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 us 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.

2 Definition: Outliers vs. Novelty

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?

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 not-unimodal data distribution. In such case the previous recipe fails. What we can do is to compare the distance againt the local density.

The local density is studies 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

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.