Distribution Free Prediction Sets

Jing Lei

Department of Statistics, CMU

Collaborators: James Robins (Harvard), Larry Wasserman (CMU)

July 2012

Outline

- Prediction sets: background and challenges.
- A new approach to nonparametric estimation of prediction sets.
- Extensions to more general statistical learning problems.

Prediction Sets: Motivation and Definition

- Prediction: observe $Y_1, \dots, Y_n \overset{iid}{\sim} P$, $Y_i \in \mathbb{R}^d$. $Y_{n+1} = ?$
- Want intervals (sets) rather than point predictions.
- Prediction set: $C \subset \mathbb{R}^d$ such that, for a given $\alpha \in (0,1)$,

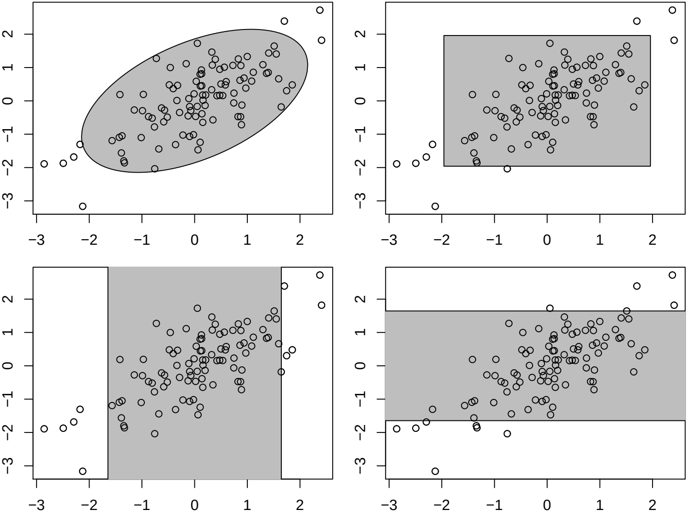
$$\mathbb{P}(Y_{n+1} \in C) = P(C) \geq 1 - \alpha.$$

- Look for estimator $C_n = C_n(Y_1, \dots, Y_n)$, such that

$$P(C_n) \geq 1 - \alpha$$

  holds with some probabilistic guarantee.
- Applications: anomaly detection, quality control, clustering.

Prediction Sets: Examples



How to Evaluate Prediction Sets?

- Validity: C_n has the desired coverage under P.

  Finite sample validity: $\mathbb{E}(P(C_n)) \geq 1-\alpha$ for all $n > 0$ and all P.

- Efficiency: C_n has small Lebesgue measure.
  1. "Oracle set": $C^{(\alpha)} = \{y : p(y) \geq t_\alpha\}$, where t_α is chosen such that $P(C^{(\alpha)}) = 1-\alpha$.
  2. Asymptotic efficiency: $\mu\left(C_n \triangle C^{(\alpha)}\right) \xrightarrow{P} 0$, where μ is the Lebesgue measure.

- Existing methods such as plug-in density level sets (Hyndman 1996; Cadre 2006) do not give finite sample validity and asymptotic efficiency at the same time.

Conformal Prediction Sets

We construct prediction sets that

1. always have finite sample validity *with no assumptions* on P;
2. are asymptotically efficient with near optimal rate under standard smoothness conditions;
3. can be easily implemented with simple parameter tuning.

The approach is based on a novel combination of conformal prediction (Vovk et al, 2009) with statistical principles.

Basic Idea

- Let $\mathbf{Y} = (Y_1, ..., Y_n)$. For any $y \in \mathbb{R}^d$, let $\mathbf{Y}^y$ be the augmented data $(Y_1, ..., Y_n, Y_{n+1})$ with $Y_{n+1} = y$.

Basic Idea

- Let $\mathbf{Y} = (Y_1, ..., Y_n)$. For any $y \in \mathbb{R}^d$, let $\mathbf{Y}^y$ be the augmented data $(Y_1, ..., Y_n, Y_{n+1})$ with $Y_{n+1} = y$.
- Let $g(\mathbf{Y}, y) \in \mathbb{R}^1$ be a function that is symmetric in each element of $\mathbf{Y}$. E.g: $g(\mathbf{Y}, y) = -|\bar{\mathbf{Y}} - y|$.

Basic Idea

- Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$. For any $y \in \mathbb{R}^d$, let $\mathbf{Y}^y$ be the augmented data $(Y_1, \ldots, Y_n, Y_{n+1})$ with $Y_{n+1} = y$.
- Let $g(\mathbf{Y}, y) \in \mathbb{R}^1$ be a function that is symmetric in each element of $\mathbf{Y}$. E.g: $g(\mathbf{Y}, y) = -|\bar{\mathbf{Y}} - y|$.
- $g(\mathbf{Y}^y, Y_i)$, $1 \leq i \leq n+1$, are called the conformity scores.

Basic Idea

- Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$. For any $y \in \mathbb{R}^d$, let $\mathbf{Y}^y$ be the augmented data $(Y_1, \ldots, Y_n, Y_{n+1})$ with $Y_{n+1} = y$.
- Let $g(\mathbf{Y}, y) \in \mathbb{R}^1$ be a function that is symmetric in each element of $\mathbf{Y}$. E.g: $g(\mathbf{Y}, y) = -|\bar{\mathbf{Y}} - y|$.
- $g(\mathbf{Y}^y, Y_i)$, $1 \leq i \leq n+1$, are called the conformity scores.
- Rank $g(\mathbf{Y}^y, Y_{n+1}) = g(\mathbf{Y}^y, y)$ among all $n+1$ scores:

$$\pi_n(y) = (n+1)^{-1} \sum_{i=1}^{n+1} \mathbb{I}\left[g(\mathbf{Y}^y, Y_i) \leq g(\mathbf{Y}^y, Y_{n+1})\right].$$

Basic Idea

- Let $\mathbf{Y} = (Y_1, ..., Y_n)$. For any $y \in \mathbb{R}^d$, let $\mathbf{Y}^y$ be the augmented data $(Y_1, ..., Y_n, Y_{n+1})$ with $Y_{n+1} = y$.
- Let $g(\mathbf{Y}, y) \in \mathbb{R}^1$ be a function that is symmetric in each element of $\mathbf{Y}$. E.g: $g(\mathbf{Y}, y) = -|\bar{\mathbf{Y}} - y|$.
- $g(\mathbf{Y}^y, Y_i)$, $1 \le i \le n+1$, are called the conformity scores.
- Rank $g(\mathbf{Y}^y, Y_{n+1}) = g(\mathbf{Y}^y, y)$ among all $n+1$ scores:

$$\pi_n(y) = (n+1)^{-1} \sum_{i=1}^{n+1} \mathbb{I}\left[g(\mathbf{Y}^y, Y_i) \le g(\mathbf{Y}^y, Y_{n+1})\right].$$

- Conformal prediction region: $C_n = \{y \in \mathbb{R}^d : \pi_n(y) \ge \alpha\}$.

Conformal Prediction with Kernel Density

- Kernel density:

$$\hat{p}_h(u) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{u - Y_i}{h}\right), \ \forall\, u \in \mathbb{R}^d.$$

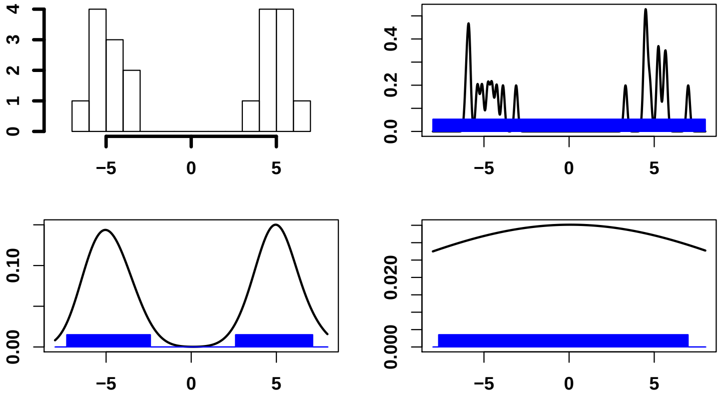
- Kernel density using augmented data $\mathbf{Y}^y$:

$$\begin{aligned}\hat{p}_h^y(u) &= \frac{1}{(n+1)h^d} \sum_{i=1}^{n+1} K\left(\frac{u - Y_i}{h}\right) \\ &= \frac{n}{n+1}\hat{p}_h(u) + \frac{1}{(n+1)h^d} K\left(\frac{u - y}{h}\right).\end{aligned}$$

- Define $g(\mathbf{Y}^y, Y_i) = \hat{p}_h^y(Y_i)$, for $1 \leq i \leq n+1$.

Example: Gaussian Mixture

Gaussian mixture with $n = 20$, $\alpha = 0.1$, $h = 0.1$, 1, and 10.



Bandwidth tuning: minimize the prediction set.

Theoretical Properties

Theorem

C_n has finite sample validity

$$\mathbb{E}(P(C_n)) \geq 1-\alpha, \quad \text{for all } n \text{ and } P.$$

Moreover, it is asymptotically efficient under regularity conditions:

$$\mu\left(C_n \triangle C^{(\alpha)}\right) = O_P\left[\left(\frac{\log n}{n}\right)^{\frac{\beta\gamma}{2\beta+d}\wedge\frac{1}{2}}\right],$$

where β and γ are smoothness parameters of p.

Proof of validity is extremely simple and uses symmetry.

Proof of efficiency is based on approximating C_n by plug-in level sets.

Further Extensions

1. Prediction with covariates: conformal nonparametric regression (Lei and Wasserman, 2012).
2. Other choices of conformity scores: Gaussian mixture density; pseudo density (Lei, Rinaldo, and Wasserman 2012)
3. Tuning parameter selection by minimizing conformal sets (e.g., high-dimensional regression, k-means clustering).
4. Classification: connection to classification with rejection (ongoing, with L. Wasserman).

Questions?