2017-05-27

Tom Rochette <tom.rochette@coreteks.org>

December 21, 2025 — 77e1b28a

0.1 Context

0.2 Learned in this study

0.3 Things to explore

1 Problems faced

2 Overview

Today I wanted to do one-class classification, that is, determine whether a new set of observations belongs to a set of training data or not, without providing it with any negative examples. In my case, I gave it 247938 examples of characters from the NIST Special Database 19, which contains characters `[A-Za-z0-9]`.

Initially I wanted to use sklearn.svm.OneClassSVM, but apparently my 28x28 images and the amount of examples I have is too large for it. on the Outlier detection with several methods, they also list sklearn.ensemble.IsolationForest as a method to do anomaly detection, which is what I tried next.

Initially, I provided IsolationForest with only positive examples. Upon testing it against random noise and characters, it returned the noise as 100% being outliers which was a good sign, but for some of the characters it returned them as outliers as well, which was odd, but whatever. Upon trying it out on some "real" data, I found out it was not really able to make the difference between characters and random drawings. Thus, I looked at IsolationForest to discover it has a 0.1 contamination value, meaning that 10% of the dataset used to fit is assumed to have outliers, which wasn't true in my case. I initially tried to set the contamination value to 0, but doing so would result in the noise examples being classified as inliners, which is not what I wanted.

The next step thus was to add a bit of noise examples, such that I would have about 0.1 contamination amount and then retrain using IsolationForest. At this point all the noise examples are properly classified as outliers and a small amount of inliners are misclassified as outliers, which is fine. However, once again, any sort of doodle would be considered as a character, which most likely meant that the decision boundary was considering these doodle as inliners instead of outliers, which would possibly mean I would need to generate additional negative examples to better define the boundary.

3 See also

4 References

- https://datascience.stackexchange.com/questions/310/one-class-discriminatory-classification-with-imbalanced-heterogenous-negative-b
- https://stats.stackexchange.com/questions/225701/what-exactly-is-a-background-class-in-a-classification-problem