Report: machine learning--unsupervised learning

Ganghao Li

lighao@bu.edu

1. Introduction to the topic

There is often a problem in real life: the lack of sufficient prior knowledge makes it difficult to manually label categories or manually label them. Naturally, we want computers to do the work for us, or at least to help. The various problems in pattern recognition are solved according to training samples whose categories are unknown (not marked), called unsupervised learning.

Two of the main methods used in unsupervised learning are principal component and cluster analysis. Cluster analysis is used in unsupervised learning to group, or segment, datasets with shared attributes in order to extrapolate algorithmic relationships. This approach helps detect anomalous data points that do not fit into either group.

Clustering divides the data set into different categories based on the data characteristics, so that the data within the category is relatively similar/correlated, and the data similarity/correlation between the categories is relatively small. When we apply clustering to unsupervised, we focus on two properties. The first property is consistency and the second property is association.

2. Analysis

Pros:

(1) No label & Clustering

Supervised learning is to choose classifiers and determine weights. Unsupervised learning is density estimation (look for descriptive data statistics), which means that unsupervised algorithms can start working as long as they know how to calculate similarity.

(2) Reduce dimension

If the supervised input is n-dimensional, the feature is considered to be n-dimensional and

usually does not have the ability to reduce dimensions. Unsupervised learning need to extract features, or use layer/item clustering to reduce the dimension of data features.

(3) Non-independent

For different scenarios, the distribution of positive and negative samples may have offset. Because of the connection between the distribution of data, as a training sample, large offsets are likely to cause noise to the classifier, but for unsupervised learning's situation is much better.

(4) Interpretable

The reason for the classification of supervised algorithms is unclear, because these rules are derived by manual modeling and cannot be self-generated. So it is difficult to apply to scenarios requiring clear rules. Unsupervised clustering is well explained. Because the elements in one group have similar features and consistency.

(5) Expandability

A n-dimensional model, if is added into a very strong feature, which is enough to break up the original classification or clustering. In supervised learning, the weights will change almost completely. The unsupervised algorithm developed by DataVisor is extremely scalable. No matter how high the weight of this multi-dimensional data is, it does not affect the original result output. The original result can still be retained, and only needs to process new dimension data. .

Cons:

(1) Accuracy and validity

Because researchers can set the label for the supervised learning in advance, which means supervised learning can be run in a more reasonable way under control. But for unsupervised learning, it learns objects in its own logic analysis without any one interfering, so it may have unpleasant results.

3. Applied field

(1) Data mining

Unsupervised learning is often used for data mining to find out what is in a large amount of unlabeled data. Its training data is unlabeled, and the training goal is to classify or distinguish observations.

(2) Abnormal detection

Unsupervised learning classifies the objects according to their features, so if the abnormal objects have little consistency with other groups, they will be detected. For example, it always be applied to anti-fraud in finance industry.

(3) Detect a segment objects

It can be used in recognizing segment objects. For example, Computer Vision in Unmanned vehicle will use this method to detect the roads, lines and traffic lights.

(4) Advertisement

Applying this method to find users' interest based on what the users usually search or browse and then advertise precisely.

4. Recommendations to systems

(1) Clustering is gathering the observations into groups, each of which contains one or several features. Proper extraction of features is the most critical aspect of unsupervised. So people who use this method should focus on this point.

(2) Compared with supervised learning, clustering is less valid and accurate, so if the data set is independent, maybe sometimes use supervised learning is better.

5. Conclusions

Although our world is almost inundated with data, a large part of it is unmarked and unorganized, which means that it is not available for most current supervised learning. So unsupervised learning can solve this problem to some extent.

The core of unsupervised learning is to classify the objects based on elements features' consistency and relativity without any label. The main method of unsupervised learning is clustering, and this method can be applied in many aspects of out lives, like anti-fraud, unmanned vehicles and advertising.

6. Summary of references

[1] Unsupervised learning definition https://en.wikipedia.org/wiki/Unsupervised_learning
[2] Roman, Victor (2019-04-21). "Unsupervised Machine Learning: Clustering Analysis". *Medium*. Retrieved 2019-10-01.
[3] Unsupervised Learning, Mathworks
[4] Machine Learning for Humans, Part 3: Unsupervised Learning, Medium
[5] Introduction to Unsupervised Learning | Algorithmia Blog