Predictive Inference

Naive prediction interval

Define $\hat{F}_n(\cdot)$ to be the empirical CDF of fitted residuals $|Y_i - \hat{\mu}(X_i)|$ $(i=1,\dots,n)$

$C(X_{n+1}) = [\hat{\mu}(X_{n+1}) - \hat{F}_n^{-1}(1-\alpha),\ \hat{\mu}(X_{n+1}) + \hat{F}_n^{-1}(1-\alpha)]$

$\hookrightarrow (1-\alpha)$-quantile

need many samples for accurate estimation

(Conformal Prediction)

Basic CP (given learned estimator, to do predictive inference)

conformality score: $S(x,y)$ eg. $S(x,y) = 1 - \hat{f}(x)_y$ for given estimator $\hat{f}$

1. $s_i = S(X_i, Y_i),\ i=1,\dots,n$. $n = \#$ calib set
2. $\hat{q}_{1-\alpha}$ = the $\lceil (n+1)(1-\alpha) \rceil / n$ quantile of the conformal scores $\{s_i\}_{i=1}^n$
3. prediction set for new input x_{n+1}:

$$T(X) = \{y : \hat{q} > S(X, y)\}$$

Full CP. $\hat{f}$ is trained with $\{(x_i, y_i)_{i=1}^n, (x_{n+1}, y)\}$ (for each $y \in \mathcal{Y}$, train a different $\hat{f}$)

compute $\sigma_i = S(\hat{f}, x_i, y_i)$ $i=1,\dots,n+1$ $\pi(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}\{\sigma_i \le \sigma_{n+1}\}$

$\Rightarrow T_\alpha(X_{n+1}) = \{y \in \mathcal{Y} : \pi(y) > \alpha\}$

Thm. $(X_i, Y_i): i=1,\dots,n+1$ are exchangeable, $\hat{q}$ is $\lceil (n+1)(1-\alpha) \rceil / n$ quantile of $\{s_i\}_{i=1,\dots,n}$. Then (smaller score is better)

$$P(Y_{n+1} \in T(X_{n+1})) \ge 1-\alpha$$

Proof: $\hookrightarrow S_{n+1} = S(X_{n+1}, Y_{n+1}) < \hat{q} = S_{\lceil (n+1)(1-\alpha) \rceil}$

Changeability $\Rightarrow P(S_{n+1} < S_k) = k/(n+1)$, $S_1 < \dots < S_n$

$\Rightarrow P(S_{n+1} < S_{\lceil (n+1)(1-\alpha) \rceil}) = \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1} \ge 1-\alpha$

(if score is continuous, $P(\cdots) \le 1-\alpha + \frac{1}{n+1}$)

Thm (iid) $\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Y_i \in C(X_i)\} - (1-\alpha)\right| \ge \varepsilon\right) \le 2\exp\left(-cn^2(\varepsilon - 2/n)_+^2\right)$

$n = \#$ calib set

Bonferroni argument

Multiple Split — Split the data M times, yielding M intervals: $C_1, \dots, C_M$ with level $1-\alpha/M$

$$\Rightarrow C^{(M)}(x) = \bigcap_{j=1}^{M} C_j(x)$$

Thm. $C^{(M)}(X)$ is wider than $C(X)$.

Full CP

$J(X_{n+1}) \leftarrow \phi$

For $y \in \mathcal{Y}$:

$\quad \hat{\mu}_y = A(\{(X_i, y_i)\}_{i=1}^{n}, (X_{n+1}, y)\})$

$\quad$ if $S(X_{n+1}, y, \hat{\mu}_y) \le \text{Quantile}(1-\alpha, \{S(X_i, y_i, \hat{\mu}_y)\}_{i=1}^{n} \cup S(X_{n+1}, y, \hat{\mu}_y))$

$\quad\quad$ then add y into $J(X_{n+1})$

output $J(X_{n+1})$

Jackknife prediction

$\hat{\mu}^{(-i)} = A(\{(X_\ell, Y_\ell), \ell \neq i\}), \; i \in [n]$ leave-one-out

$R_i = S(x_i, y_i, \hat{\mu}^{(-i)})$

$\hat{q}$ = the $\lceil (1-\alpha)\cdot n \rceil$-th smallest value in $\{R_i\}_{i=1}^{n}$

$C_{jack}(x) = \{y \mid S(x, y, \hat{\mu}) < \hat{q}\}$ no gaurantee for out-of-sample coverage

Property: $\mathbb{P}(Y_i \in C_{jack}(X_i)) \ge 1-\alpha, \; i = 1, \dots, n$

Adaptive.CP

$S(X,Y) = \sum_{j=1}^{k} \hat{f}(X)_{\pi_j}$, where $Y = \pi_k$, $\hat{f}(X)_{\pi_1} \geq \hat{f}(X)_{\pi_2} \geq \cdots$

(utilize the softmax out of all classes instead of only true class)

Output: $\mathcal{T}(X) = \{\pi_1, \ldots, \pi_k\}$, $k = \inf\{k : \sum_{j=1}^{k} \hat{f}(X)_{\pi_j} \geq \hat{q}\}$

$\hookrightarrow = \{y : S(x,y) \leq \hat{q}\}$

$\hookrightarrow \sum_{j=1}^{k} \hat{f}(x)_{\pi_j} \leq \hat{q}$, $\pi_k = y$, $\hat{f}(x)_{\pi_1} \geq \hat{f}(x)_{\pi_2} \geq \cdots$

$k = \max\{k : \sum_{j=1}^{k} \hat{f}(x)_{\pi_j} < \hat{q}\}$

$S(x,y) \geq \hat{q}$ 且S尽可能小 $k=m$, $\sum_{j=1}^{m-1} \hat{f}(x)_{\pi_j} < \hat{q} < \sum_{j=1}^{m} \hat{f}(x)_{\pi_j}$ …

统计希望保守 $S(x,m) \leq \hat{q}_{1-\alpha}$

$\downarrow$

$m \in \mathcal{T}(X)$ error I ≤ 0.05 $\sum_{j=1}^{k} \hat{f}(x)_{\pi_j} \leq \hat{q}$, $\pi_k = m$

Quantile regression

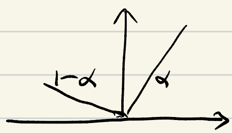
Learn γ quantile of $Y|X$ for any X: $t_\gamma(x) \leftarrow \hat{t}_\gamma(x)$

expect the interval $[\hat{t}_{\alpha/2}(x), \hat{t}_{1-\alpha/2}(x)]$ to have $1-\alpha$ coverage

but we don't know how accurate the quantile regression is.

pinball loss: $\rho_\gamma(f, Y) = (Y - f(x)) \cdot \gamma \cdot \mathbb{1}\{Y > f(x)\} + (f(x) - Y) \cdot (1-\gamma) \cdot \mathbb{1}\{Y \le f(x)\}$

e.g. $\rho_{0.5} = |f(x) - Y|$



$$\Rightarrow l_\alpha(y, f_{\alpha_{lo}}(x), f_{\alpha_{hi}}(x)) = \rho_{\alpha_{lo}}(y, f_{\alpha_{lo}}(x)) + \rho_{\alpha_{hi}}(y, f_{\alpha_{hi}}(x))$$

(low: α_{lo}, high: α_{hi})

interval score loss

$$l_\alpha^{int}(y, f_{\alpha_{lo}}(x), f_{\alpha_{hi}}(x)) = (\underbrace{f_{\alpha_{hi}}(x) - f_{\alpha_{lo}}(x)}_{\text{encourage short interval}}) + \frac{2}{\alpha}(f_{\alpha_{lo}}(x) - y)\,\mathbb{1}\{y < f_{\alpha_{lo}}(x)\} + \frac{2}{\alpha}(y - f_{\alpha_{hi}}(x))\,\mathbb{1}\{y > f_{\alpha_{hi}}(x)\}$$

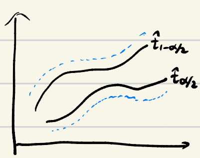
$$\mathbb{E}_{\alpha \sim U(0,1)}[l_\alpha^{int}(\cdot)]$$

Conformalized Q-regression

$s(x, Y) = \max\{\hat{t}_{\alpha/2}(X) - Y,\ Y - \hat{t}_{1-\alpha/2}(X)\}$

still, set $\hat{q} = \text{quantile}(s_1, \dots, s_n;\ \lceil (n+1)(1-\alpha) \rceil / n)$

we have $\mathcal{T}(X) = [\hat{t}_{\alpha/2}(X) - \hat{q},\ \hat{t}_{1-\alpha/2}(X) + \hat{q}]$



$$\left(\begin{aligned} s(x,y) < \hat{q} &\Leftrightarrow \max\{\hat{t}_{\alpha/2}(x) - y,\ y - \hat{t}_{1-\alpha/2}(x)\} < \hat{q} \\ &\Leftrightarrow \begin{cases} \hat{t}_{\alpha/2}(x) - y < \hat{q} \\ y - \hat{t}_{1-\alpha/2}(x) < \hat{q} \end{cases} \Leftrightarrow \hat{t}_{\alpha/2}(x) - \hat{q} < y < \hat{t}_{1-\alpha/2}(x) + \hat{q} \end{aligned} \right)$$

Estimate Uncertainty

$\hat{\mu}(x), \hat{\sigma}(x)$: use
- Gausian likelihood loss
- ensemble
- dropout
- random input perturbation
- ...

Or, train $\hat{r}(x)$ to fit $|Y-\hat{f}(x)|$, expect $[\hat{f}-\hat{r}, \hat{f}+\hat{r}]$ to have good coverage

$$s(X,Y) = \frac{|Y-\hat{f}(x)|}{u(x)}$$, u could be $\hat{\sigma}(\cdot)$ or $\hat{r}(\cdot)$

↳ multiplicative correction factor of uncertainty scalar

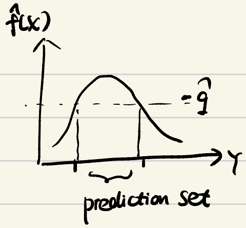
$$T(x) = [\hat{f}(x) - u(x)\hat{q}, \hat{f}(x) + u(x)\hat{q}]$$

Conformal Bayes

Say $\hat{f}(x)_Y$ is the estimated value of posterior distribution:

$$S(X,Y) = -\hat{f}(X)_Y$$

seems same as standard CP



Evaluation

Coverage

$$C_j = \frac{1}{n'} \sum_{i=1}^{n'} \mathbb{1}\{ Y_i^{(val)} \in \mathcal{T}(X_i^{(val)}) \}, \quad j=1,\dots,R, \quad n' = \#\text{ valid set}$$

(random split R times)

the histogram of these C_j should be centered at $1-\alpha$.

center or slightly larger

保证 coverage >0.95下，让 set size 尽可能小

Set size

plot the histograms of set size

① small average set size $\Rightarrow$ precise

② wide spread of the histogram $\Rightarrow$ the sets adapts to the difficulty of examples

(就是不要成点分布)

Evaluate Adaptiveness

return larger sets for harder examples

Conditional coverage: $P(Y \in \mathcal{T}(X) \mid X) \geq 1-\alpha$ ← CP is not gaurantee to satisfy!

stronger than marginal coverage

metric

$$\min_{g \in \{1,\dots,G\}} \frac{1}{|I_g|} \sum_{i \in I_g} \mathbb{1}\{ Y_i \in \mathcal{T}(X_i) \}$$

g indexes the group

I_g contains examples in group g

this should be $1-\alpha$ (slightly larger)

反正比 $1-\alpha$ 小太多不行

① (feature-stratified coverage)

FSC metric

group by value of some feature

② (set-stratified coverage)

SSC metric

general grouping

$L=0 \Rightarrow \hat{C}(x)=\{y_0\} \Rightarrow \underline{\mathbb{P}(Y=y_0 \mid X=x)=1-\alpha}$

↳ $Y|X=x$ 连续

Conditional coverage

Thm. $Y|X=x$ is continuous, $\mathbb{P}(Y\in\hat{C}(X) \mid X=x)=1-\alpha$ for all x

then $V \perp\!\!\!\perp L$, where $V=\mathbb{1}\{Y\in\hat{C}(x)\}$, $L=|\hat{C}(x)|$ ⟶ regression problem

Proof.

$$\mathbb{P}(V=v) = \int \underbrace{\mathbb{P}(V=v \mid X=x)}_{\text{与}x\text{无关},\ =\begin{cases}1-\alpha & \text{if } v=1\\ \alpha & \text{if } v=0\end{cases}} P_X(x)\,dx = \begin{cases}1-\alpha & \text{if } v=1\\ \alpha & \text{if } v=0\end{cases}$$

$$\mathbb{P}(V=v \mid L=\ell) = \int_{\{x:\,L=\ell\}} \mathbb{P}(V=v \mid X=x, L=\ell)\, P_{X|L}(x|\ell)\,dx$$

$$= \int_{\{x:\,L=\ell\}} \mathbb{P}(V=v \mid X=x)\, P_{X|L}(x|\ell)\,dx$$

$$= \cdots = \begin{cases}1-\alpha & \text{if } v=1\\ \alpha & \text{if } v=0\end{cases}$$

$\Rightarrow \mathbb{P}(V=v \mid L=\ell) = \mathbb{P}(V=v)$

$\Rightarrow V \perp\!\!\!\perp L$