Real-world-robustness of tree-based classifiers

Christoph Schweimer
cschweimer@know-center.at
Know-Center GmbH
Graz, Austria

Sebastian Scher
sscher@know-center.at
Know-Center GmbH
Graz, Austria

ABSTRACT

The concept of trustworthy AI has gained widespread attention lately. One of the aspects relevant to trustworthy AI is robustness of ML models. In this study, we show how to exactly compute the recently introduced measure of real-world-robustness - a measure for robustness against naturally occurring distortions of input data - for tree-based classifiers under the assumption that the natural distortions are given as probability distributions. The idea is to extract the decision rules of a trained tree-based classifier, separate the feature space into non-overlapping regions and determine the probability that a data sample with distortion returns its predicted label. The original method works for all black box classifiers, but is only an approximation and only works if the input dimension is not too high, whereas our proposed method returns exact results.

1 INTRODUCTION

Robustness of machine learning models is a recently widely investigated topic. One extensively studied topic is adversarial robustness [26], which deals with small particular manipulations of the input to cause misclassifications. These systematic manipulations are called adversarial attacks, and several algorithms have been developed [5, 19, 25] to find the nearest counterfactual [16, 18, 29] (closest point to the input according to a distance metric that leads to misclassification) of data samples in various scenarios. Especially the area of adversarial attacks on images is highly researched since it can cause a variety of safety concerns, e.g., in medical image processing and classification [15, 17]. [30] argue that adversarial attacks are unnatural and not meaningful since the adversarial data samples are very unlikely to occur in real-world applications. They therefore developed a method to generate natural adversarial examples [14] with generative adversarial networks [12]. Adversarial examples are found in the latent space and mapped back into the original feature space. The distance from the input data sample to the generated adversarial data samples is measured in the latent space, not in the original feature space. [20] go a step further and argue that data samples that are generated with adversarial attacks shall not be called natural adversarial examples. In their interpretation, natural adversarial examples occur in the real-world and lead to a misclassification “without an evident cause”, i.e., are caused by natural noise (e.g., from cameras or by natural changes of the input). [28] investigated the robustness of classifiers against blurs in images. [13] create benchmarks for the robustness of image classifiers with several robustness metrics on corrupted images (ImageNet-C). These alterations to the images are being referred to as common corruptions (e.g., noise and blurring in images), opposed to adversarial attacks and can be interpreted as natural adversarial examples. The robustness of trained classifiers is measured by evaluating the performance on unseen test data and the computation of the corruption error and variations thereof. In adversarial examples, robustness is usually measured by some distance metric, e.g., the Euclidean distance, and the distance to the closest counterfactual is used as the robustness metric. With this interpretation of robustness, one single adversarial example in the close vicinity of a test sample can have a huge influence in the evaluation of classifiers (see [23] for a visual illustration).

[7] investigate robustness of smoothed classifiers with Gaussian noise, which come from a base classifier. A Monte-Carlo sampling approach is used to determine the most probable outcome of a data sample under the Gaussian noise. Even though the probability for each class can be computed, it is not used as the measure of robustness.

A different definition of robustness was introduced in [23] and termed real-world-robustness. It describes a general framework to compute the robustness of individual predictions of a trained machine learning model against natural distortions, e.g., data-processing errors, noise or measurement errors in the input data, opposed to individual adversarial samples. The real-world-robustness $R_\mu$ of the prediction $f$ of an $N$-dimensional data sample $\mu$ with distortion $\mu(x)$ given as a probability density function (PDF) is defined via a binary function $f'$,

$$f'(\mu, \epsilon) = \begin{cases} 0, & f(\mu + \epsilon) = f(\mu) \\ 1, & f(\mu + \epsilon) \neq f(\mu) \end{cases}$$

with $\epsilon \sim p_\mu(x)$ and an integral that determines the probability $P$ for a different prediction compared to the data sample $\mu$,

$$P(f(\mu + \epsilon) \neq f(\mu)) = \int_{\mathbb{R}^N} f'(\mu, \epsilon) \, d\epsilon.$$  \hspace{1cm} (2)

The real-world-robustness $R_\mu$ of the data sample $\mu$ with distortion $\epsilon \sim p_\mu(x)$ is then computed by

$$R_\mu = 1 - P.$$  

In words, real-world-robustness is the probability that the prediction (classification) of an input sample does not change under the given uncertainty of the input sample. [23] showed how real-world-robustness can approximately be computed for any black-box classifier with a Monte-Carlo based method, under the constraint that the input feature space is not too high dimensional. They additionally provide a detailed discussion for the justification of this definition of real-world-robustness, and how it differs from adversarial robustness. This definition of real-world-robustness clearly distinguishes the measure from corruption-robustness, which measures the rate of misclassification in test sets with included errors, but does not give a conditional probability of misclassification for single test samples. The practical relevance of real-world-robustness are settings in which the uncertainty of input samples is known (or at least approximately known), and in which one needs a measure
of how likely it is that a misclassification will occur due to the random error in the input data. This could for example be applications that use multivariate sensor data with measurement noise (e.g., temperature measurements at different locations).

Another research direction deals with robust adversarial training of machine learning models. [21] present a comprehensive survey on robust adversarial training by introducing the fundamentals and a general theoretical framework, and by summarising different training methodologies against various attack scenarios. [27] introduce a training framework by adding an adversarial sample detection network to improve the classifier. In tree-based models, a training framework to learn robust trees against adversarial attacks has been developed by [4], and [11] investigate the robustness of Decision Trees with symmetric label noise in the training data. [5] propose a robustness verification algorithm for tree-based models to find the minimal distortion in the input that leads to a misclassification.

In this paper, we show how to precisely compute real-world-robustness for tree-based classifiers (Decision Trees, Random Forests and XGBoosted classifiers), under the assumption that the uncertainty of the input test samples can be described by certain statistical distributions. This is possible because the decision boundaries of tree-based classifiers are explicitly given (in contrast to, e.g., neural network classifiers). We extract the decision rule of each decision node of a tree-based classifier to separate the input feature space into non-overlapping regions. We then determine the probability that a random data sample wrt. the given uncertainty, which is modelled as a probability distribution, around a test sample lies in a region that has the same label as the prediction of the test sample itself. Our measure of robustness is a probability which differs from distance metrics that are commonly used for adversarial robustness.

The paper is structured as follows. First, we describe how to compute real-world-robustness for single Decision Trees in Section 2. Then the approach is extended to Random Forest classifiers and XGBoosted classifiers in Section 3. In Section 4, we present experimental results and Section 5 concludes the paper.

2 ROBUSTNESS OF DECISION TREES
At first we show how to compute real-world-robustness for a trained Decision Tree (DT) classifier [22] with a categorical target variable. We extract the decision rule from each decision node of a trained DT to split the input feature space into non-overlapping regions. For two-dimensional inputs, the regions are rectangles and for higher dimensions, the regions are hyperrectangles. For ease of description, we call the regions boxes [5]. To determine the robustness of the prediction of a data sample with uncertainty (e.g., noise or measurement errors), we classify the data sample with the trained DT and compute the probability that a random sample wrt. the given uncertainty is in a box that has the same label as the data sample. Taking the sum over the computed probabilities returns the robustness of the DT classification for that particular data sample.

2.1 Segmentation of the feature space
We have a trained DT without prior knowledge about the input features $X_i$. Each decision node in a DT is a decision rule of the form $X_i \leq \tau_{ij}$, where $\tau_{ij}$ marks the $j^{th}$ decision rule of feature $X_i$. Note that $\tau_{ij}$ is unique for each $j$ in a DT. We extract the decision rule from each decision node in the tree, add it to the decision rule set $\tau_1$ of the associated feature $X_i$ and sort each $\tau_i$ in ascending order. The elements of $\tau_i$ split one dimension of the input feature space into non-overlapping segments and the individual elements of two sets $\tau_j$ and $\tau_k$ are orthogonal to each other. Using two successive elements of each decision rule set $\tau_i$ to split the input feature space creates one individual box. If a feature $X_i$ is bounded, we expand $\tau_i$ to its minimum and/or maximum values, otherwise we expand $\tau_i$ to negative and positive infinity to cover the entire input feature space, i.e., if $\tau_i = \{60, 80, 100\}$, the expanded unbounded set is $\tau_i' = \{-\infty, 60, 80, 100, \infty\}$. This results in a total number of boxes $n_b$ given by

$$n_b = \prod_i^N(|\tau_i^j| - 1) = \prod_i^N(|\tau_i| + 1), \quad (3)$$

where $N$ is the number of input features and $|\tau_i|$ is the number of decision rules for feature $X_i$.

2.2 Robustness
We use the created boxes to determine the robustness $R_{\mu}$ of the prediction of an input data sample $\mu$ with associated uncertainty $p_{\mu}(\tilde{x})$. We determine the predicted label of $\mu$ as well as the labels of all boxes by classifying their centre with the DT. To compute the robustness of the prediction of $\mu$, it suffices to only consider the boxes that have the same label as $\mu$ itself, denoted as $B_{\mu}$, since data samples in these boxes return the same result as $\mu$. We determine the probability mass $p_B$ that each box $B \in B_{\mu}$ is covering wrt. the given uncertainty $p_{\mu}$ around the data sample $\mu$. Taking the sum

![Figure 1: Illustration of boxes of a trained binary Decision Tree with two input features $X_1, X_2$ and a data sample $\mu$.](image)

The elements of $\tau_i$ split one dimension of the input feature space into non-overlapping segments and the individual elements of two sets $\tau_j$ and $\tau_k$ are orthogonal to each other. Using two successive elements of each decision rule set $\tau_i$ to split the input feature space creates one individual box. If a feature $X_i$ is bounded, we expand $\tau_i$ to its minimum and/or maximum values, otherwise we expand $\tau_i$ to negative and positive infinity to cover the entire input feature space, i.e., if $\tau_i = \{60, 80, 100\}$, the expanded unbounded set is $\tau_i' = \{-\infty, 60, 80, 100, \infty\}$. This results in a total number of boxes $n_b$ given by

$$n_b = \prod_i^N(|\tau_i^j| - 1) = \prod_i^N(|\tau_i| + 1), \quad (3)$$

where $N$ is the number of input features and $|\tau_i|$ is the number of decision rules for feature $X_i$.

2.2 Robustness
We use the created boxes to determine the robustness $R_{\mu}$ of the prediction of an input data sample $\mu$ with associated uncertainty $p_{\mu}(\tilde{x})$. We determine the predicted label of $\mu$ as well as the labels of all boxes by classifying their centre with the DT. To compute the robustness of the prediction of $\mu$, it suffices to only consider the boxes that have the same label as $\mu$ itself, denoted as $B_{\mu}$, since data samples in these boxes return the same result as $\mu$. We determine the probability mass $p_B$ that each box $B \in B_{\mu}$ is covering wrt. the given uncertainty $p_{\mu}$ around the data sample $\mu$. Taking the sum
over the determined probability masses returns the robustness of the prediction of \( \mu \). Figure 1 illustrates the classified boxes (two labels) of a trained DT with two input features \( X_1, X_2 \) and a data sample \( \mu \).

In case the uncertainty distribution of the data sample \( \mu \) is analytically tractable (e.g., a multivariate normal distribution with uncertainty \( \Sigma \)), exact solutions can be computed. We determine the probability mass \( m_B \) that each box \( B \in B_\mu \) is covering by integrating the probability density function \( p_\mu(\mathbf{x}) \) of the underlying uncertainty distribution between the lower and upper boundaries of each box. Taking the sum over all computed probability masses gives the robustness \( R_\mu \) of the prediction of the data sample \( \mu \).

\[
R_\mu = \sum_{B \in B_\mu} \int_{B_{\text{low}}}^{B_{\text{upp}}} p_\mu(\mathbf{x}) \, d\mathbf{x},
\]

where \( B_{\text{low}} \) denotes the lower boundaries and \( B_{\text{upp}} \) denotes the upper boundaries of a box.

2.2.1 One-dimensional feature space. For one-dimensional data samples \( \mu \) with uncertainty given as a probability distribution, we can just integrate the PDF of the uncertainty probability distribution between the lower and upper boundary of each box \( B \in B_\mu \) if it is analytically tractable and take their sum to get the robustness of the prediction of \( \mu \).

2.2.2 Input data with multivariate normal uncertainty. For an \( N \)-dimensional data sample \( \mu \) with multivariate normal uncertainty \( \Sigma \), we integrate the multivariate normal probability density function

\[
p_\mu(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N|\Sigma|}} \exp \left( -\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right)
\]

with the method by [10] between the lower and upper boundaries of each box \( B \in B_\mu \) and take the sum over the probability masses, which returns the robustness of the prediction of \( \mu \).

2.2.3 Input data with mixed uncertainty distributions. For an \( N \)-dimensional data sample \( \mu \) with correlated features and the uncertainty in different dimensions given by different distributions, computing the robustness via analytical integration is not always possible, since integrating the joint PDF might not be analytically tractable - except if the variables are all uncorrelated to each other. However, in certain cases, analytical integration of correlated non-Gaussian distributions is possible using the reversed NORTA (normal to anything) principle [3], where a joint probability distribution with given correlation matrix is transformed into a multivariate normal distribution in two steps. In the first step, we apply the cumulative distribution function (CDF) of the respective distributions that model the uncertainty w.r.t. the data sample \( \mu \) on the decision boundaries in each dimension, which transforms them into the \([0, 1]^N\)-space. In the second step, we use the inversion method [9] by applying the inverse of the CDF of the standard normal distribution on each transformed decision boundary in the \([0, 1]^N\)-space. We can now solve Equation 4 by integrating the multivariate normal probability density function (Equation 5) with \( \mu = 0 \) and the covariance matrix \( \Sigma \) with Ones on the main diagonal elements and the correlation coefficients between the features on the off-diagonal elements between the transformed lower and upper boundaries of each box \( B \in B_\mu \) to compute the robustness of the prediction of a data sample \( \mu \).

For an \( N \)-dimensional data sample with independent features, the robustness can be computed directly via analytical integration, if the PDF of the uncertainty distribution in each dimension is analytically tractable. We determine the covered probability mass of each box \( B \in B_\mu \) by taking the product of the covered probability masses in each dimension and then take the sum over the probability masses \( m_B \) per box.

2.2.4 Runtime Improvement. Computing the robustness of low-dimensional inputs and shallow trees is fast since the number of boxes is small. For high-dimensional inputs and deep trees, the number of boxes (see Equation 3) and simultaneously the runtime increases with each input dimension. One approach to decrease the runtime for the robustness computation of the prediction of a data sample \( \mu \) with multivariate normal uncertainty \( \Sigma \) is to only consider the boxes that are inside of or intersect with the bounding box of, e.g., the 99% confidence hyperellipsoid around a data sample \( \mu \). Experiments have shown that the resulting robustness computations are much faster since the probability mass for less boxes needs to be computed, while the difference in the results is negligible. There is also a theoretical upper bound for the error of the resulting robustness (at max 1 percentage point). Another approach to speed up computations, which has not been tested, would be to compute the probability mass \( m_B \) of each box with the same label as \( \mu \) in parallel.

![Figure 2: Boxes inside of or intersecting with the bounding box of the 99% confidence ellipse around a data sample with multivariate normal uncertainty, and more transparent boxes outside of the bounding box.](image-url)
Figure 2 represents a two-dimensional data sample \( \mu \) with multivariate normal uncertainty, where the boxes that are outside of the 99% confidence ellipse are more transparent. Only using the boxes that are inside of or intersect with the bounding box of the 99% confidence ellipse to determine the robustness of the prediction of \( \mu \) speeds up computations, while the difference in the results is negligible (see Section 4).

3 ROBUSTNESS OF OTHER TREE-BASED METHODS

We now extend the approach to determine the robustness of a DT to more advanced tree-based methods. We look at Random Forests, which consist of multiple trees, and at XGBoosted trees.

3.1 Random Forest

A Random Forest (RF) [2] extends a Decision Tree model and consists of multiple DTs. Each tree is trained on a bootstrapped dataset with the same size as the original training dataset, i.e., a sample can be part of the training set for one tree multiple times, whereas other samples are not part of that training set. We extract the decision rule of each decision node in all individual trees, merge them per input feature \( X_i \) to form the decision rule set \( \tau_i \) and create boxes to compute the robustness of the prediction of a data sample.

Analogous to the DT setup, we have a trained RF without prior knowledge about the input features \( X_i \) or the individual trees. For each tree in the RF, we extract the decision rule \( X_i \leq \tau_{ij} \) of each decision node, add it to the decision rule set \( \tau_i \) of the associated feature \( X_i \) and sort each \( \tau_i \) in ascending order, as was done for DTs. Since a RF consists of multiple DTs, decision nodes in different trees can have the same decision rule \( X_i \leq \tau_{ij} \) leading to duplicate entries in \( \tau_i \). We eliminate all but one of the duplicate entries \( \tau_{ij} \) such that each \( \tau_{ij} \) is unique. To create boxes for robustness computations, we use the method described in Section 2.1 for DTs.

After creating the boxes, we compute the robustness of the prediction of a data sample \( \mu \) in a RF, analogous to the approach for DTs described in Section 2.2. First we determine the label of each box by classifying its centre with the RF, not with the single trees. Then we classify \( \mu \) with the RF to determine its label and take the sum over all probability masses \( m_B \) of the boxes \( B \in B_\mu \) (boxes with the same label as \( \mu \)).

Computing the robustness in a RF is computationally more expensive than in a DT, since a RF consists of multiple trees, leading to more decision rules and a higher number of boxes (see Equation 3). In Section 2.2.4, we described an approach to approximate the robustness of the prediction of a data sample with multivariate normal uncertainty in a DT by only considering the boxes that are inside of or intersect with the bounding box of the 99% confidence hyperellipsoid. Experiments with various sizes of RFs (number of trees and depth of trees) have shown that computing the robustness of the prediction of a data sample is much faster when only considering these boxes, while the robustness results only differ marginally (see Section 4).

Figure 3 contains the boxes of a trained RF (Figure 3d) with two input features \( X_1, X_2 \), two labels and three associated DTs (Figures 3a, 3b, 3c), as well as a data sample with multivariate normal uncertainty. We see that each individual DT is represented by different decision boundaries and different boxes. Overlaying the individual Figures of the DTs gives the representation for the trained RF (Figure 3d). We also observe that the data sample is classified differently in DT1 (red label) compared to the other two DTs and the RF (yellow label).

At first glance, it would appear simpler to compute the real-world-robustness of a RF by first computing it individually for each DT of the forest, and then combining the results. This is, however, not possible, as averaging the robustness of the individual trees cannot account for interdependencies (e.g., even if tree A and tree B have the same robustness, the contributions from different parts of the feature space might be different, and averaging them would lead to false results).

3.2 XGBoosted Decision Tree model

Similar to a Random Forest, an XGBoosted Decision Tree model [6] also consists of multiple trees. Instead of training the trees from bootstrapped datasets, the individual trees build on each other. We again extract the decision rules of each decision node in all trees and combine them per feature \( X_i \) to form the associated decision rule set \( \tau_i \). We create boxes which are used to compute the robustness of the prediction of a data sample. Since trees in an XGBoosted Decision Tree model build on each other, there are less different decision rules than in a RF, when the number of trees and depth of the trees is the same. This leads to a smaller number of boxes and thus computations are faster.

4 EXPERIMENTAL RESULTS

In this Section we summarise experimental results for the computation of real-world-robustness for tree-based classifiers. Experiments were carried out on 1 core of an Intel(R) Xeon(R) 6248 CPU @ 2.50GHz processor with 256GB RAM. We are using 2 datasets for the experiments, the Iris flower dataset [1] and the MNIST dataset [8]. For illustration purposes, the uncertainty distribution in most experiments is prescribed as a multivariate normal distribution since we can compute exact solutions directly. We have not made comparisons to other notions of robustness since most other robustness techniques (e.g., adversarial examples) use a distance metric between a data sample and counterfactuals, whereas our method returns a probability. The repository with codes will be made available with the camera ready version.

4.1 Results with 99% confidence hyperellipsoid

In Section 2.2.4 we described a method to decrease the runtime of the robustness computation while only having negligible differences in the results. We trained a Decision Tree with a maximum depth of 4 on the Iris flower dataset with a train/test split of 90/10, yielding 15 test samples. To compute the robustness of the prediction of the test samples, the uncertainty distribution of the test samples is given as a multivariate normal distribution with 0.1 on the main diagonal and Zeros on the off-diagonals of the covariance matrix. This uncertainty distribution was chosen for illustration purposes. In real applications, the uncertainty needs to be prescribed based on the exact application setting, requiring domain knowledge. In the experiments, we determine the robustness of the prediction of the test samples computed with all boxes...
Figure 3: Boxes of 3 Decision Trees with two-dimensional input (a) - (c) and the resulting combined boxes (d) of the associated binary Random Forest.

and the robustness, where we only look at the boxes that are inside of or intersect with the bounding box of the 99% confidence hyperellipsoid around the test samples. Table 1 lists the robustness of the 15 test samples computed with all boxes and the boxes that are inside of or intersect with the bounding box of the 99% confidence hyperellipsoid. Even though the number of boxes diverges heavily for some test samples, we see that the robustness results are very similar and only start to differ in the fifth decimal place. Figure 4 shows the robustness results of the 15 test samples in comparison. We observe that the data points are almost on the straight dashed red line with slope 1, which illustrates the similarity between the results. We achieved $R^2$-scores exceeding 0.9999 in each test run (equivalent to the interpretation of the $R^2$-score in a linear regression model), showing the similarity between the results and that
Table 1: Comparison of robustness results for 15 test samples of the Iris dataset using all boxes and only the boxes that are inside of or intersect with the bounding box of the 99% confidence hyperellipsoid around a test sample for robustness computation

| Test sample | All boxes | 99% boxes |
|-------------|-----------|-----------|
|             | Boxes - Robustness | Boxes - Robustness |
| 1           | 8 - 0.99640        | 2 - 0.99640    |
| 2           | 8 - 0.99999        | 1 - 0.99998    |
| 3           | 8 - 0.99955        | 2 - 0.99954    |
| 4           | 13 - 0.93596       | 13 - 0.93596   |
| 5           | 13 - 0.85704       | 13 - 0.85704   |
| 6           | 8 - 0.99113        | 6 - 0.99114    |
| 7           | 8 - 0.99986        | 4 - 0.99985    |
| 8           | 8 - 0.99955        | 2 - 0.99954    |
| 9           | 13 - 0.94685       | 13 - 0.94685   |
| 10          | 11 - 0.97147       | 11 - 0.97147   |
| 11          | 13 - 0.98848       | 5 - 0.98847    |
| 12          | 13 - 0.91796       | 13 - 0.91796   |
| 13          | 8 - 0.95900        | 2 - 0.95900    |
| 14          | 13 - 0.76319       | 13 - 0.76319   |
| 15          | 8 - 0.99115        | 4 - 0.99114    |

Figure 4: Illustration of robustness results for 15 test samples of the Iris dataset using all boxes and only the boxes that are inside of or intersect with the bounding box of the 99% confidence hyperellipsoid around a test sample. The dashed red line is straight.

it suffices to only consider the boxes that are inside of or intersect with the bounding box of the 99% confidence hyperellipsoid to compute the robustness of the prediction of the test samples.

4.2 Runtime Analysis

Considering only the boxes that are inside of or intersect with the bounding box of the 99% confidence hyperellipsoid around a test sample not only returns accurate results for the robustness of the prediction of a test sample, but is also much faster. With the Iris flower dataset, the runtime difference for the test samples was negligible since it only contains 4 input features. We therefore conducted experiments with the MNIST dataset. We resized the images from $28 \times 28$ to $5 \times 5$ pixels, normalised them, flattened them into a vector and used the 25-dimensional vector as input to train a Random Forest with 5 trees and a maximum depth of 3 per tree on the training set (60,000 training samples). To evaluate the runtime of the robustness computation, we determined the robustness of the first 10 images of the test set (10,000 test samples). The uncertainty distribution of the test samples is given as a multivariate normal distribution with 0.001 on the main diagonal elements and Zeros on the off-diagonal elements of the covariance matrix for illustration purposes. The resulting runtimes and number of boxes for the robustness computation of the 10 test samples are listed in Table 2. We see that the number of boxes that are inside of or intersect with the bounding box of the 99% confidence hyperellipsoid is much smaller and that the runtimes are much shorter, compared to using all boxes. The difference in the results is negligible as we again achieved $R^2$-scores exceeding 0.9999, further indicating that it suffices to only use the boxes that are inside of or intersect with the bounding box of the 99% confidence hyperellipsoid to compute the robustness of the prediction of test samples.

4.3 Comparison to approximate robustness computation

[23] introduced the term real-world-robustness, but their method is based on random sampling and therefore only returns approximate results. With this comes the limitation that it only works when the feature dimensionality is not too high. We compare the two methods (Random Sampling (sampling 1,000,000 times) against 99% confidence hyperellipsoid around a test sample) by computing the robustness of the prediction of data samples on trained Decision Trees. The DTs have a depth between 4 and 6 and were trained on the MNIST dataset. We perform the same data preprocessing steps as for the runtime analysis, but resize the images to various sizes, ranging from $3 \times 3$ to $10 \times 10$ such that the DTs have input feature dimensions ranging from 9 to 100, and train them on the training set. The uncertainty is given as a multivariate normal distribution with 0.0001 on the main diagonal elements and Zeros on the off-diagonal

Table 2: Comparison of runtimes for 10 test samples of the MNIST dataset using all boxes and only the boxes that are inside of or intersect with the bounding box of the 99% confidence hyperellipsoid around a test sample for robustness computation

| Test sample | All boxes | 99% boxes |
|-------------|-----------|-----------|
|             | Boxes - Runtime | Boxes - Runtime |
| 1           | 19,118 - 488s | 296 - 8s   |
| 2           | 24,616 - 589s | 3,848 - 99s|
| 3           | 32,564 - 502s | 1,824 - 46s|
| 4           | 39,636 - 697s | 52 - 1s    |
| 5           | 13,356 - 274s | 1,512 - 39s|
| 6           | 32,564 - 518s | 1,824 - 45s|
| 7           | 13,356 - 241s | 48 - 2s    |
| 8           | 48,128 - 927s | 380 - 10s  |
| 9           | 32,564 - 468s | 120 - 3s   |
| 10          | 5,736 - 153s  | 24 - 1s    |
Table 3: Comparison of results for 9 trained Decision Trees on the MNIST dataset using the boxes that are inside of or intersect with the bounding box of the 99% confidence hyperellipsoid around a test sample and using the method in [23] for robustness computation

| Resized Image Size | DT depth | $R^2$-score |
|--------------------|----------|-------------|
| 3 x 3              | 4        | 0.99990     |
| 3 x 3              | 5        | 0.99999     |
| 3 x 3              | 6        | 0.99999     |
| 5 x 5              | 5        | 0.99999     |
| 5 x 5              | 6        | 0.99998     |
| 8 x 8              | 4        | 0.99999     |
| 8 x 8              | 5        | 0.99999     |
| 10 x 10            | 4        | 0.99999     |
| 10 x 10            | 5        | 0.99998     |

![Robustness results with XGBoosted Decision Tree model](image)

Figure 5: Illustration of robustness results for 15 test samples of the Iris dataset for correlated features with mixed distortions.

4.4 Robustness computation for correlated features and mixed distortions

In case the features are correlated and the uncertainty distribution of data samples is best modelled by different probability distributions in different dimensions, we can compute the robustness of the prediction of a data sample $\mu$ by utilising the reversed NORTA principle presented in Section 2.2. To illustrate the approach, we trained an XGBoosted Decision Tree model with a maximum depth of 4 and 6 trees on the Iris flower dataset with a train/test split of 90/10. For illustration purposes the uncertainty of the 4 features is given by different distributions - normal for feature 1 (sepal length), exponential for feature 2 (sepal width), $\chi^2$ for feature 3 (petal length) and lognormal for feature 4 (petal width) - with different parameters. After transforming the distributions, the correlation matrix is given with Ones on the main diagonal and values between 0.1 and 0.3 on the off-diagonals. The robustness results of the 15 test samples are visualised in Figure 5 and we observe that several test samples have a robustness of exactly 1. This is caused by uncertainty distributions with a lower bound (e.g., exponential or $\chi^2$ distribution).

In case the value of a test sample in a dimension that is modelled by such a probability distribution is higher than the highest value for the decision boundaries in that dimension, the data point in this dimension will always be above the highest value of the decision boundary. For a visual explanation, if the data sample $\mu$ in Figure 1 was in the upper right box (or a neighbouring box thereof) and the uncertainty in both dimensions was best modelled by two exponential distributions, the data sample would have a robustness of 1 since no smaller values could be sampled.

5 CONCLUSION

In this paper, we presented a method to precisely determine the real-world-robustness (robustness against natural distortions in the input) of tree-based classifiers. We extract the decision rules from a trained classifier to separate the input feature space into non-overlapping regions, called boxes, and integrate the underlying probability distribution that models the uncertainty of a test sample between the lower and upper boundaries of each box to determine their covered probability mass. Taking the probability sum over the boxes that have the same label as the test sample returns the real-world-robustness of the prediction of the test sample. We presented this approach for Decision Trees in detail and discussed the extension to Random Forests and XGBoosted trees. The method gives a precise measure of the real-world-robustness of the prediction of individual data samples, introduced by [23], with tree-based classifiers. This differs from other robustness measures which use various techniques to look for adversarial examples and use the distance between the test sample and the closest counterfactual as the measure for robustness. It also differs from robustness to common corruptions, which measures the average rate of misclassification of a test set with perturbed samples.

One limitation of our approach is that the uncertainty distribution needs to be modelled as a probability distribution and that the PDF of the distribution must be analytically tractable to compute exact solutions. In case the uncertainty distribution is more complex (e.g., because it is not given as a probability distribution, but by a stochastic function that models some process), we can in principle...
use numerical integration techniques to integrate the probability mass over the boxes for classifiers with explicit decision boundaries (such as tree-based models) and find approximate solutions. We carried out some initial experiments in that direction, solving Equation 4 with the quadpy package [24] for numerical integration. We observed however that the results with numerical integration techniques are often unstable and unreliable, especially when a data sample is in a large box, i.e., there is a big gap between the lower and upper boundaries in at least one dimension of the box, compared to cases that are actually analytically solvable. While it could be possible that with more carefully choosing types and parameters of the numerical routines the results could be better, this shows that simply using off-the-shelf integration routines is not a feasible option.

A second limitation of our approach is the high amount of data storage space needed. Equation 3 shows the number of boxes that are being created for a trained classifier. Classifiers with high input feature dimension and especially Random Forests quickly exceed the available storage space, such that the robustness cannot be computed anymore. With a more powerful machine, experiments on tree-based classifiers with high input feature dimension could be carried out and still return exact solutions.

The method presented in this paper is applicable because tree-based classifiers have the convenient property of having explicitly described decision boundaries that form hyperrectangles. Future research should be dedicated to extending the presented approach to more advanced classifiers with complicated decision boundaries, such as (nonlinear) Support Vector Machines and Neural Networks, and to find solutions for computing the real-world-robustness of high dimensional classifiers.

REFERENCES

[1] Anderson, E. The Species Problem in Iris. Annals Of The Missouri Botanical Garden. 23 pp. 457-569 (1936)
[2] Breiman, L. Random Forests. Machine Learning, 45, 5-32 (2001), http://dx.doi.org/10.1023/A
[3] Cario, M. & Nelson, B. Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix. (Northwestern University,1997,4)
[4] Chen, H., Zhang, H., Boning, D. & Hsieh, C. Robust Decision Trees Against Adversarial Examples. Proceedings Of The 36th International Conference On Machine Learning, 97 pp. 1122-1131 (2019,6,9), https://proceedings.mlr.press/v97/chen19m.html
[5] Chen, H., Zhang, H., St, S., Li, Y., Boning, D. & Hsieh, C. Robustness Verification of Tree-based Models. Advances In Neural Information Processing Systems. 32 (2019), https://proceedings.neurips.cc/paper/2019/file/cd9508fdaa5e139b6e9c329001cf1459-Paper.pdf
[6] Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings Of the 22nd ACM SIGKDD International Conference On Knowledge Discovery And Data Mining. pp. 785-794 (2016), https://doi.org/10.1145/2939672.2939785
[7] Cohen, J., Rosennfeld, E. & Kolter, Z. Certified Adversarial Robustness via Randomized Smoothing. Proceedings Of The 36th International Conference On Machine Learning, 97 pp. 1310-1320 (2019,6,9), https://proceedings.mlr.press/v97/cohen19c.html
[8] Deng, L. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. Signal Processing, Magazine, IEEE. 29 pp. 141-142 (2012,11)
[9] Devroye, L. Non-Uniform Random Variate Generation. (Springer-Verlag,1986)
[10] Genz, A. Numerical computation of multivariate normal probabilities. Journal Of Computational And Graphical Statistics. 1. 141-149 (1992)
[11] Ghosh, A., Manwani, N. & Sastry, P. On the Robustness of Decision Tree Learning Under Label Noise. Advances In Knowledge Discovery And Data Mining. pp. 685-697 (2017)
[12] Goodfellow, I., Bengio, Y. & Courville, A. Deep Learning. (MIT Press,2016), http://www.deeplearningbook.org
[13] Hendrycks, D. & Dietterich, T. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. ArXiv. abs/1903.12261 (2019)