Out line

1. Open set classification
2. EVT
   - ① "from mean to extremes"
   - ② limit distribution
   - ③ MDA and GEV
   - ④ GPD
3. EVM algorithm
   - ① algorithm
   - ② limitation
4. The GPD classifier
   - ① Settings
   - ②-Th 3
   - -Th 4 (cow)
   - -Hill estimator
   - -Th 5
   - ③ algorithm
5. GEV classifier
6. Application

---

1. Open set classification (开集)

Supervised learning { regression, classification.

- Data: $\{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^p$, $y_i \in \{C_1, \cdots, C_J\}$
- classifier $f: X \to y$, $f \in F$

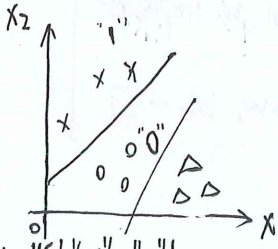
{ training set : validation set : testing set :

$\hat{f}$ hyper parameters

Recall, Precision, AUC, ...

⇒ samples are drawn from the same distribution.

eg1. $P = 2$, $y_i \in \{$"O", "X"$\}$



training set : $y \in$ {"X", "O"}

testing set { ~~closed set~~ {"X", "O"} — closed set

{"X", "O", "△", ···} open set

unknown classes

all possible classes are present during the training phase.

eg 2. face recognition.

- Identification: Figure → ID 肯定是, 是谁!
- Verification: 是否是员工?

- Open set classification
  - ① distinguish between known and unknown classes.
    - — novelty detection : rare classes
    - — outlier detection
  - ② work in an incremental way
    - — life-long learning : new classes eg. KNN
  - ③ few hyper parameters.
    - ∴ cross-validation is not available.

"we know nothing of the unknown classes."

"why extreme value theory?"

"extreme features", rather than the average ones, are the most important for discriminating between different objects."

CV.

eg. A tire of a car can fail in two ways : $X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} F$

- partial sums exceed some threshold

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - EX_i}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

- partial maxima ... [Block Maxima]

$M_n = \max(X_1, \cdots, X_n)$ → unusually large (or small) levels.

$$\frac{M_n - a_n}{b_n} \xrightarrow{d} G(x) \quad (n \to \infty)$$

the behavior of X_i is usually unknown.
↓
exact calculation on M_n is impossible
↓
suitable assumptions, $n \to \infty$, $M_n \to \square$.

a_n, b_n ?

let $x^* := \sup\{x : F(x) < 1\}$

then $\max(X_1, \cdots, X_n) \xrightarrow{P} x^*$, $n \to \infty$

since $P\{M_n \le x\} = P\{X_1 \le x, \cdots, X_n \le x\}$

$= F^n(x) \to 0$, for $x < x^*$

$\to 1$, for $x \ge x^*$

we need a normalization.

Suppose $\exists\ a_n > 0$, $b_n \in \mathbb{R}$, $n = 1, 2, \cdots$

such that $\frac{M_n - b_n}{a_n}$ has a nondegenerate limit distribution

as $n \to \infty$,

i.e., $\lim_{n\to\infty} F^n(a_n x + b_n) = G(x)$

The class of F satisfying this condition is called the

Maximum domain of attraction of G

MDA (G)

$F \in MDA(G)$

~~$G(x) = \exp\{-[1+\xi x - \mu]\}$~~

$$G(z) = \exp\left\{-\left[1+\xi\frac{z-\mu}{\sigma}\right]^{-\frac{1}{\xi}}\right\},\ z \in \mathbb{R} : 1+\xi\frac{z-\mu}{\sigma} > 0$$

$\xi, \mu \in \mathbb{R}$, $\sigma > 0$. (shape, location, scale)

① $\xi = 0$, Gumbel

$$G(z) = \exp\left\{-\exp\left[-\frac{z-\mu}{\sigma}\right]\right\} \quad -\infty < z < \infty$$

② $\xi > 0$, Frechet

$$G(z) = \begin{cases} \exp\left\{-\left[\frac{z-\mu}{\sigma}\right]^{-\frac{1}{\xi}}\right\}, & z > \mu \\ 0, & z \le \mu \end{cases}$$

③ $\xi < 0$, reversed Weibull

$$G(z) = \begin{cases} \exp\left\{-\left[-\frac{z-\mu}{\sigma}\right]^{-\frac{1}{\xi}}\right\}, & z < \mu \\ 1, & z \ge \mu \end{cases}$$

EVT-II.

Th1. Block Maxima

$G(x) = \exp\{-[1+\xi\frac{x-\mu}{\sigma}]^{-\frac{1}{\xi}}\}$, $x \in \mathbb{R}: 1+\xi\frac{x-\mu}{\sigma} > 0$

$\xi \in \mathbb{R}$: shape, $\mu \in \mathbb{R}$, $\sigma > 0$ | ξ: different representations of extreme value behavior.

① $\xi = 0$, Gumbel, Type 1

$G_1(x) = \exp\{-\exp[-\frac{x-\mu}{\sigma}]\}$

Exponential decay in the tail of $F(x)$

$F \in MDA(G_1)$ iff $\lim_{x\to\infty} \frac{1-F(t+R(x))}{1-F(x)} = e^{-t}$, for all $t>0$.

② $\xi > 0$, Fréchet, Type 2

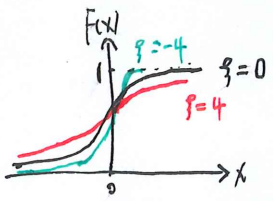
$F \in MDA(G_2)$ iff $\lim_{x\to\infty} \frac{1-F(tx)}{1-F(x)} = t^{-\frac{1}{\xi}}$, for all $t>0$.

Polynomial decay in the tail of $F(x)$. "fat tail"

③ $\xi < 0$, reversed Weibull, Type 3. (Lemma 1)

$F \in MDA(G_3)$ iff $\lim_{x\to x^*} \frac{1-F(x^*-\frac{1}{x})}{1-F(x^*-\frac{1}{x})} = t^{-\frac{1}{\xi}}$, for all $t>0$

$x^* = \sup\{x: F(x)<1\}$ $F(x)$ has a finite upper endpoint of x^*.



$P\{X > x_p\} = p$, $p = 0.01$: 百年一遇, 大坝.

Th2. Peak over threshold

$Y := X-u \mid X>u$: more efficient use of data.

$H(y) = 1-(1+\xi\frac{y}{\sigma})^{-\frac{1}{\xi}}$, $y>0$; $1+\xi\frac{y}{\sigma}>0$

$\tilde{\sigma} = \xi(u-\mu)$

⇓

need to choose u.

4. The extreme value machine (Rudd et al. 2018)

- $x_i \in \mathbb{R}^p$, $y_i \in \{C_1, \dots, C_N\}$, $i = 1, \dots, n$

  x_0 : a new point (known or unknown)

- $M^{(i)} = \min_{j: y_j \neq y_i} D_j^{(i)} = \min_{j: y_j \neq y_i} \frac{1}{2}\|x_i - x_j\|$.

  since $\overline{M}^{(i)}$ is bounded above by zero.

  that is, assuming $z^* = 0$. (lemma1)

$$\overline{M}^{(i)} = -M^{(i)} \xrightarrow{d} W^{(i)}(z) = \begin{cases} \exp\left\{-\left(\frac{\theta - z}{\sigma_i}\right)^{-\frac{1}{\xi}}\right\}, & \text{if } z < \theta \\ 1, & \text{if } z \geq \theta \end{cases}$$

(fit it to the k largest $-D_j^{(i)}$ for each point x_i)

- $(\hat{\sigma}_i, \hat{\xi})$, $\hat{W}^{(i)}$

"x_0 落入任一训练样本点的 Margin distance，即 unknown"

then, label x_0 as known if $\max_{i=1,\dots,n} \hat{W}^{(i)}(-\|x_0 - x_i\|) \geq \delta$

If x_0 is likely enough to fall in at least one of the margins, of the n training data, then label it as known.

---

limitation.

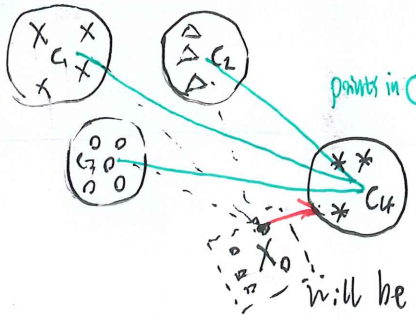
① $z^* = 0$: all classes in the training set share the same support.

(for each i) ⇒ It's impossible two classes are perfectly separated.

② Estimating the upper tail of $\overline{M}^{(i)}$ based on the k largest observed $-D_j^{(i)}$ for each x_i should rely on Th2. GP instead of GEV. peaks over threshold ✓ Block Maxima.

③ δ should be chosen by fixing the type-I error.

④ a non-justified premium : classes far from all the others.



points in C_4 will have a large margin distance.

x_0 will be misclassified to C_4

⇓

- We cannot use the geometry given by the known classes to do open set classification.

⇓

对 unknowns 什么都不假设

对 unknown - 无所所知，故不可假设距离关系！

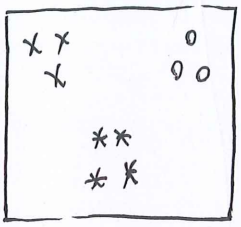
5. GPDC

Intuition: Test points that are extremely far from the training classes are more likely to be unknown.

$\{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^p$, $y_i \in \{C_1, \cdots, C_J\}$

x_0, $D_i = \|x_i - x_0\|$ (consider Euclidean distance for simplicity)

then the lower endpoint of the distribution D of the D_i $\begin{cases} = 0 : x_0 \text{ is known} \\ \neq 0 : \text{unknown.} \end{cases}$



training set

$\int_{x \in A} f_{C_j}(x)dx$: the probability of a point in C_j falls in the set $A \subseteq \mathbb{R}^p$

training data generating process: $f(x) = \sum_{j=1}^J w_j f_{C_j}(x)$, $w_j \in [0,1]$, $\sum_{j=1}^J w_j = 1$

↓ approximate directly f.

Ball $B_\delta(x_0)$

$P\{B_\delta(x_0)\} = P(D < \delta) = P(-D > -\delta)$

Th 3. Consider $D_1, \cdots, D_n$ and assume $x_0 \in \text{supp}(f)$, let $R = -D$.

↓ that is, (the upper endpoint of $-D$ is zero.) $r^* = \sup\{r : F(r) < 1\}$

$\hat{\xi}_n$ (MLE) Let $Y := R_{(n)} - u \mid R > u$, $H(y) = 1 - (1 + \xi \frac{y}{\sigma})^{-\frac{1}{\xi}}$, $y > 0$, $1 + \xi \frac{y}{\sigma} > 0$

$$h(y) = \frac{1}{\sigma}(1 + \xi \frac{y}{\sigma})^{-\frac{1}{\xi} - 1}$$

suppose $y_1, \cdots, y_k$ are k excesses of the threshold u,

since $\xi \neq 0$, $\log L(y_1, \cdots, y_k; \sigma, \xi) = -k \log \sigma - (1 + \frac{1}{\xi}) \sum_{i=1}^k \log(1 + \xi \frac{y_i}{\sigma})$

as $r^* = u - \frac{\sigma}{\xi} = 0 \Rightarrow \sigma = u\xi$ ~~$= -k \log \sigma - (1 + \frac{1}{\xi})$~~

$\Rightarrow$ $\log L(R_1, \cdots, R_n; \sigma, \xi) = -k \log u\xi - \frac{1}{\xi} \sum_{i=1}^k \log \frac{R_{(n+1-i)} - u}{u}$, where $R_{(n)} \geq R_{(n-1)} \geq \cdots \geq R_{(1)}$

$\hat{\xi}_n = \arg\max_\xi -k \log \xi - \frac{1}{\xi} \sum_{i=1}^k \log \frac{R_{(n+1-i)}}{u}$

$= \frac{1}{k} \sum_{i=1}^k \log \frac{R_{(n+1-i)}}{u}$

Th 4. let $u = R_{(n-k)}$, assume $k = k(n) \to \infty$, $\frac{k(n)}{n} \to 0$, as $k \to \infty$

the shape parameter of the distribution of $-D$ is $\xi = -\frac{1}{p}$,

and $\hat{\xi}_n \xrightarrow{P} -\frac{1}{p}$, where $p \in N$. $\quad \hat{\xi}_n = \frac{1}{k}\sum_{i=1}^{k} \log \frac{R_{(n+1-i)}}{u}$

proof.

① the distribution of $-D$ is in the max-domain of attraction of a GEV distribution with $\xi = -\frac{1}{p}$.

$R = -D$, $F_R \in MDA(GEV_\xi)$

$$P\{R > -\delta\} = P\{D < \delta\} = P\left(\frac{1}{D} > \frac{1}{\delta}\right)$$
$$= P\{B_\delta(x_0)\} = \int_{B_\delta(x_0)} f(x)\,dx$$
$$= f(x_0) V\{B_\delta(x_0)\} + o(1)$$
$$= f(x_0)\,\pi^{\frac{p}{2}} \cdot \delta^p / \Gamma\left(\frac{p}{2}+1\right) + o(1)$$

$$V\{B_\delta(x_0)\} = \pi^{\frac{p}{2}} \delta^p / \Gamma\left(\frac{p}{2}+1\right)$$

$$t^* := \sup\{t : F(t) < 1\} = 0$$

$$\lim_{t \to \infty} \frac{1 - F(t^* - \frac{t}{r})}{1 - F(t^* - \frac{1}{r})} = \lim_{r \to \infty} \frac{1 - F(-\frac{t}{r})}{1 - F(-\frac{1}{r})}$$
$$= \lim_{r \to \infty} \frac{f(x_0)\,\pi^{\frac{p}{2}} \cdot (-\frac{t}{r})^p + o(1)}{f(x_0)\,\pi^{\frac{p}{2}} \cdot (-\frac{1}{r})^p + o(1)}$$
$$= t^p, \text{ for all } t > 0$$

$\Rightarrow p = -\frac{1}{\xi}$, $\xi = -\frac{1}{p}$ $\qquad \xi = \frac{1}{p} > 0$

② The upper tail of $\frac{1}{D}$ is in the max-domain of attraction of a GEV distribution with

$Z = \frac{1}{D}$, $z^* = \sup\{z : F(z) < 1\} = \infty$

$F_Z \in MDA(G_\xi)$

$$\lim_{z \to \infty} \frac{1 - F(tz)}{1 - F(z)} = \lim_{z \to \infty} \frac{f(x_0)\,\pi^{\frac{p}{2}} (tz)^{-p} / \Gamma\left(\frac{p}{2}+1\right) + o(1)}{f(x_0)\,\pi^{\frac{p}{2}} (z)^{-p} / \Gamma\left(\frac{p}{2}+1\right) + o(1)}$$
$$= t^{-p}$$

$\Rightarrow -p = -\frac{1}{\xi}$, $\tilde{\xi} = \frac{1}{p} > 0$

A positive function C, on $[0, +\infty)$ is

① slowly varying at ∞, if $\lim_{x \to \infty} \frac{C(tx)}{C(x)} = 1$, for all $t > 0$.

② regularly varying at ∞ with index $\rho \in R$.

if $\lim_{x \to \infty} \frac{C(tx)}{C(x)} = t^\rho$, $t > 0$.

③ $\hat{\xi}_n = \frac{1}{k}\sum_{i=1}^{k} \log \frac{R_{(n+1-i)}}{u}$

$= -\frac{1}{k}\sum_{i=1}^{k} \log \left(\frac{-1/R_{(n+1-i)}}{-1/R_{(n-k)}} \right) \to \frac{1}{D}$, $\tilde{\gamma} = \frac{1}{p}$

(2006)

De Haan L. Extreme value theory : an introduction

⇓

Hill estimator

$\hat{\xi}_{Hill} \xrightarrow{P} \tilde{\gamma} = \frac{1}{p}$

k: the number of the excedances above u. 超多参数.

$H_0: \xi_n = -\frac{1}{p}$ $(\gamma^* = 0)$

$H_1: \xi_n = 0$ $(\gamma^* < 0)$, $x_0 \notin \text{supp}(f)$

Th5. Suppose $\gamma^* < 0$, $u = R_{(n-k)}$, $k = k(n) \to \infty$, $\frac{k(n)}{n} \to 0$, as $n \to \infty$

then $\hat{\xi}_n = \frac{1}{k}\sum_{i=1}^{k} \log \frac{R_{(n+1-i)}}{R_{(n-k)}} \to 0$.

proof [page10]. $0 \le -\hat{\xi}_n = \frac{1}{k}\sum_{i=1}^{k} \log \frac{-R_{(n-k)}}{-R_{(n+1-i)}}$

$\le k \cdot \frac{1}{k} \log \frac{-R_{(n-k)}}{-\gamma^*} \to 0.$

$R_{(n+1-i)} < \gamma^*$

$-R_{(n+1-i)} > -\gamma^*$

$\Rightarrow$ the first test $\begin{cases} H_0: p\hat{\xi}_n = -1 \\ H_1: p\hat{\xi}_n = 0 \end{cases}$

choose $s > -1$, if $p\hat{\xi}_n < s$, mark x_0 as possibly known, go to next step.

$s = -0.8$, $p\hat{\xi}_n < s$, 靠近 -1.

else if $p\hat{\xi}_n > s$, unknown.

(更靠近 0)

why $p\hat{\xi}_n$ instead of $\hat{\xi}_n$?

$-\frac{1}{p} \to 0$, as $p \to \infty$

$p\hat{\xi}_n \sim N(-1, 1)$

What if P_{ξ_n} is close to -1? (can't reject H_0)

From Th3, let $u = R_{(n-k)}$

$$P\left(\frac{-D}{R} > x\right) \approx \frac{k}{n}\left[1 - \hat{F}(x - R_{(n-k)})\right]$$

$$\Leftrightarrow \quad D < -x \quad = \frac{k}{n}\left(x/R_{(n-k)}\right)^{-\frac{1}{\xi_n}}$$

proof:

$$P\{-D > x,\ x > R_{(n-k)}\} = P(-D > x \mid x > R_{(n-k)})\, P(x > R_{(n-k)})$$

$$\approx \frac{k}{n}\left[1 - P(-D \le x \mid x > R_{(n-k)})\right]$$

$$= \frac{k}{n}\left[1 - \hat{H}(x - R_{(n-k)})\right]$$

$$= \frac{k}{n}\left(1 + \xi_n \frac{x - R_{(n-k)}}{\sigma}\right)^{-\frac{1}{\xi_n}}$$

$$\text{let } \sigma = u\xi,\ u = R_{(n-k)}$$

$$= \frac{k}{n}\left(\frac{x}{R_{(n-k)}}\right)^{-\frac{1}{\xi_n}}$$

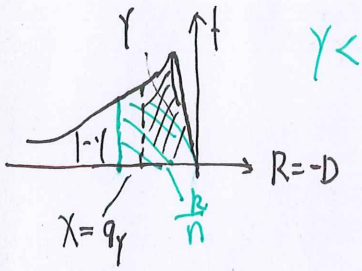
$$\gamma \approx \frac{k}{n}\left[1 - \hat{F}(x - R_{(n-k)})\right]$$

$$1 - \frac{n\gamma}{k} = \hat{H}(x - R_{(n-k)})$$

⊗ $q_\gamma = R_{(n-k)} + \hat{H}^{-1}(1 - \frac{n\gamma}{k})$

$$= R_{(n-k)}\left(\frac{n\gamma}{k}\right)^{-\xi}$$

$q_\gamma < 0$ 为 $(1-\gamma)$-quantile of $-D$



$\gamma < \frac{k}{n}$

(γ 固定，距离小，概率小)

$$f = H(y) = 1 - (1 + \xi\frac{y}{\sigma})^{-\frac{1}{\xi}}$$

$$y = \hat{H}^{-1}(f) = \frac{\sigma}{\xi}\left[(1-f)^{-\xi} - 1\right] = u\left[(1-f)^{-\xi} - 1\right]$$

$-q_\gamma \downarrow$, $q_\gamma \uparrow$, q_γ 靠近0, $f(x_0) \uparrow$

$-q_\gamma \uparrow$, $q_\gamma \downarrow$, q_γ 远离0, unknown

$$P\{B_{-q_\gamma}(x_0)\} = 1 - \gamma \ ?$$

So, if $-q_\gamma > t > 0$, mark x_0 as unknown.

otherwise, known.

$-q_\gamma$ 越小，$f(x_0)$ 处密度越大.

(尾部!)

5.3. GPDC Algorithm.

~~1. $-D_1, \cdots, -D_n$~~

input: ~~$\{$~~ $\{X_i\}_{i=1}^n$, X_0.

1. $-D_1, \cdots, -D_n$

2. $\hat{\xi}_n = -\frac{1}{k}\sum_{i=1}^{k}\log\left(\frac{R_{(n+1-i)}}{R_{(n-k)}}\right)$

3. fir test: $\begin{cases} H_0 : P\{\hat{\xi}_n = -1 \\ H_1 : P\{\hat{\xi}_n = 0. \end{cases}$

~~if $P\{\hat{\xi}_n < s$, where $s \geq -1$,~~

if $P\{\hat{\xi}_n > s$, where $s > -1$, X_0 is unknown.

else go next

4. $q_Y = R_{(n-k)} + \hat{H}^{-1}(1-\frac{1}{k})$

($1-\frac{1}{n}$)-quantile of $-D$)

5. if $-q_Y > t$, where $t > 0$, X_0 is unknown,

else X_0 is known.

① k: the number of upper order statistics

bias-variance trade-off

② s.t. : control type-I error.

a jackknife fashion (instead of using test statistic under H_0)

$\hat{\xi}_n^{(1)}, \cdots, \hat{\xi}_n^{(n)}$, $1-\frac{\alpha}{2}$ quantiles $\longrightarrow s$

$-q_Y^{(1)}, \cdots, -q_Y^{(n)}$, $1-\frac{\alpha}{2}$ quantiles $\longrightarrow t$.

(备) 两重检验, Bonferroni's correction.

~~设 $H_1, \cdots, H_m$~~

Let $H_1, \cdots, H_m$ be a family of hypotheses,

the familywise error rate (FWER) 族错误率.

$$FWER = P\{\bigcup_{i=1}^{m_0}(P_i \leq \frac{\alpha}{m})\}$$

出现至少一次错误 $\leq \sum_{i=1}^{m_0} P\{P_i \leq \frac{\alpha}{m}\}$

错误的概率

$$\leq m_0 \frac{\alpha}{m}$$

$$\leq m \frac{\alpha}{m} = \alpha.$$

P_i: 每一假设的 P 值.

m_0: 实际为真的零假设总数.

6. GEV classifier

① for each X_i, calculate

$$D_i^{min} = \min_{j \neq i} \|X_i - X_j\|$$

$O(n \log n)$

$$-D_i^{min} = \max_{j \neq i} \|X_i - X_j\| \leq 0.$$

② fit $\hat{W}$ to $-D_1^{min}, \cdots, -D_n^{min}$

③ $-d_0^{min} = \max_{i=1,\cdots,n} -\|X_i - X_0\|$,

H_0: X_0 is known

H_A: X_0 is unknown.

④ $\hat{W}(-d_0^{min}) < \alpha$: reject H_0. $\Rightarrow$ X_0 is unknown.

$P(-D^{min} < -d_0^{min})$.

K-D 树.

对于给定的维度, 查找时间 $O(\log N)$