defined and a composite* $s(X, Y) := (s_1(X, Y), ..., s_K(X, Y))$ *is defined. Let* σ, $(\mathcal{S}_C^{(1)}, \mathcal{S}_C^{(2)})$, *and* $U = \{u_m\}_{m=1}^{M}$ *be as defined by lines 3-4 of Algorithm 1 for some* $\alpha \in (0, 1)$. *Denote by* $\{\tilde{q}_m\}_{m=1}^{M}$ *the parameters defined by lines 4-9 of Algorithm 1 and by* T *the scoring function* $T(s; \mathcal{S}_C^{(1)}, U) = \max_{m=1,...,M}(u_m^\top s / \tilde{q}_m)$ *for any score vector* $s \in \mathbb{R}_+^K$. *Then, denoting by* $\mathcal{C}(X') = \{y \in \mathcal{Y} \mid T(s(X', y); \mathcal{S}_C^{(1)}) \leq \hat{t}\}$, $\mathcal{P}(Y' \in \mathcal{C}(X')) \geq 1 - \alpha$, *where the probability is over the joint draws of* $\mathcal{D}_C$, (X', Y'), *and* σ.

Experiments

We now apply CSA across both classification and regression tasks, comparing to naive conformalization, conformal aggregation strategies ($\mathcal{C}^M$, $\mathcal{C}^R$,$\mathcal{C}^U$), and aggregation via a single weight vector (VFCP):

Table 1. Classification results are shown across tasks for various values of α with coverages in the top (grey) and average prediction set sizes (white) in the bottom.

Dataset/α	ResNet	VGG	DenseNet	VFCP	$\mathcal{C}^M$	$\mathcal{C}^R$	$\mathcal{C}^U$	Ensemble	CSA
ImageNet	0.901 (0.005)	0.902 (0.003)	0.902 (0.003)	0.899 (0.004)	0.938 (0.003)	0.909 (0.004)	0.9 (0.004)	0.899 (0.004)	0.9 (0.003)
($\alpha = 0.10$)	137.004 (1.98)	136.116 (2.206)	120.096 (2.427)	46.063 (1.089)	87.337 (1.604)	82.746 (1.692)	131.856 (2.378)	69.123 (1.317)	**34.006 (0.924)**
($\alpha = 0.05$)	0.95 (0.003)	0.949 (0.004)	0.952 (0.002)	0.95 (0.003)	0.975 (0.002)	0.954 (0.004)	0.95 (0.003)	0.949 (0.002)	0.95 (0.003)
	220.022 (2.072)	229.523 (3.076)	208.658 (2.016)	78.108 (2.004)	166.933 (2.157)	143.323 (2.932)	220.491 (2.773)	112.161 (2.115)	**59.574 (3.382)**
($\alpha = 0.01$)	0.99 (0.001)	0.991 (0.001)	0.989 (0.002)	0.99 (0.001)	0.997 (0.001)	0.991 (0.002)	0.99 (0.002)	0.99 (0.002)	0.99 (0.002)
	491.952 (6.353)	726.028 (12.157)	459.399 (6.739)	**194.691 (4.579)**	580.592 (7.715)	532.155 (24.829)	559.188 (7.07)	299.453 (6.526)	**201.32 (46.509)**

Table 2. The results for five regression tasks for $\alpha = 0.05$. Again, the average coverages (grey rows) and prediction set lengths (white rows) are given.

Dataset/α	OLS	LASSO	RF	XGBoost	$\mathcal{C}^M$	$\mathcal{C}^R$	$\mathcal{C}^U$	Ensemble	*Single-Stage*	CSA
361234	0.97 (0.011)	0.966 (0.011)	0.939 (0.002)	0.954 (0.006)	0.956 (0.011)	0.948 (0.01)	0.96 (0.013)	0.95 (0.006)	*0.955 (0.013)*	0.957 (0.01)
($\alpha = 0.05$)	9.673 (0.160)	9.645 (0.154)	10.080 (0.160)	9.157 (0.052)	9.196 (0.123)	8.703 (0.086)	9.524 (0.056)	17.759 (0.275)	*7.646 (0.073)*	**7.688 (0.181)**
361235	0.947 (0.0)	0.945 (0.005)	0.968 (0.016)	0.95 (0.005)	0.955 (0.016)	0.897 (0.005)	0.953 (0.011)	0.932 (0.021)	*0.745 (0.011)*	0.984 (0.005)
($\alpha = 0.05$)	20.961 (0.651)	24.241 (0.246)	**10.096 (0.587)**	11.387 (0.452)	11.782 (0.057)	–	16.088 (0.118)	15.823 (1.272)	*6.162 (0.458)*	11.695 (0.266)
361236	0.975 (0.008)	0.975 (0.008)	0.961 (0.0)	0.948 (0.012)	0.948 (0.012)	0.938 (0.012)	0.965 (0.008)	0.934 (0.004)	*0.94 (0.004)*	0.963 (0.004)
($\alpha = 0.05$)	4.44e4 (1.17e3)	4.45e4 (1.23e3)	5.08e4 (3.86e2)	4.10e4 (1.22e3)	4.32e4 (1.00e3)	4.09e4 (1.09e3)	4.44e4 (8.52e2)	6.05e4 (2.41e3)	*3.10e4 (2.48e3)*	**3.34e4 (1.28e3)**
361237	0.969 (0.023)	0.969 (0.023)	0.981 (0.0)	0.923 (0.0)	0.954 (0.015)	0.9 (0.008)	0.969 (0.023)	0.885 (0.038)	*0.8 (0.015)*	0.977 (0.008)
($\alpha = 0.05$)	44.019 (0.990)	44.069 (1.115)	27.035 (1.014)	–	26.524 (1.244)	–	31.967 (1.118)	–	*14.473 (0.503)*	**23.145 (0.199)**
361241	0.954 (0.001)	0.956 (0.001)	0.944 (0.005)	0.957 (0.002)	0.954 (0.002)	0.923 (0.0)	0.952 (0.0)	0.949 (0.001)	*0.917 (0.006)*	0.951 (0.001)
($\alpha = 0.05$)	19.133 (0.062)	20.245 (0.095)	18.102 (0.055)	18.482 (0.062)	17.958 (0.062)	–	18.932 (0.034)	29.548 (0.191)	*15.199 (0.427)*	**17.328 (0.097)**

Discussion

This work suggests many directions for future work. One is in extracting insights of the relative predictor uncertainties from the data-driven relation $\lesssim$. Another is the integration of CSA with end-to-end decision-making frameworks to automatically construct regions.

yppatel@umich.edu