On Anomaly Ranking and Excess-Mass Curves

Nicolas Goix Anne Sabourin Stéphan Clémençon

TELECOM ParisTech — INSTITUT Mines-Télécom

Motivation

No ROC curve in *unsupervised* learning !

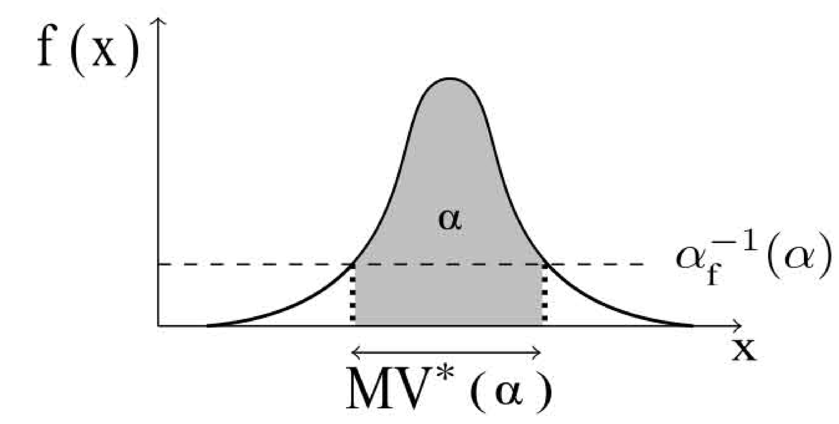
- Anomaly Detection: needs for ranking the observations according to their degree of abnormality. Scoring function:
  $$s: \mathcal{X} \to \mathbb{R}_+$$
  The "smaller" the score s(X), the more "abnormal" the observation X is viewed

  Linked with **density level sets estimation** ([1], [2])
- How to know if a score is good or not? How to compare two scoring functions ?
- How to build such scoring function ?
- **Need for a Criterion** to evaluate the quality and to be optimized in the process of building a scoring function.

Existing Work

- Mass-Volume Curve ([3]) $\alpha \in (0,1) \mapsto MV_s(\alpha) = \lambda_s \circ \alpha_s^{-1}(\alpha)$

  where $\alpha_s(t) = \mathbb{P}(s(X) \geq t)$ and $\lambda_s(t) = Leb(\{x \in \mathcal{X}, s(x) \geq t\})$
- Optimal MV curve:
  $MV^* := MV_f$



Drawbacks:

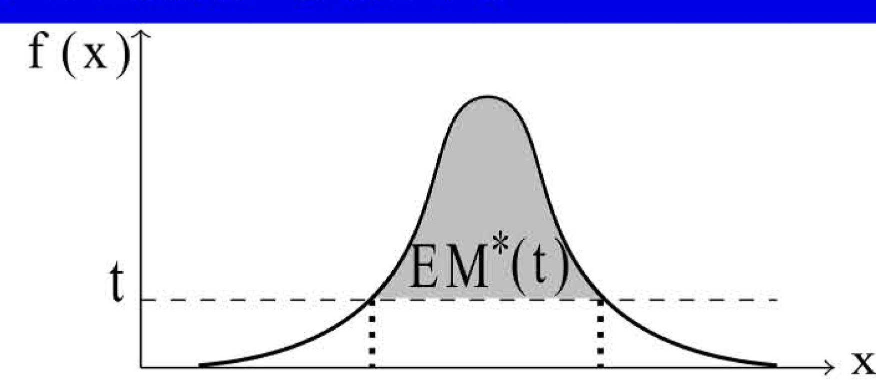
- As an evaluation criterion: pseudo-inverse may be hard to compute
- As a building criterion: produce level sets which are not necessarily nested.

A novel criterion: the Excess-Mass Curve

- EM curve of a scoring function $s: \mathcal{X} \to \mathbb{R}_+$:

$$EM_s(t) := \sup_{u \geq 0} \mathbb{P}(s(X) \geq u) - tLeb(s \geq u)$$

$$EM^* := EM_f(t) = \mathbb{P}(f(X) \geq t) - tLeb(f \geq t)$$



Excess-mass functional ([4], [5])

For all score s:
$EM^* \geq EM_s$

$\|EM^* - EM_s\|_\infty$

- $EM^*(t) - EM_s(t) \leq \|f\|_\infty \inf_{u>0} Leb(\{s > u\} \Delta \{f > t\})$
  ➡ pseudo distance between ***the level sets induced by s*** and ***those induced by the true underlying distribution f.***
- ➡ pseudo distance between s and f "in the space of scoring function"! (In this space, $s = T \circ s$ for every increasing transform T)
  $\|EM^* - EM_s\|_\infty \leq C \inf_{T \nearrow} \|f - T \circ s\|_\infty$

- **Practical criterion:** $$\widehat{EM}_s(t) = \sup_{u \geq 0} \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{s(X_i) \geq u} - tLeb(s \geq u)$$

➡ **Evaluation criterion**: s_1 better than s_2 if $\widehat{EM}_{s_1} \geq \widehat{EM}_{s_2}$ ⟶ Main drawback: Monte Carlo for $Leb(s \geq u)$

➡ **Criterion for M-estimation:** $\underset{s}{\text{maximize}}\ \widehat{EM}_s$ ⟶ $s \in ?$ maximize $\widehat{EM}_s(t)$ $\forall t$?

M-estimation

Formulation

- Optimal solution of $\underset{s}{\text{maximize}}\ EM_s$: any function s with $\Omega_t^* := \{f \geq t\}$ as level sets. For instance : $s(x) = \int_{t=0}^{+\infty} \mathbb{I}_{x \in \Omega_t^*} a(t)dt$ with a(t) > 0.

Fix $0 < t_K < t_{K-1} < \ldots < t_1$

$$\underset{\Omega_{t_1} \subset \ldots \subset \Omega_{t_N}}{\text{maximize}}\ \widehat{EM}_{s_N} \quad \text{where } s_N(x) = \sum_{k=1}^N (t_k - t_{k+1})\mathbb{I}_{x \in \Omega_{t_k}}$$

(then we have: $\Omega_{t_k} = \{s \geq t_k\}$)

⇕

$$\widehat{\Omega}_{t_k} = \text{argmax}_{\Omega \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{X_i \in \Omega} - t_k \text{Leb}(\Omega)\ ;\ s_N(x) = \sum_{k=1}^N (t_k - t_{k+1})\mathbb{I}_{x \in \widehat{\Omega}_{t_k}}$$

Algorithm : for k = 1, ... ,N
$$\begin{cases} \widehat{\Omega}_{t_k} = \underset{\Omega \supset \widehat{\Omega}_{t_{k-1}}}{\text{argmax}} \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{X_i \in \Omega} - t_k \text{Leb}(\Omega) \\ t_{k+1} = \frac{t_1}{(1+\frac{1}{\sqrt{n}})^k} \end{cases}$$

Theoretical results

Assumptions:

The density f has **no flat parts**, is **bounded**.
The class $\mathcal{G}$ has **VC-dim** $< \infty$

For bias control:

There exists a countable subcollection of $\mathcal{G}$, $F = \{F_i\}_{i \geq 1}$ say, forming a partition of $\mathcal{X}$ and such that $\sigma(F) \subset \mathcal{G}$.

Theorem: Let an integer $N > 0$ and s_N the scoring function returned by the algorithm. Then with proba larger than $1 - \delta$:

$$\sup_{t \in]0, t_1]} |EM^*(t) - EM_{s_N}(t)| \leq \left[A + \sqrt{2\log(1/\delta)}\right]\frac{1}{\sqrt{n}} + \|f - f_F\|_{L^1} + o_N(1),$$
$$\sup_{t \in]0, t_1]} |EM^*(t) - EM_{s_\infty}(t)| \leq \left[A + \sqrt{2\log(1/\delta)}\right]\frac{1}{\sqrt{n}} + \|f - f_F\|_{L^1}$$

where $o_N(1) = 1 - EM^*(t_N)$. The support of f may be non compact.

Bias: $\|f - f_F\|_{L^1}$

f_F the best approximation (for the L_1-norm) of f by piecewise functions on F,
$f_F(x) := \sum_{i \geq 1} \mathbb{I}_{x \in F_i} \frac{1}{Leb(F_i)} \int_{F_i} f(y)dy$.

Experiments

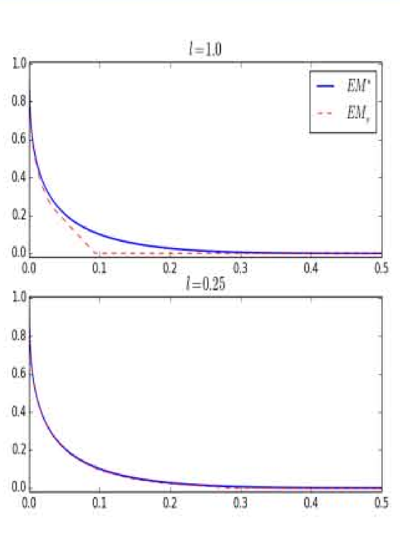


Figure 4: Optimal and realized EM curves

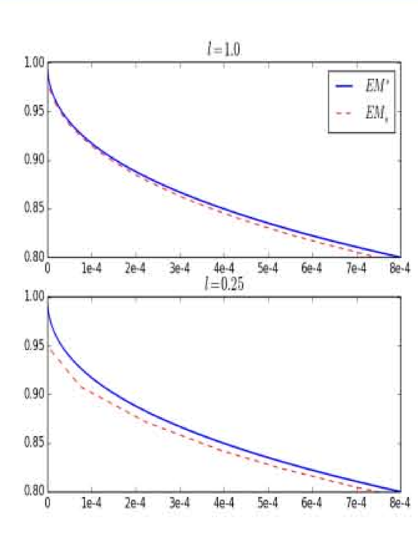


Figure 5: Zoom near 0

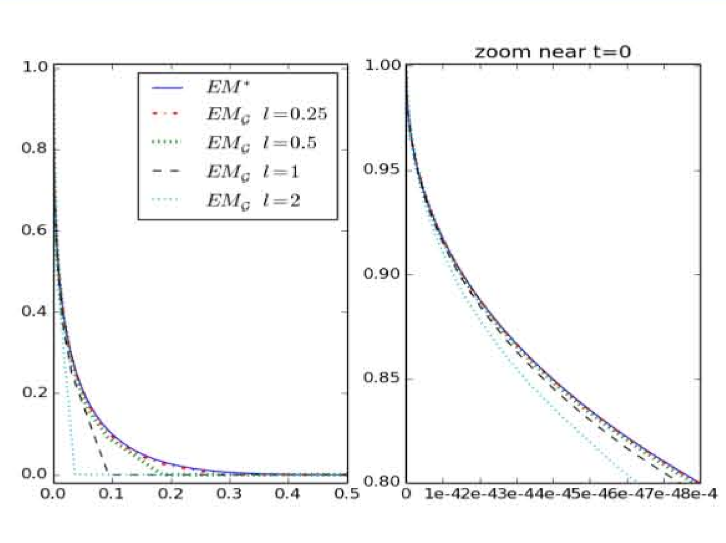


Figure 6: EM_G for different l

References

[1] C. Scott and R. Nowak. Learning Minimum Volume Sets. *Journal of Machine Learning Research*, 7:665–704, 2006.

[2] J.P. Vert and R. Vert. Consistency and convergence rates of one-class svms and related algorithms. *JMLR*, 6:828–835, 2006.

[3] S. Clémençon and J. Jakubowicz. Scoring anomalies: a M-estimation approach. 2013.

[4] W. Polonik. Measuring mass concentrations and estimating density contour cluster-an excess mass approach. *The annals of Statistics*, 23(3):855–881, 1995.

[5] D.W. Müller and G. Sawitzki. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, 86(415):738–746, 1991.