Calibrated Out-of-Distribution Detection with Conformal P-values

Lihua Lei (@lihua_lei_stat)

Department of Statistics, Stanford University

ICML 2021 Workshop on Uncertainty & Robustness in Deep Learning

Collaborators

Stephen Bates

Emmanuel Candès

Yaniv Romano

Matteo Sesia

Score-based out-of-distribution detection

Idea: learn a "resemblance" score $\hat{s}(X)$ from the in-distribution data and threshold it

- Classical methods:
  - One-class SVM Schölkopf, Williamson, Smola, Shawe-Taylor, and Platt ('99)
  - Isolation Forest Liu, Ting, and Zhou. ('08)
  - ...
- Deep OOD detection:
  - Softmax score Hendrycks and Gimpel ('17), ODIN Liang, Li, and Srikant ('18)
  - Mahalanobis distance-based confidence score Lee, Lee, Lee, and Shin ('18)
  - ...

Score-based out-of-distribution detection

Idea: learn a "resemblance" score $\hat{s}(X)$ from the in-distribution data and threshold it

- Classical methods:
  - One-class SVM Schölkopf, Williamson, Smola, Shawe-Taylor, and Platt ('99)
  - Isolation Forest Liu, Ting, and Zhou. ('08)
  - ...
- Deep OOD detection:
  - Softmax score Hendrycks and Gimpel ('17), ODIN Liang, Li, and Srikant ('18)
  - Mahalanobis distance-based confidence score Lee, Lee, Lee, and Shin ('18)
  - ...
- This talk: statistical techniques to **calibrate the threshold** with certain **Type-I error control**
- Can wrap around any score-based OOD detection method (**complementarity**)

Calibrated out-of-distribution detection

Given $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$ for some unknown P, test $H_j : X_{n+j} \sim P$, $j = 1, \ldots, m$

Calibrated out-of-distribution detection

Given $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$ for some unknown P, test $H_j : X_{n+j} \sim P$, $j = 1, \ldots, m$

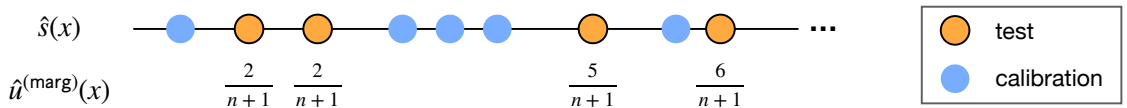
What statistical error(s) to control?

- Type-I error for a single test point X_{n+1}: $\mathbb{P}(X_{n+1} \text{ is detected as OOD}) \leq 5\%$ if $X_{n+1} \sim P$
- Type-I error for testing the global null $\cap_{j=1}^{m} H_j$
- False discovery rate (FDR, i.e., 1 - Expected Precision) for $H_1, \ldots, H_m$
- ...

Marginal conformal p-value (Vovk et al. '03, '05)

(Split-)conformal p-values with "resemblance" score $\hat{s}(\cdot)$ obtained from an independent training set

$$p(x) = \frac{\text{Rank of } \hat{s}(x) \text{ in the set } \{\hat{s}(x), \hat{s}(X_1), \ldots, \hat{s}(X_n)\}}{n+1}$$



Statistical properties of marginal conformal p-value

- $\mathbb{P}(p(X_{n+1}) \leq \alpha) \leq \alpha$ if $X_{n+1} \sim P \implies \alpha$ is a calibrated threshold for a single point

- Corrected Fisher's combination test for testing whether there is one OOD sample:

$$\mathbb{P}\left(\underbrace{-2\sum_{i=1}^{m} \log\left[p(X_{n+i})\right]}_{\text{aggregation of evidence}} \geq \underbrace{\chi^2(2m; 1-\alpha)\sqrt{1+\frac{m}{n}} - 2\left(\sqrt{1+\frac{m}{n}} - 1\right)m}_{\text{calibrated threshold}} \right) \leq \alpha$$

- Benjamini-Hochberg (BH) procedure for controlling FDR: $\mathbb{E}[\text{FP}/(\text{TP}+\text{FP})] \leq \alpha$,

$$\text{Claim } X_{n+i} \text{ as OOD if } p(X_{n+i}) \leq \underbrace{\frac{\alpha R_{\text{BH}}}{m}}_{\text{calibrated threshold}}$$

Conditional error control

- Marginal errors average over both training data $\{X_1, \ldots, X_n\}$ and test data $\{X_{n+1}, \ldots, X_{n+m}\}$

$$\mathbb{E}_{X_1,\ldots,X_n,X_{n+1},\ldots,X_{n+m}}[\mathrm{Err}] \leq \alpha$$

- Conditional error control (a.k.a. PAC-learning):

$$\mathbb{P}_{X_{n+1},\ldots,X_{n+m}}(\mathrm{Err} \leq \alpha \mid X_1, \ldots, X_n) \geq 1 - \delta$$

- Conditional error control might be more appropriate for inductive OOD detection

Conditional-calibrated p-values

Adjusting conformal p-value as $h \circ p(x)$ such that, if X_{n+1} is in-distribution, then

$$h \circ p(X_{n+1}) \succeq \mathrm{Unif}([0,1]) \mid X_1, \ldots, X_n, \quad \text{with prob. } 1-\delta$$

Conditional-calibrated p-values

Adjusting conformal p-value as $h \circ p(x)$ such that, if X_{n+1} is in-distribution, then

$$h \circ p(X_{n+1}) \succeq \mathrm{Unif}([0,1]) \mid X_1, \ldots, X_n, \quad \text{with prob. } 1-\delta$$

$h \circ p(X_{n+1}), \ldots, h \circ p(X_{n+m})$ are independent conditional on $X_1, \ldots, X_n$

$\rightsquigarrow$ Conditional validity of single point testing, Fisher's combination test, BH procedure...

Conditional-calibrated p-values via generalized Simes' inequality

Goal: preserving small marginal p-values, i.e., $h(u)$ is small for small u

Conditional-calibrated p-values via generalized Simes' inequality

Goal: preserving small marginal p-values, i.e., $h(u)$ is small for small u

Theorem (Generalized Simes Inequality, Sarkar ('08))

h_{Simes} yields a valid calibration-conditional p-value where

$$h_{Simes}\left(1-\frac{i}{n+1}\right) = 1-\delta^{1/k}\left(\frac{i\cdots(i-k+1)}{n\cdots(n-k+1)}\right)^{1/k}.$$

Initially used for multiple hypothesis testing; used here in a non-standard way

Type-II error?

- Type-II error control is possible when labelled OOD samples are available
- Unsupervised approach: treat OOD samples as “in-distribution”
- Supervised approach: use probabilistic classification to replace the resemblance score

Augmenting machine learning with statistical reasoning

Empowering statistical reasoning with machine learning