Extreme Value Theory for Open Set Classification
GPD and GEV Classifiers

Edoardo Vignotto, Sebastian Engelke
{edoardo.vignotto, sebastian.engelke}@unige.ch

UNIVERSITÉ DE GENÈVE

Introduction

Classification tasks usually assume that all possible classes are present during the training phase. This is restrictive if the algorithm is used over a long time and possibly encounters samples from unknown classes. To overcome this problem, we propose two new algorithms that, relying on approximations from extreme value theory, are able to identify test data coming from unknown classes. We show the effectiveness of our classifiers in simulations and on the LETTER and MNIST datasets.

General Setting

- Training data: $x_i \in \mathbb{R}^p$, $i = 1, \dots, n$, with class labels $y_i \in \{C_1, \dots, C_J\}$.
- Each class is described by a density function f_{C_j} defined on $\mathbb{R}^p$.
- Training distribution: $f(x) = \sum_{j=1}^{J} w_j f_{C_j}(x)$, for weights $w_j \in [0, 1]$, with $\sum_{j=1}^{J} w_j = 1$.
- Test data: a new point x_0 that we want to mark as known, i.e., $x_0 \sim f$, or unknown, i.e., $x_0 \sim f_0$, for a new density f_0.
- Type I error: we want to control the probability α of making a type I error, i.e., the probability of marking as unknown a known point.

Theorems for GPD Classifier

Theorem 1 *Denote the distances between x_0 and the points in the training set with $D_1, \dots, D_n \sim D$ and assume that x_0 is from a known class, that is, the upper end point of $-D$ is zero. Under mild conditions, the distribution of $R_i - u$, where $R_i = -D_i$, above a high threshold $u < 0$ (close to 0) can be approximated by a GPD distribution $\widehat{H}$ with log-likelihood*

$$\log L(R_1, \dots, R_n; \xi) \propto -k \log \xi - \frac{1}{\xi} \sum_{i=1}^{k} \log \left(\frac{R_{(n+1-i)}}{u} \right),$$

where $R_{(n)} \geq R_{(n-1)} \cdots \geq R_{(1)}$ are the order statistics of the R_i and k is the number of exceedances above u.
The maximum likelihood estimator of ξ is then

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

Theorem 2 *Assume the same conditions as in Theorem 1. Choose the threshold u to be the order statistic $R_{(n-k)}$, and assume that $k = k(n) \to \infty$ and $k(n)/n \to 0$, as $n \to \infty$. Then the statistic in (1) converges in probability*

$$\widehat{\xi}_n \xrightarrow{P} -1/p,$$

where p is the dimension of the predictor space.

Theorem 3 *Assume the same conditions as in Theorem 2, only that x_0 is from an unknown class and the upper endpoint of $-D$ is $r^* < 0$. Then the statistic in (1) converges almost surely to zero, that is, $\widehat{\xi}_n \xrightarrow{a.s.} 0$.*

GEV Classifier

1. For each $i = 1, \dots, n$, compute the distance D_i^{min} between the training point x_i and the nearest training point to it.
2. Fit a reversed Weibull distribution $\widehat{W}$ to $-D_1^{min}, \dots, -D_n^{min}$.
3. Compute the distance d_0^{min} between x_0 and the nearest training point.
4. If $\widehat{W}(-d_0^{min}) < \alpha$ mark x_0 as unknown, otherwise mark it as known.

References

Vignotto E. and Engelke S., Extreme value theory for open set classification - GPD and GEV classifiers, *arXiv prepint arXiv:1808.09902*, 2018.

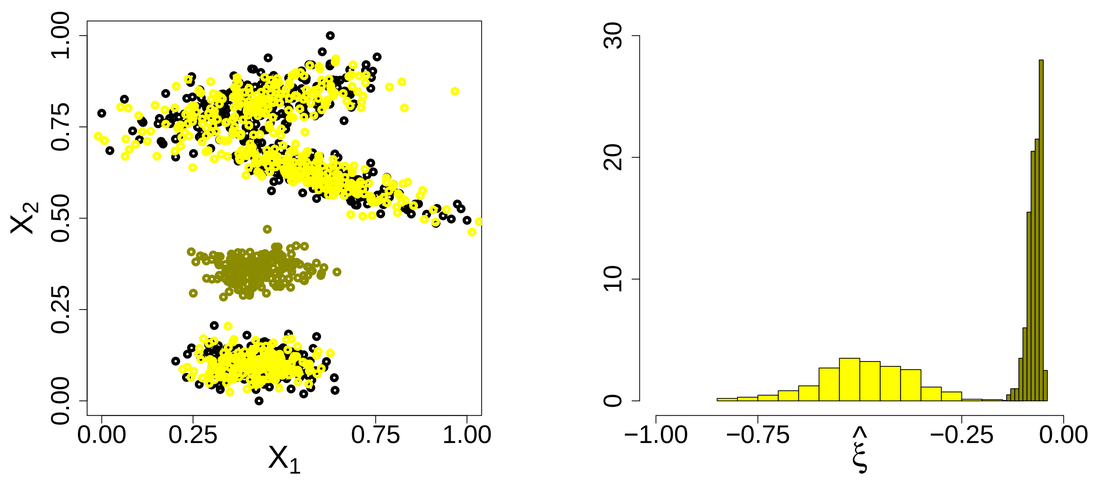
GPD Classifier

1. Compute the negated distances $-D_1, \dots, -D_n$ between x_0 and each point in the training set.
2. Estimate $\widehat{\xi}_n$ using only the biggest k negated distances $R_{(n)}, \dots, R_{(n+1-k)}$.
3. If $p\widehat{\xi}_n$ is smaller than a given threshold $s > -1$, mark x_0 as possibly known and go to the next point, otherwise mark it as unknown and exit the algorithm.
4. Compute a high quantile of $-D$ by $q_k = \widehat{H}^{-1}(1 - 1/k)$ of the estimated GPD $\widehat{H}$.
5. If the radius $-q_k$ is bigger than a given threshold $t > 0$ mark x_0 as unknown, otherwise mark it as known.

The threshold s and t are chosen executing the algorithm using all the points in the training set as unknown one after the other in a jackknife fashion and assuring an α probability of making a type I error.

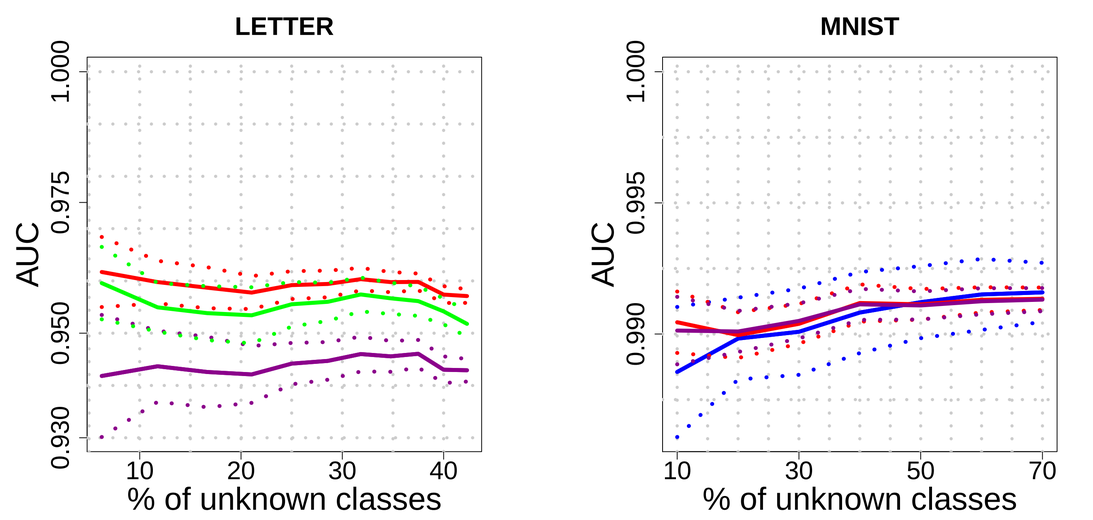
Results

Simulated data



On the left, the simulated dataset for the toy example: training data (in black), the known (bright yellow) and unknown examples (dark yellow) from the test set. On the right, the $\widehat{\xi}_n$ estimates for known (bright yellow) and unknown test data (dark yellow) for this dataset.

LETTER and MNIST datasets



Results obtained on the LETTER and MNIST datasets varying the amount of unknown test data for the GPD (red line) and the GEV (magenta line) classifiers and two competitors, the EVM (green line) and the One-Class SVM (blue line). Dotted lines represent one standard deviation confidence intervals. The evaluation measure is the Area Under the ROC curve (AUC).