Deployment of unsupervised learning in a search for new physics with ATLAS Open Data

Analysis of two lepton final state from OpenData at ATLAS

Sakarias Frette

Spring 2022

Abstract

University of Oslo

CONTENTS

I. INTRODUCTION

II. THEORY

Some of theory sections about anomaly detection and machine learning algorithms are based on previous work done in other courses, such as [3] and [2].

A. Anomaly detection

Anomaly detection is a tool with a wide range of uses, from time series data, fraud detection or anomalous sensor data. Its main purpose is to detect data which does not conform to some predetermined standard for normal behavior. The predetermined standard varies from situation to situation, and can be set by the context it self, and what is expected as an anomaly. We typically classify anomalies in three categories[1]:

1. Point anomalies
2. Contextual anomalies
3. Collective anomalies

For the purpose of this report we will mostly consider collective anomalies, as anomalies in the standard model has to be collective to claim anything due to noise etc.

In high energy physics we can using machine learning separate anomaly detection into two categories, supervised and unsupervised searches.

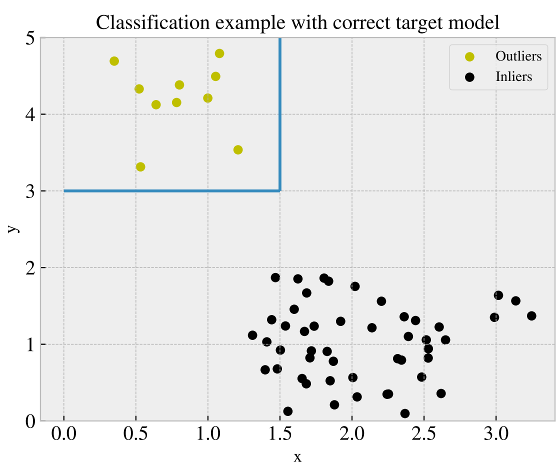

Figure 1: Example of supervised classification of anomaly where the target model is correct. Here the machine learning model manages to correctly identify the anomalies.

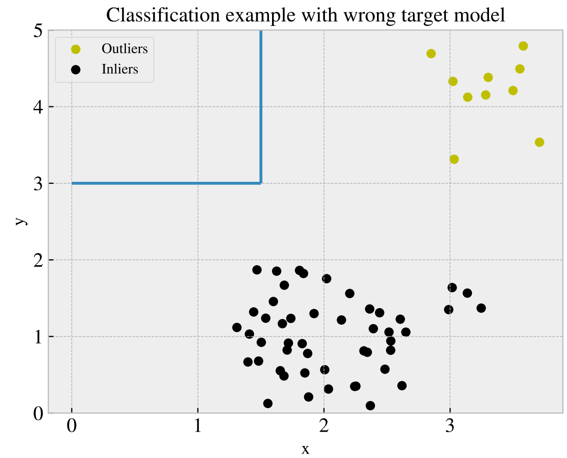

Figure 2: Example of supervised classification of anomaly where the target model is wrong. Here the machine learning model manages to wrongly identify the anomalies.

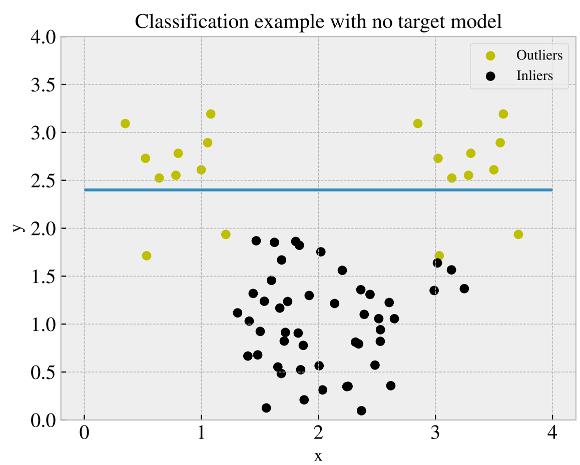

Figure 3: Example of unsupervised classification of anomaly with no target model. Here the machine learning model manages to with good precision identify the anomalies.

B. Stacked auto encoder

One method to attack the anomaly detection problem is the so called (stacked) auto encoder. The idea is based on reconstruction, and has been implemented for denoising of images, image compression, and anomaly detection. An auto encoder is a subgroup of feed forward neural networks, and the goal is to compress the information into fewer variables, called the latent space, which then can explain much of the data through decoding that information. This allows the algorithm to learn the most important components of the data. An illustrative image of an auto encoder is shown below.

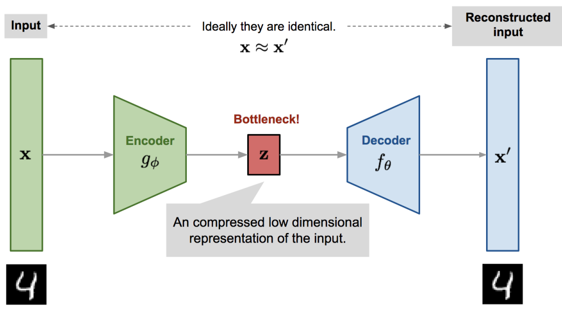

Figure 4: The architecture of an autoencoder with image reconstruction as the example use. Source, accessed 28.04.22

We see here in figure 4 that the input data is deconstructed through an encoder to a lower dimensional space called the latent space, depicted here as $\mathbf{z}$, and then reconstructs the data with the decoder. It is important to note that the number of layers, nodes per layer and activation function per layer for the encoder does not need to match the structure of the decoder. The only requirement is that the input and output layer has the same shape. The end result is the reconstructed data $\mathbf{x'}$. The training of the model is parameterized in the following way. We define the decoder as

$$\mathbf{z} = g_\phi(\mathbf{x}),$$

and the reconstructed information as

$$\mathbf{x'} = f_\theta(g_\phi(\mathbf{x})),$$

where the parameters (ϕ, θ) are tuned to reconstruct the data as close to the original data as possible.

C. Monte Carlo simulated data

D. Beyond standard model physics

III. IMPLEMENTATION

A. Handling of data

1. *Cuts and pre-selection*

2. *Choice of features*

- Which features to pick, amount etc
- How to handle features with no value/nan value (preserve laws of physics)

B. Tuning and training

IV. RESULTS AND DISCUSSION

A. Semi unsupervised

- Difference between brad and narrow search
- Bias
- bias with features to choose
- difference of how well the model detects different signal models in mc data signals
- any interesting in actual data?

V. CONCLUSION

REFERENCES

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar. "Anomaly Detection: A Survey". In: *ACM Comput. Surv.* 41 (July 2009). DOI: 10.1145/1541880.1541882.

[2] Sakarias Frette and William Hirst. "Seaching for Higgs through supervised and unsupervised machine learning algorithms". In: (2022). URL: https://github.com/WilliamHirst/Advanced-Machine-Learning/blob/main/article/Project_1.pdf.

[3] Sakarias Frette, William Hirst, and Mikkel Metzch Jensen. "A computational analysis of a dense feed forward neural network for regression and classification type problems in comparison to regression methods". In: (2022), p. 5. URL: https://github.com/Gadangadang/Fys-Stk4155/blob/main/Project%202/article/Project_2_current.pdf.