1 Introduction

In the last few chapters, we have tried to develop a model to estimate the product cost. There are two main problems that can hinder the estimator to achieve a high performance, namely, presence of **outliers** and **novelty**. These two concepts prevent the model to make a good estimation either by decieving or by not appropriety extrapolating to the new data points. What an outlier and a novelty datapoints have in common is that both deviate substantially from the statistic metrics of the dataset.

2 Definition: Outliers vs. Novelty

`Study last three papers and sklearn documentation`

Notice that the target value of outliers is a random number in the same expected range. This is a sensible assumption, otherwise if the target has a same form of functionality of the features, even out of range, lead to the same trained model.

Outlier detection helps to make a more robust machine learning model in different ways. However imagine that the generating provess evolve over time and that can be a source of change in data pattern. On the other hand, a comparatively sudden change in the pattern does not necessarily mean that the data point is a outlier.

Imagine a factory, if the factory introduce a new product into its production line, there would be a some new data points that are not fallen into the region of past point bu they populate a new region in the feature space. And here comes the actual difference between outliers and novelty. Since there is no unique process of generating outliers (if it would then you can simply make a predictive model for them.) they are scattered in the feature space. On the other hand novelties due to their underlying process occupy cetrain limited in the feature space with high frequency.

1. Add a formal definition for outlier and novelty.
2. Make few actual or fictitious, relevant or errelevant examples from industry.
3. How these two concepts are different from theoritical and practical perspective?

Here we identify two specific forms of outliers:

- Outliers are uniformly distributated in range of inliers
- Outliers are isolated in the feature space

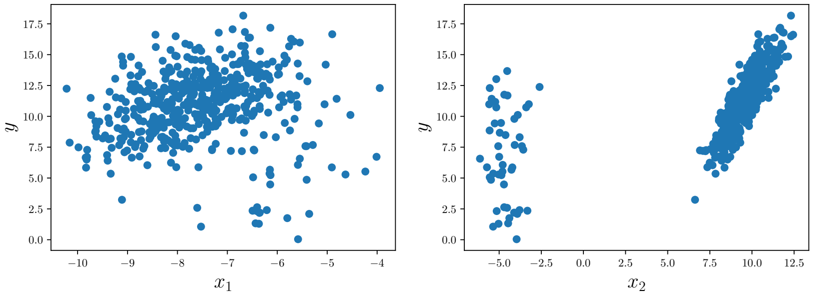

Figure 1: Example of two bi-modal dataset with two features. We can identify two isolated density dataset.

3 Approaches

Here we take two main approches toward the problem. (i)**Supervised** and (ii) **Unsupervised** methods.

4 Unsupervised Approaches

What are the advantages?

4.1 Simple statistical tools

Comparing the distance of data point to the center of data and variation of data.

4.2

Extend the idea from previous section for a non-unimodal data distribution. In such cases the previous recipe fails. What we can do is to compare the distance with the local density.

The local density is determined by **KNN**

4.3 Isolation Forest

Outliers have shorter tree branch length compare to the outliers

5 Supervised

5.1 Robust Regression on clustered data

The approach that we present in this section is to build a robust regression by mean of clustering. The process starts by applying a *proper* clustering method to the dataset. Two different outcomes are expected:

1. There are few clusters that contain the majority of the dataset.
2. Dataset is distributed among clusters almost uniformly.

If the outcome 1 happens, the regression model of interest will be trained on the sorted clusters based on their population from large to small. The test error will be measured every time a new cluster is added to the training dataset. The accumulated clusters with minimum generalization error will be the final dataset to train a regression model.

If the outcome 2 happens, then there would be no preferance among the clusters to start with. The best course of action would be similar to *cross validation* by taking one cluster out and train on the rests. The best model can be reached either by taking average of parameters if a model is linear, or exclude the cluster that when it is added to the training dataset increase the generalization error dramaticly. The later approach is tacken when avering of parameters does not make any sense such as case of decision tree.

(number of cluster is large to make sure that the elbow criteria is passed).

5.2 Curse of Dimensionality

When the number of features is too large, the clustering method lose its reliability.

5.3 Algorithm

- Clustering with large number of clusters
- Apply clustering, make sure you pass elbow criteria.
- Sort cluster based on their population.
- Start training from high population accumulatively.
- Measure generalization error.
- Plot error agains clusters.

6 Summary

References

[1] Breunig, Markus M., et al. "LOF: identifying density-based local outliers." *Proceedings of the 2000 ACM SIGMOD international conference on Management of data.* 2000.

[2] Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. “Isolation forest.” *2008 eighth ieee international conference on data mining. IEEE,* 2008.

[3] Schölkopf, Bernhard, et al. “Estimating the support of a high-dimensional distribution.” *Neural computation 13.7* (2001): 1443-1471.

[4] Hodge, Victoria, and Jim Austin. “A survey of outlier detection methodologies.” *Artificial intelligence review 22.2* (2004): 85-126.

[5] Sabokrou, Mohammad, et al. “Adversarially learned one-class classifier for novelty detection.” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2018.

[6] Carreno, Ander, Inaki Inza, and Jose A. Lozano. “Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework.” *Artificial Intelligence Review* (2019): 1-20.