Distribution-Free, Risk-Controlling Prediction Sets

Ramya Ramalingam

Department of Computer and Information Science
University of Pennsylvania

April 13th, 2022

Table of Contents

Definitions

Goal: Generate prediction sets with high-probability guarantees.

Definition 1 *(Risk-Controlling Prediction Set)*: Let $\mathcal{T}$ be a function which takes values in $\mathcal{X} \to \mathcal{Y}'$. $\mathcal{T}$ is (α, δ)-risk-controlling if:

$$\mathbb{P}(R(\mathcal{T}) \leq \alpha) \geq 1 - \delta$$

Set-Up

- Given a dataset $((X_1, Y_1), (X_2, Y_2), \cdots (X_m, Y_m))$ - split into training (size $m - n$) and calibration data (size n).
- User uses training data to train a predictive model $\hat{f}$.
- Define loss function on prediction-sets $L(y, \mathcal{S}) : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}_{\geq 0}$ which satisfies a nesting property:

$$\mathcal{S} \subset \mathcal{S}' \implies L(y, \mathcal{S}) \geq L(y, \mathcal{S}')$$

- Use calibration data to select best prediction-set model $\mathcal{T}$ from a parametric set of such models.

Basic Procedure

1. Assume a collection of functions $\{\mathcal{T}_\lambda\}_{\lambda\in\Lambda}$ with the following nesting property:
$$\lambda_1 < \lambda_2 \implies \mathcal{T}_{\lambda_1}(x) \subset \mathcal{T}_{\lambda_2}(x)$$
2. Assume a point-wise *upper confidence bound* (UCB) $\hat{R}^+$ for risk so that for all λ:
$$\mathbb{P}(R(\lambda) \leq \hat{R}^+(\lambda)) \geq 1-\delta \tag{1}$$
3. Choose $\hat{\lambda}$ as:
$$\hat{\lambda} = \inf\{\lambda \in \Lambda \mid \hat{R}^+(\lambda') < \alpha \quad \forall \lambda' \geq \lambda\} \tag{2}$$

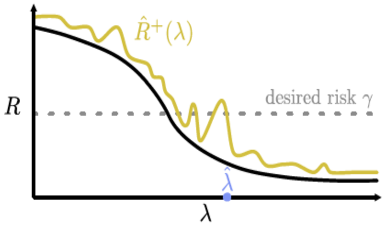


Figure 3: Visualization of UCB calibration.

Theorem 1 *(validity of UCB calibration)*: Let R be a continuous non-increasing function such that $R(\lambda) \leq \alpha$ for some α. For $\hat{\lambda}$ as defined above, $\mathcal{T}_{\hat{\lambda}}$ is a (α, δ)-risk controlled prediction set.

Proof. Define λ^* as:

$$\lambda^* = \inf\{\lambda \in \Lambda \mid R(\lambda) \leq \alpha\}$$

If $R(\hat{\lambda}) > \alpha$, then $\hat{\lambda} < \lambda^*$. Then by definition, $\hat{R}^+(\lambda^*) < \alpha$. Since $R(\lambda^*) = \alpha$, this happens will probability less than δ (due to the UCB guarantee). So,

$$\mathbb{P}(R(\hat{\lambda}) \leq \alpha) \geq 1 - \delta$$

Penn Engineering

Concentration Inequalities for UCB

Simple Hoeffding Bound (Bounded Loss)

Define empirical risk $\hat{R}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} L(Y_i, \mathcal{T}_\lambda(X_i))$.

Assuming $L(y, \mathcal{S}) \in [0, 1]$ for all $y \in \mathcal{Y}, \mathcal{S} \in \mathcal{Y}'$, use Hoeffding's inequality:

$$\mathbb{P}(\hat{R}(\lambda) - R(\lambda) \leq -x) \leq \exp\{-2nx^2\}$$

which gives us the following UCB:

$$\hat{R}^+_{\text{sHoef}}(\lambda) = \hat{R}(\lambda) + \sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)}$$

Concentration Inequalities for UCB

Proposition 2: Let $g(t; R)$ be a non-decreasing function in $t \in \mathbb{R}$ for every R, and:

$$\mathbb{P}(\hat{R}(\lambda) \leq t) \leq g(t; R(\lambda))$$

Then $\hat{R}^{+}(\lambda) = \sup\{R \mid g(\hat{R}(\lambda); R) \geq \delta\}$ is a valid UCB.

Proof. Note that:

$$\mathbb{P}(R(\lambda) > \hat{R}^{+}(\lambda)) \leq \mathbb{P}(g(\hat{R}(\lambda); R(\lambda)) < \delta) \leq \mathbb{P}(G(\hat{R}(\lambda)) < \delta)$$

where G is the CDF for $\hat{R}(\lambda)$. Define $G^{-1}(\delta) = \sup\{x : G(x) \leq \delta\}$. Then,

$$\mathbb{P}(G(\hat{R}(\lambda)) < \delta) \leq \mathbb{P}(\hat{R}(\lambda) < G^{-1}(\delta)) \leq \delta$$

and so

$$\mathbb{P}(R(\lambda) > \hat{R}^{+}(\lambda)) \leq \delta$$

as desired.

Concentration Inequalities for UCB

Using concentration inequalities:

Tight Hoeffding Bound [Hoeffding, 1963]: For any $t \leq R(\lambda)$,

$$\mathbb{P}(\hat{R}(\lambda) \leq t) \leq \exp\{-nh_1(t; R(\lambda))\}$$

where $h_1(t; R) = t \log(t/R) + (1-t)\log((1-t)/(1-R))$.

Bentkus Inequality [Bentkus, 2004]: If loss is bounded above by one, then:

$$\mathbb{P}(\hat{R}(\lambda) \leq t) \leq e\mathbb{P}(\mathrm{Binom}(n, R(\lambda)) \leq \lceil nt \rceil)$$

where $\mathrm{Binom}(n, p)$ is a binomial random variable with sample size n and success probability p.

Concentration Inequalities for UCB

Prop 5 [Waudby-Smith and Ramdas, 2020]: Let $L_i(\lambda) = L(Y_i, T_\lambda(X_i))$ and

$$\hat{\mu}_i(\lambda) = \frac{1/2 + \sum_{j=1}^{i} L_j(\lambda)}{1+i}, \quad \hat{\sigma}_i^2(\lambda) = \frac{1/4 + \Sigma_{j=1}^{i}(L_j(\lambda) - \hat{\mu}_j(\lambda))^2}{1+i},$$

$$\nu_i(\lambda) = \min\left\{1, \sqrt{\frac{2\log(1/\delta)}{n\hat{\sigma}_{i-1}^2(\lambda)}}\right\}, \quad \mathcal{K}_i(R;\lambda) = \prod_{j=1}^{i}\{1 - \nu_j(\lambda)(L_j(\lambda) - R)\}$$

Then,

$$\hat{R}_{\mathrm{WSR}}^{+}(\lambda) = \inf\left\{R \geq 0 \mid \max_{i \in [n]} \mathcal{K}_i(R;\lambda) \geq \frac{1}{\delta}\right\},$$

is a $(1-\delta)$ UCB.

Concentration Inequalities for UCB

Proof Sketch. Define $\mathcal{K}_i = \mathcal{K}_i(R(\lambda), \lambda)$ and the set of σ-fields $\mathcal{F}_i = \sigma(L_1(\lambda), L_2(\lambda), \cdots L_i(\lambda))$.

Show $\{\mathcal{K}_i : i \in [n]\}$ is a non-negative martingale with respect to filtration $\{\mathcal{F}_i : i \in [n]\}$.

$$\mathbb{E}[\mathcal{K}_i \mid \mathcal{F}_{i-1}] = \mathcal{K}_{i-1}\mathbb{E}[1 - \nu_i(\lambda)(L_i(\lambda) - R(\lambda)) \mid \mathcal{F}_{i-1}] = \mathcal{K}_{i-1}$$

Use Ville's maximal inequality:

$$\mathbb{P}\left(\max_{i \in [n]} \mathcal{K}_i \geq \frac{1}{\delta}\right) \leq \delta.$$

Penn Engineering

Greedy Algorithm for Set-Predictors

Define *conditional risk density* ρ_x as $\rho_x(y, \mathcal{S}) = L(y, \mathcal{S}) p_{Y|X=x}(y)$. Then,

Algorithm 1:

Input: λ, estimate of conditional risk density $\hat{\rho}_x$, and stepsize $d\zeta$.

$\mathcal{T} \leftarrow \emptyset$

$\zeta \leftarrow \mathsf{C}$

while $\zeta > -\lambda$ **do**

$\quad \zeta \leftarrow \zeta - d\zeta$

$\quad \mathcal{T} = \mathcal{T} \cup \{y' \in \mathcal{T}^C \mid \hat{\rho}_x(y', \mathcal{T}) > \zeta\}$

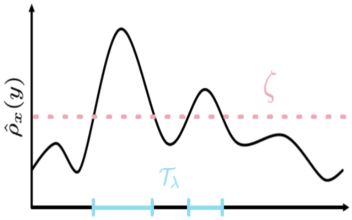
return $\mathcal{T}$

Penn Engineering

Optimality of Greedy Set-Predictors

Consider loss of the form: $L(y, \mathcal{S}) = L_y \mathbb{1}_{y \notin \mathcal{S}}$. Then,

$$T_\lambda(x) = \{y' \in \mathcal{Y} \mid \hat{\rho}(y', \emptyset) \geq \zeta(\lambda)\}$$



Theorem 7 *(Optimality of greedy set-predictors)*: Let $\mathcal{T}'$ be any set-predictor such that $R(\mathcal{T}') \leq R(T_\lambda)$. Then for the given loss, and assuming knowledge of the true probability density $p_{Y|X=x}(y)$, it follows that:

$$\mathbb{E}[|\mathcal{T}_\lambda(X)|] \leq \mathbb{E}[|\mathcal{T}'(X)|].$$

Proof. Since $R(\mathcal{T}') \leq R(T_\lambda)$,

$$\int_X \int_{\mathcal{T}'(x)} \rho_x(y) dy \, dP(x) \geq \int_X \int_{\mathcal{T}_\lambda(x)} \rho_x(y) dy \, dP(x)$$

$$\implies \int_X \int_{\mathcal{T}'(x) \setminus \mathcal{T}_\lambda(x)} \rho_x(y) dy \, dP(x) \geq \int_X \int_{\mathcal{T}_\lambda(x) \setminus \mathcal{T}'(x)} \rho_x(y) dy \, dP(x)$$

By construction of the greedy set-predictors, for any $y \in T_\lambda(x)$, $\rho_x(y) \geq \zeta(\lambda)$ and vice-versa for $y \notin T_\lambda(x)$. So,

$$\int_X \int_{\mathcal{T}'(x) \setminus \mathcal{T}_\lambda(x)} 1 dy \, dP(x) \geq \int_X \int_{\mathcal{T}_\lambda(x) \setminus \mathcal{T}'(x)} 1 dy \, dP(x)$$

which proves that $\mathbb{E}[|\mathcal{T}_\lambda(X)|] \leq \mathbb{E}[|\mathcal{T}'(X)|]$.

Experiments - Classification with class-varying loss

Loss: $L(y, \mathcal{S}) = L_y \mathbb{1}_{y \notin \mathcal{S}}$

Parametric Set of Functions: $\mathcal{T}_\lambda(x) = \{y \mid \hat{\pi}_x(y) > -\lambda\}$ for some trained classifier $\hat{\pi}_x : \mathcal{Y} \to [0, 1]$ and $\lambda \in [-1, 0]$.

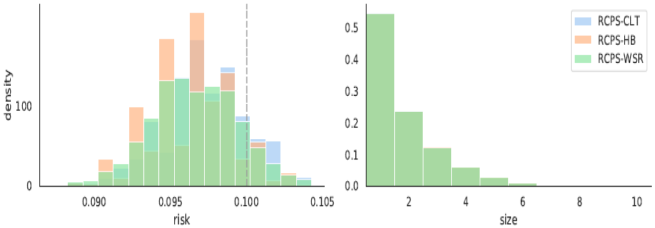


Figure: Plots of risk and prediction-set sizes across 100 different splits of data.

Experiments - Multi-Label Classification

Since each image can take several labels at once, both y and $\mathcal{S}$ have the same domain $\mathcal{D} = 2^{\{1,2,\cdots,K\}}$.

Loss: $L(y, \mathcal{S}) = 1 - \frac{|y \cap \mathcal{S}|}{|y|}$

Parametric Set of Functions:

$\mathcal{T}_\lambda(x) = \{z \in \{1, 2, \cdots K\} \mid \hat{\pi}_x(z) > -\lambda\}$ for some trained classifier $\hat{\pi}_x : \mathcal{Y} \to [0, 1]$, and $\lambda \in [-1, 0]$.

Figure: Examples of generated prediction sets for multi-label classification.

Experiments - Multi-Label Classification

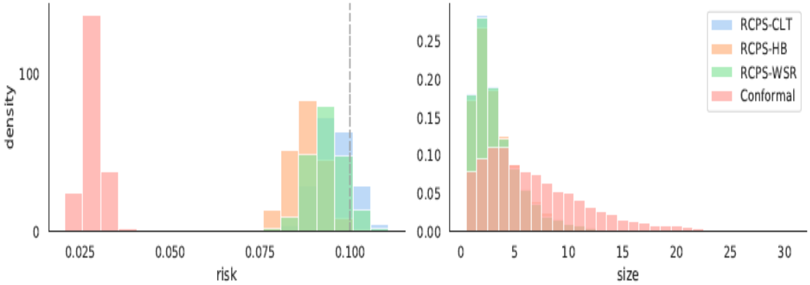


Figure: Plots of risk and prediction-set sizes across 1000 different splits of data.

Experiments - Image Segmentation

Define a connected components function $h : \mathcal{Y} \to 2^{\mathcal{Y}}$.

Loss: $L(y, \mathcal{S}) = \frac{\sum_{y' \in h(y)} |y' \setminus S| / |y'|}{|h(y)|}$

Parametric Set of Functions: $\mathcal{T}_\lambda = \{(i, j) \mid\mid \hat{f}(x)_{i,j} \geq -\lambda\}$ for a learned model $\hat{f} : \mathbb{R}^{d_1 \times d_2} \to [0, 1]^{d_1 \times d_2}$, and $\lambda \in [-1, 0]$.

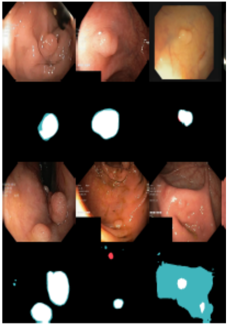


Figure: Examples of generated prediction sets for image segmentation.

Experiments - Image Segmentation

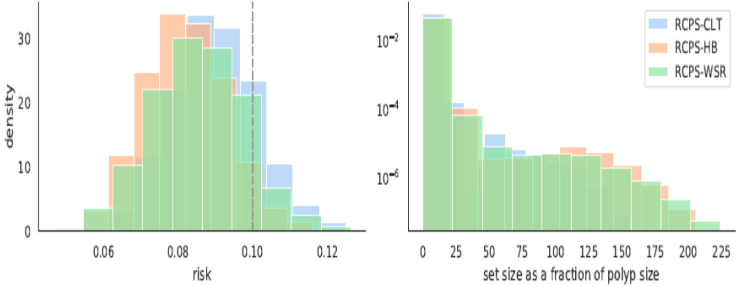


Figure: Plots of risk and (normalized) prediction-set sizes across splits of data.

Extensions / Further Work

- Ranking - define loss as a function of several points.
  - Learn ranking rule $\hat{r} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ from training data.
  - For $y_1, y_2 \in \{1, 2, \cdots K\}$ and $\mathcal{S} \in 2^{\mathbb{R}}$,

    $$L(y_1, y_2, \mathcal{S}) = \mathbb{1}_{\{\sup \mathcal{S} < 0\}} \mathbb{1}_{\{y_1 > y_2\}} + \mathbb{1}_{\{\inf \mathcal{S} > 0\}} \mathbb{1}_{\{y_1 < y_2\}}$$
  - Select $T_\lambda : \mathcal{X} \times \mathcal{X} \to 2^{\mathbb{R}}$ from some parametrized collection of set-predictors.
- Adversarial Robustness
  - Encode possible error due to perturbations into loss function.
  - $R^{(\mathrm{rob})}(\mathcal{T}) = \mathbb{E}\left[\sup_{x' \in \mathcal{B}_\epsilon(X)} L(Y, \mathcal{T}(x'))\right]$

References

Bates S., Angelopoulos A., Lei L., Malik J., and Jordan M. I., "Distribution-Free, Risk-Controlling Prediction Sets."

W. Hoeffding, "Probability inequalities for sums of bounded random variables."

V. Bentkus, "On Hoeffding's inequalities."

I. Waudby-Smith and A. Ramdas, "Variance-adaptive confidence sequences by betting."

V. Vovk, "Conditional Validity of inductive conformal predictors."

Penn Engineering