The hidden assumptions and pitfalls of density based anomaly detection

FARIED ABU ZAID, appliedAI

August 11, 2022

Anomaly detection is a notoriously ill-defined problem. The notion of an anomaly is arguably subjective since it depends to some extent on the downstream task. Nevertheless, there have been several attempts to provide exact descriptions that can act as definition of an anomaly. In particular, density based anomaly scores are very popular as they appeal to the intuition that anomalies appear in rarely observed regions of the feature space. Inspired by work by Charline Le Lan and Laurent Dinh, I want to discuss in this article why even perfect density models cannot guarantee to provide good anomaly detection results and why I think we still need them.

Contents

In [LD21] the authors challenge some commonly used density based approaches to anomaly detection. They criticize the representation dependence of density scoring methods. The paper is interesting not only because of its content but also because of its history. It received a quite strong rejection at ICLR 2020 although the reviewers honor the attempt to challenge current practices. The rejection was mainly because the reviewers did not agree with the main principle that is formulated in the paper: Anomaly detection techniques should generally be invariant under reparametrization with continuous invertible functions. The ICLR review is available online on OPENREVIEW and the discussion between the authors and the reviewers is an interesting addition to read alongside the paper.

The paper was eventually published in the journal ENTROPY and presented at the I Can't Believe it's not Better workshop at NEURIPS 2020. I actually first saw the well delivered presentation at this workshop and it really made me think again about the fundamental setup of anomaly detection. However, I would eventually agree with the ICLR review. In this post I want to present the content of the paper alongside some thoughts that occurred to me while reading in the hope that other people might also benefit from revisiting these fundamental questions about anomaly detection.

1 Density based definitions of anomaly

A popular way of approaching anomaly detection is to view it as a binary classification task where at training time only samples from one class, the nominal one, are available (one-class classification). Simply put, one aims to partition the feature space $\mathscr{X}$ into two subsets $\mathscr{X}_{\text{in}}$ and $\mathscr{X}_{\text{out}}$ where $\mathscr{X}_{\text{in}}$ denotes the nominal region and $\mathscr{X}_{\text{out}}$ the anomalous region. Within this approach, we assume that the nominal points are drawn from a probability distribution P_X and demand that $\mathcal{X}_{\text{in}}$ covers the majority of the density mass, say 99%. However, this alone does not uniquely determine the partition as there are obviously infinitely many ways to cover 99% of the probability mass of a continuous distribution.

One can additionally argue that the density of the nominal distribution in the anomalous region should generally be smaller than in the nominal region by appealing to the intuition that anomalies should produce unusual observations. We can then define $\mathcal{X}_{\text{in}} = \{x \in \mathcal{X} \mid p_X(x) > \lambda\}$ and $\mathcal{X}_{\text{out}} = \{x \in \mathcal{X} \mid p(x) \leq \lambda\}$ to be the upper level and lower level sets with respect to some density threshold λ that is chosen such that $P_X(\mathcal{X}_{\text{in}}) = 0.99$. This definition appears for instance in Bishop [Bis94].

Intuitively, the idea might seem pretty solid since densities are related to probabilistic frequencies which seems to match our intuition that anomalies occur in unlikely areas of the feature space. However, a direct attribution of high density with high frequency seems faulty in high dimensions when considering the Gaussian annulus theorem[1].

Most of the probability mass of a high dimensional Gaussian is concentrated in a thin annulus around the surface of a hypersphere of radius $\sqrt{d}$. This might seem unintuitive at first glance but becomes clearer if we recall that the length of a sample vector is distributed like $\sqrt{\sum^{d}_{i=1} X_i^2}$ and where the X_i are independent standard normal distributions. Since $\sum_{i=1}^{d} X_i^2$ follows a chi-squared distribution by definition, $\sqrt{\sum^{d}_{i=1} X_i^2}$ follows a chi distribution with d degrees of freedom. The mean of $\sqrt{\sum^{d}_{i=1} X_i^2}$ is therefore $\sqrt{2}\,\frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \approx \sqrt{k}$. The chi distribution has a relatively low variance of $k - \left(\sqrt{2}\,\frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}\right)^2$.

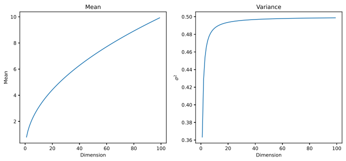

Figure 1. Mean and variance of a χ distribution as function of the degrees of freedom

Hence, for large d we will barely ever observe anything close to the origin. It was argued therefore that one might want to count it to the anomalous region. The highest density of the Gaussian is however still obtained at the origin. In order to account for such phenomena, the notion of a typical set has been introduced.

Definition 1. *(ϵ-typical set [CT06]). For a random variable X and $\epsilon > 0$ the ϵ-typical set $A_{\epsilon}^{(N)}(X) \subseteq \mathcal{X}^N$is the set of all sequences that satisfy*

$$\left| H(X) + \frac{1}{N}\sum_{i=1}^{N} \log(p(x_i)) \right| \leq \epsilon,$$

where $H(X) = -E[\log(p(X))]$ is the (differential) entropy.

[1] Gaussian Annulus Theorem [BHK20]. For every spherical d-dimensional Gaussian with variance 1 in each direction and any $\beta < \sqrt{d}$ at most $3\,e^{-c\beta^2}$ of the probability mass lies outside the annulus $\sqrt{d} - \beta \leq |x| \leq \sqrt{d} + \beta$ where c is a fixed constant.

The definition of typical sets is useful for dealing with phenomena like the Gaussian annulus theorem since $\lim_{N\to\infty} P(A_\epsilon^N(X)) = 1$ for any $\epsilon > 0$. Hence, for large N the ϵ-typical set will contain most of the mass with respect to the joint probability measure.

1.1 Is the center of a high dimensional Gaussian anomalous?

Let us first note that the Gaussian annulus theorem does not state that the area that is enclosed by the annulus is disproportionally rarely observed. Rather, it follows from the geometry of the high dimensional space that the volume close to the surface of the sphere is relatively large.[2] For large d, the ϵ-annulus of radius $\sqrt{d}$ contains many times the volume of the enclosed area. However, the probability of the enclosed area under an isotropic Gaussian is still larger than the probability of any subset of the annulus of the same volume. In a sufficiently (i.e. very) large dataset the area around the origin will in fact have the highest density of data points. The relationship between the d-th order growth of the volume of the sphere with the radius and the exponential decay of the density function of the Gaussian with the radius creates annulus phenomenon.

Nevertheless, points close to the origin are in some sense very dissimilar to the vast majority of the observed points in terms of distance to the origin. I think this is a very good example where a subjective notion of rareness is tied to a non-Euclidean notion of similarity. In fact, if we consider the variable $Y = |X|$ then we obtain density based anomalous regions that are in line with the intuition that only the annulus should be counted as nominal. One should be aware that behind this mapping is an arbitrary notion of equivalence, or more generally of similarity, in terms of length of a vector. In this space densities look substantially different. We also loose information because we equate all vectors of same length. We already see that the statement that anomalous regions coincide with low densities can only be true under certain representations because it depends on how well distances in the space capture our intuitive notion of similarity.

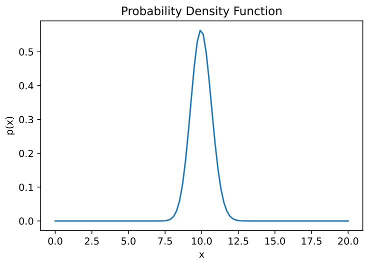

Figure 2. Probability density function of a χ distribution with 100 degrees of freedom.

2 The role of parameterization

The main point of the paper is to stress that densities are highly dependent on the feature space, up to a point where one can arbitrarily interchange high and low densities via reparametrization. This fact remains true even if we restrict ourselves to continuous bijective maps where the image of the map contains the same information about the represented event as the input. This leads the authors to raise doubts about whether density scoring based approaches are reasonable for anomaly detection in general.

Let us first have a look at the above claim. It is actually a simple consequence of how densities transfer via continuous bijective maps. Let

[2] Recall that the volume of a d-dimensional sphere with radius r is $V_d(r) = \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)} r^d$. Hence for $r > \epsilon$, the ratio of the volume of an ϵ-annulus of a sphere to the enclosed volume is $\frac{V_d(r+\epsilon) - V_d(r-\epsilon)}{V_d(r-\epsilon)} = \frac{(r+\epsilon)^d}{(r-\epsilon)^d} - 1$, which goes to ∞ as $d \to \infty$.

X be a continuous random variable, $f\colon \mathscr{X} \to \mathscr{X}'$ a continuous invertible function on X, and $Y = f(X)$. Then the pdf of X, $p_X(x)$, transfers via f to the pdf on Y, $p_Y(y)$. However, we need to take into account the way f locally stretches or compresses the space. This is reflected in the change of variables formula.

How severely even continuous invertible transformations can alter a density function is demonstrated by the Knothe-Rosenblatt rearrangement. There are only two mild assumptions that have to be made. The densities of the two distributions should be larger 0 everywhere and all cumulative conditional densities $P_{X_i|X_{<i}}(X_i \leq x_i|\, x_1, \ldots, x_{i-1})$ should always be differentiable in $(x_1,\ldots,x_i)$. As a consequence one can transform any two such continuous distributions into each other using a continuous invertible map.[3]

Example 2. Let X and Y be two continuous random variables with the above-mentioned properties. With the Knothe-Rosenblatt construction we obtain two continuous bijective maps f_X, f_Y such that $f_X(X)$ and $f_Y(Y)$ are uniformly distributed over $[0,1]$. We claim that $f_Y^{-1} f_X(X)$ has the same distribution as Y. Indeed, letting $h := f_Y^{-1} \circ f_X$:

$$\begin{aligned} p_{h(X)}(h(x)) &= p_X(x)|h'(x)|^{-1} \\ &= (p_X(x)|f_X{}'(x)|^{-1})\,|(f_Y^{-1})'(f_X(x))|^{-1} \\ &= (p_X(x)|f_X'(x)|^{-1})\,|f_Y{}'(h(x))| \\ &= p_Y(h(x)). \end{aligned}$$

Note that the inverse of a continuous function is not necessarily continuous if the domain and range have different dimensions but with the two additional assumptions mentioned above, one can show that f_X and f_Y are differentiable bijections. The inverses are therefore also diffeomorphisms and in particular continuous. The examples show that a purely density based approach to anomaly detection can lead to completely different results depending on the nature and transformations of observed features. To see this even more clearly, it is important to observe how the densities of a distribution can change relative to each other under continuous invertible maps. In the paper the authors construct several more fine-grained examples that show how one can alter densities even locally in an almost arbitrary fashion or explicitly interchange the densities of two given points. Even if we accept a density based definition in the original (real-world) probability space, the data we collect is already a transformation thereof, i.e. a random variable. The authors illustrate this on the example of images where the depicted object is the true sample and the images are the transformations we observe. This becomes even more severe if the images are given in a compressed format. Therefore, a density based approach to anomaly detection must necessarily rely on the assump-

[3] **Knothe-Rosenblatt rearrangement** [Kno57, Ros52]: Any continuous distribution with the above mentioned properties can be transformed into a uniform distribution using a continuous invertible map. Let $X_1, \ldots, X_n$ be continuous random variables with joint distribution $P_{X_1,\ldots,X_n}$. Consider the map $f(x_1, \ldots, x_n) = (y_1, \ldots, y_n)$ with $y_i = P_{X_i|X_{<i}}(X_n \leq x_i|\, x_1, \ldots, x_{i-1})$. Note that $\frac{\partial f}{\partial x^T}$ is lower triangular since y_i does not depend on x_j for $i < j$. Further, the ith component on the main diagonal of $\frac{\partial f}{\partial x^T}(x_1, \ldots, x_n)$ is $p_{X_i|X_{<i}}(x_i|\, x_1, \ldots, x_{i-1})$ since it is the derivative of the corresponding conditional cumulative distribution. Hence, the determinant of $\frac{\partial f}{\partial x^T}(x_1, \ldots, x_n)$ is simply the product of the conditional densities, which equals $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ because of the product rule. With this observation we can show that $(Y_1, \ldots, Y_n) = f\,(X_1, \ldots, X_n)$ is uniformly distributed over the n-dimensional unit hypercube because for all $y_1, \ldots\, y_n \in [0,1]$: $p_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = p_{X_1,\ldots,X_n}(f^{-1}(\bar{y}))\left|\frac{\partial f}{\partial x^T}(f^{-1}(\bar{y}))\right|^{-1} = 1.$

tion that low density areas actually correspond to anomalous regions in the presented feature space. Note that this has nothing to do with how well our density model captures the true distribution. In fact, the same problems arise if the true density function of the nominal data under a certain representation is available.

The authors stress a very important point which is certainly not new but often overlooked in practice: One should keep in mind that the bijectivity of the feature selection function is already a strong assumption that will be violated in many practical scenarios. Network intrusion detection, for instance, is often performed on just a few connection statistics, which are in no way sufficient for uniquely characterising every possible connection [TBLG09]. Further, we will usually have only limited knowledge about the transformation that led to the observed data. The direct attribution of anomalous regions with low densities becomes therefore arbitrary and needs additional justification.

3 Reparametrization invariant approaches

We can obtain a reparametrization invariant definition of the anomalous region if we model the anomaly detection problem fully probabilistically and use Bayesian inference. This is known as Huber's contamination model [Hub64]. Since we accept that even in the anomalous region the density of the nominal distribution is not 0, a sample that lies in $\mathscr{X}_{\text{out}}$ is not necessary an anomaly. Therefore, one might rather think of being anomalous as a binary random variable O. The probability that a given sample is an anomaly is given by $P_{O|X}(o|\,x)$. In this case, it seems more plausible to choose the anomalous regions based on a threshold on $P_{O|X}(o|\,x)$. This leads to a mixture model $(1-\epsilon)\,D_{\text{in}} + \epsilon D_{\text{out}}$ between the nominal distribution D_{in} and the distribution D_{out} of the anomalies. Here $D_{\text{in}} = P(X|\,O=0)$, $D_{\text{out}} = P(X|\,O=1)$ and $\epsilon = P(O=1)$ is the prior for observing an anomaly. We can now define $\mathscr{X}_{\text{in}} = \{x \in \mathscr{X} \mid P_{O|X}(1|\,x) \leq \lambda\}$ and $\mathscr{X}_{\text{out}} = \{x \in \mathscr{X} \mid P_{O|X}(1|\,x) > \lambda\}$ for some threshold $\lambda \in [0,1]$. This definition is invariant under reparametrization. Indeed, for any continuous invertible transformation of X we can compute with Bayes' rule that

$$
\begin{aligned}
P_{O|f(X)}(o|\,f(x)) &= \frac{p_{f(X)|O}(f(x)|\,o)\,P_O(o)}{p_{f(X)}(f(x))}\\
&= \frac{p_{X|O}(x|\,o)\left|\frac{\partial f}{\partial x^T}\right|^{-1} P_O(o)}{p_X(x)\left|\frac{\partial f}{\partial x^T}\right|^{-1}}\\
&= \frac{p_{X|O}(x|\,o)\,P_O(o)}{p_X(x)}\\
&= P_{O|X}(o|\,x).
\end{aligned}
$$

While this analysis is quite pleasing from a theoretical point of view, it is arguably of limited use in practice. Often times, the fraction of anomalies ϵ is simply too small to learn a reasonable mixture model. It might even happen that we have no anomalous samples at all. Therefore, we need to make assumptions about ϵ and $p_{X|O}(x|1)$. One usually adds additional assumptions to justify scoring by the nominal density [KS12]. More precisely, we assume that that anomalies are:

- Outlying: D_{in} and D_{out} do not overlap too much.
- Sparse: $\epsilon \ll \frac{1}{2}$.
- Diffuse: D_{out} is not too spatially concentrated.

It might seem a little disappointing that we end up with the same density based method. However, these assumptions explicitly state under which conditions we can expect good results from such a method. It explicitly incorporates spatial assumptions about the nature of anomalies, especially the first and the last assumption take explicit reference to spatial aspects of anomalies. This is even more present in other approaches where for instance anomalies are defined in terms of distances, e.g. nearest neighbor distance ratios [BKNS00]. A distance based approach emphasizes even more that the properties we are looking for are not invariant under reparametrization.

The paper also mentions another way out: The comparison against a reference distribution D_{ref}. One can take the ratio of the density with respect to the nominal distribution and the reference distribution [GT07]. If both - nominal and reference distribution - are transformed under the same continuous invertible function then the effect of the transformation will be canceled out in their ratio. The reference distribution allows us to model knowledge about the feature space. In this language we can quantify our assumptions and explicitly integrate them into our calculations. The drawback of this approach is again that we cannot expect to have complete knowledge about the transformation that the data has undergone. It can be very hard to define a good reference distribution under these circumstances.

4 Reparametrization invariance as a principle?

The previous analysis led the authors to the conclusion that any anomaly detection technique should be invariant under reparametrization. They formulate this as a principle.[4]

The proposal of this principle has been heavily criticized by the reviewers and led eventually to the rejection at ICLR. I also feel that this requirement is too strong. In this section, I want to point out a few things that become apparent if we switch from the perfect model regime to the learning from data regime.

[4] Formulation in the paper: "In an infinite data and capacity setting, the result of an anomaly detection method should be invariant to any continuous invertible reparametrization f."

4.1 Why the principle is too strong

We now consider the scenario where anomalies are indeed defined as lower level sets for some density threshold λ with respect to the distribution of the nominal data in some fixed base representation. We want to show that no algorithm can learn the anomalous region if we allow that the data to be transformed by an arbitrary continuous invertible function. Intuitively, this follows from the fact that we can arbitrarily transform any distribution. However, this true for any algorithm that purely learns from nominal data, even in the infinite data regime. To make this a little more precise, let us formulate some properties that we can safely assume for a learning algorithm $\mathcal{A}$:

1. The algorithm takes a dataset $X \sim D$ and outputs an anomalous region $\mathcal{A}(X)$ (represented by a model).
2. As the size of the dataset grows towards infinity, the algorithm converges to a solution $\mathcal{A}(D)$ which only depends on the distribution D of the data.
3. The limit solutions are measurable and bounded with respect to some base measure μ on the feature space, e.g. the Lebesgue measure if $\mathscr{X} = \mathbb{R}^d$.

We make the assumptions 2 and 3 mostly out of convenience. Note that in this setup there is an implicit assumption about the algorithm being deterministic. If we consider a stochastic algorithm we have to fix the “random seed” of the algorithm and changing it would lead to a different algorithm. What is important is that the result only depends on the presented data. The algorithm is free to incorporate assumptions about the data but these assumptions need to tied to the algorithm and must be independent of the distribution from which the algorithm is actually drawn. These assumptions are quite common when one wants to talk about general limitations of learning algorithms. In fact, they are inspired by the extended Bayesian formalism, which was also used by Wolpert in his no free lunch theorem [Wol02]. While the use of an improper uniform prior in the no free lunch theorem can certainly be debated, the formal framework in which he conducts his proof is well suited to answer general questions like ours.

I want to argue that if we restrict ourselves to reparametrization invariant approaches then we cannot - not even in the infinite data regime - guarantee to capture the anomalous region closely in terms of precision with respect to the base measure in the feature space (we use the base measure because we have no knowledge about the distribution of the anomalies within the anomalous region or the frequency of anomalies at test time). This holds even in cases were substantial knowledge about the base representation is available. Let us illustrate this with a little story about an unfortunate scientist. The example shows that different situations become indistinguishable when we transform the distributions with the Knothe-Rosenblatt rearrangement.

An unfortunate measurement

Imagine some scientists want to measure the state of an obscure particle that they have just discovered. They know that the state of the particles are uniformly distributed in $[0,1]$ except for one interval $\left[\frac{i}{n}, \frac{i+1}{n}\right)$ where the density must be 50% smaller when compared to the other intervals. They do not know where the exact spot is located, so they decide to conduct an experiment and build a density model.

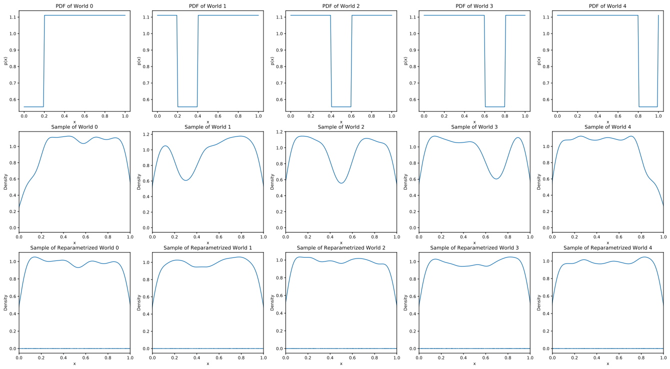

Figure 3. Visualization of the experiment. We sample from all possible worlds and fit a kernel density estimator in the original feature space (something the scientist can of course not do). After that we apply the cdf to the sample and fit again a kde. In the original feature space the situations are clearly distinguishable but after reparametrization the situations are indistinguishable from the data.

They can draw from an infinite supply of particles but unfortunately they cannot measure the state directly. All they can do is to compare the state of two particles and see which one has the higher value. They can compare one particle with as many as they want, and therefore they decide to take one particle at a time and compare it with many newly drawn ones. Finally, they record the fraction of particles that had a smaller state. Can the data help our scientists to find the anomalous region? Certainly not!

What they are actually recording is the value of the cumulative density function. If X is the state of a particle our scientist records $f(X) = \mathrm{CDF}_X(X)$. As we have previously seen, $f(X)$ will be uniformly distributed over $[0,1]$ irrespective of where the actual anomalous region is located.

Let us formalize the situation a little further:

- The scientists know that $Y = \mathrm{CDF}_{X_i}(X_i)$ for some $i < n$ and have some prior P_i on i.

- Given the dataset of i.i.d. samples from Y they build a posterior belief about i.

But since the likelihood of D does not depend on i there is nothing they can learn about i from observing D:

$$\begin{aligned} P_i(i|\,D) &= \frac{p_{D|i}(D|\,i)\,P_i(i)}{p_D(D)} \\ &= \frac{p_D(D)\,P_i(i)}{p_D(D)} \\ &= P_i(i). \end{aligned}$$

Therefore, no matter what they will try to do with the data it will not help them to identify the anomalous region. But that should make us doubt whether we should demand that the result of any anomaly detection method is invariant under reparametrization. Any algorithm that tries to learn only from D must learn the same anomalous region $\mathcal{X}_{\text{out}}$ in the infinite data regime, regardless of the value of i.[5]

One can take this argument to the extreme and derive the following normal form for reparametrization invariant learning algorithms.

PROPOSITION 3. *For every learning algorithm that is invariant under reparametrization and every $n \in \mathbb{N}^+$ there is some set $O \subseteq [0,1]^n$ such that for any continuous probability distribution D over $\mathbb{R}^n$ (which fulfills 2) the algorithm outputs $f_D^{-1}(O)$ in the infinite data regime when presented with data independently drawn from D. The function f_D denotes here the Knothe-Rosenblatt construction w.r.t. D.*

That means the algorithm can essentially be specified by some set $O \subseteq [0,1]$ which determines the outcome for almost any possible input. Importantly, this set O is chosen before the data is seen. In the case of $\mathcal{X} = \mathbb{R}$ the algorithm has a preselected set of percentiles that are anomalous and he just "estimates" the correct values for them from the data. Hence, the result has almost nothing to do with the problem at hand!

As the previous example shows, this unavoidably leads to failures of the algorithm even for "trivial" instances. Note that the same type of argument can be applied to many other notions of anomalies (including those that don't solely depend on the distribution of nominal data).

5 Conclusion

Given the success of non-reparametrization invariant methods in anomaly detection, the principle seems unreasonably restrictive. However, I agree that we should have a reference framework for the test scenario where the problem definition is reparametrization invariant. Such a framework could be the aforementioned mixture model of nominal and anomaly distribution. In practice, we mostly have to live with the fact that only nominal data is available for training. Hence, I would rather stress that most notions of anomalies are tied to a distance metric (e.g. implicitly when estimating the density).

[5] We can try to bound the average fraction of $f_i^{-1}(\chi_{\text{out}})$ that intersects with the anomalous region $\left[\frac{i}{n}, \frac{i+1}{n}\right)$. Indeed, we can compute for the average case that: $\frac{1}{n}\sum_{i=0}^{n-1} \frac{\mu\left(f_i^{-1}(\chi_{\text{out}}) \cap \left[\frac{i}{n}, \frac{i+1}{n}\right)\right)}{\mu(f_i^{-1}(\chi_{\text{out}}))} \leq \frac{1}{n}\sum_{i=0}^{n-1} \frac{c\mu\left(f_i\left(\left[\frac{i}{n}, \frac{i+1}{n}\right)\right) \cap \chi_{\text{out}}\right)}{\frac{c}{2}\mu(\chi_{\text{out}})} \leq \frac{2}{n}\frac{\sum_{i=0}^{n-1} \mu\left(f_i\left(\left[\frac{i}{n}, \frac{i+1}{n}\right)\right) \cap \chi_{\text{out}}\right)}{\sum_{i=0}^{n-1} \mu\left(f_i\left(\left[\frac{i}{n}, \frac{i+1}{n}\right)\right) \cap \chi_{\text{out}}\right)} = \frac{2}{n}$. The first inequality holds since $\left|\frac{\partial f_i}{\partial x}(x)\right| = \begin{cases} \frac{c}{2} \text{if } x \in \left[\frac{i}{n}, \frac{i+1}{n}\right) \\ c \, \text{else} \end{cases}$ for some $c > 0$ and therefore $\frac{c}{2}\mu(f(X)) \leq \mu(X) \leq c\mu(f(X))$ for all measurable subsets X of $[0,1]$. The second inequality holds since we have $f_i\left(\left[\frac{i}{n}, \frac{i+1}{n}\right)\right) \cap f_j\left(\left[\frac{j}{n}, \frac{j+1}{n}\right)\right) = \emptyset$ for $i \neq j$ (checking this is a good exercise). Therefore, any algorithm must produce a largely wrong result with $\frac{\mu\left(f_i^{-1}(\chi_{\text{out}}) \cap \left[\frac{i}{n}, \frac{i+1}{n}\right)\right)}{\mu(f_i^{-1}(\chi_{\text{out}}))} \leq \frac{2}{n}$ in at least one situation. Since this is unavoidable, we ask that the algorithm produce the same bad result even if we present the real state X_i instead of $\text{CDF}_{X_i}(X_i)$.

I think the more interesting question is whether we can learn representations that are particularly well suited for anomaly detection. It is known that deep density models such as normalizing flows do not necessarily map out-of-distribution data into low density areas of the latent space [KIW20, ZGR21]. I think this will continue to be a major research direction in anomaly detection for the upcoming years.

Nevertheless, I believe the paper to be a valuable contribution. The authors remind us that some common approaches to anomaly detection should be used with more care. They rest on assumptions that are not explicitly formulated and lack theoretical justification. Their article definitely motivated me to revise these foundational questions with greater care.

Let us wrap up this article with a few takeaway messages:

- Anomaly detection is notoriously ill-defined and arguably subjective to a certain degree. When applying an anomaly detection method, one needs to ensure that the selected approach actually captures the notion of anomaly in the application. The main goal of this article is to convince you that this is not intrinsic to the algorithms.

- Density based approaches are particularly fragile because the result can be almost arbitrarily changed by "simple" continuous invertible transformations (as they are routinely applied to data).

- Additionally, even in the original - usually high-dimensional - probability space low densities might not coincide with anomalous regions because rareness might be tied to a non-Euclidean notion of similarity. This indicates that anomaly detection approaches might have the hidden assumption that Euclidean distance is a suitable measure of similarity.

- Mixture models or likelihood ratio scores against a reference distribution allow to encode missing information about potential anomalies to interpret the densities consistently across all possible reparametrizations.

- However, mixture models / reference distributions need to be defined in a representation specific fashion which is not always possible.

- I agree with the ICLR reviewers that even with an infinite supply of data and capacity no algorithm can guarantee to eventually learn the anomalous region from nominal data if arbitrary reparametrizations are allowed. Therefore, one should not try to enforce reparametrization invariance as a general principle.

- We conclude that feature engineering has an even more crucial role in anomaly detection than in other areas of machine learning. It is the practitioners' responsibility make sure that the representation of the data and the selected approach are suitable for each other.

- Learning good representations for anomaly detection remains a major challenge in machine learning research.

BIBLIOGRAPHY

[BHK20] Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. Cambridge University Press, Cambridge, 2020.

[Bis94] C. M. Bishop. Novelty detection and neural network validation. *IEE Proceedings - Vision, Image and Signal Processing*, 141(4):217–222, aug 1994. Publisher: IET Digital Library.

[BKNS00] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: identifying density-based local outliers. *ACM SIGMOD Record*, 29(2):93–104, may 2000.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, 2006.

[GT07] Thomas L. Griffiths and Joshua B. Tenenbaum. From mere coincidences to meaningful discoveries. *Cognition*, 103(2):180–226, may 2007.

[Hub64] Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. Publisher: Institute of Mathematical Statistics.

[KIW20] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Why normalizing flows fail to detect out-of-distribution data: 34th Conference on Neural Information Processing Systems, NeurIPS 2020. *Advances in Neural Information Processing Systems*, 2020-December, 2020.

[Kno57] Herbert Knothe. Contributions to the theory of convex bodies. *Michigan Mathematical Journal*, 4(1):39–52, jan 1957. Publisher: University of Michigan, Department of Mathematics.

[KS12] JooSeuk Kim and Clayton D. Scott. Robust kernel density estimation. *The Journal of Machine Learning Research*, 13(1):2529–2565, sep 2012.

[LD21] Charline Le Lan and Laurent Dinh. Perfect density models cannot guarantee anomaly detection. *Entropy*, 23(12):1690, dec 2021. ArXiv: 2012.03808.

[Ros52] Murray Rosenblatt. Remarks on a Multivariate Transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, 1952. Publisher: Institute of Mathematical Statistics.

[TBLG09] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. A detailed analysis of the KDD CUP 99 data set. In *Proceedings of the Second IEEE international conference on Computational intelligence for security and defense applications*, CISDA'09, pages 53–58. Ottawa, Ontario, Canada, jul 2009. IEEE Press.

[Wol02] David H. Wolpert. The Supervised Learning No-Free-Lunch Theorems. In Rajkumar Roy, Mario Köppen, Seppo Ovaska, Takeshi Furuhashi, and Frank Hoffmann, editors, *Soft Computing and Industry: Recent Applications*, pages 25–42. Springer, London, 2002.

[ZGR21] Lily Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding Failures in Out-of-Distribution Detection with Deep Generative Models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12427–12436. PMLR, jul 2021. ISSN: 2640-3498.