The machine learning landscape

Different types of machine learning:

1. Supervised learning

In supervised learning, the training data you feed to the algorithm includes the desired solutions. A good example of supervised learning is how the spam filter works. It is trained with many emails along with their class and the filter must learn how to classify the emails using the labels attached to the emails.

k-Nearest neighbors, the algorithm we used in week 1, is also an algorithm that uses supervised learning. In the example of week 1, the product ID was the label.

2. Unsupervised learning

Virtually the same as supervised learning, with the difference being that the training data is now unlabeled, and the system learns without it being coached by a label.

A good example that I found of unsupervised learning:

Say you have a lot of data about your blog's visitors. You may want to run a clustering algorithm to try to detect groups of similar visitors (Figure 1-8). At no point do you tell the algorithm which group a visitor belongs to: it finds those connections without your help. For example, it might notice that 40% of your visitors are males who love comic books and generally read your blog in the evening, while 20% are young sci-fi lovers who visit during the weekends, and so on.

Clustering is one way of unsupervised learning. This would look something like this:

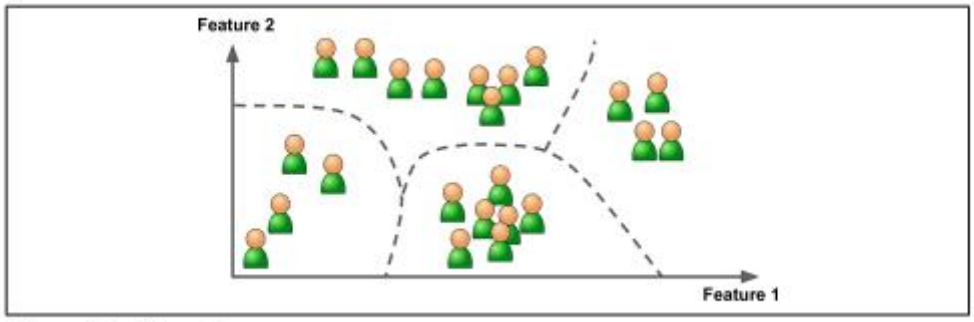

Figure 1-8. Clustering

One of the key tasks of unsupervised learning is something called dimensionality reduction, which basically means that it wants to simplify the data without losing too much information. This is achieved by something called feature extraction, which takes 2 characteristics that have much in common, and essentially merging them together.

Another key task is something called anomaly detection. This means that it will scan your data, and look for oddities. For example, if you have a dataframe of different dogs, and most dogs weigh around 20-40 kilograms, and there is suddenly a dog that weighs 100 kilograms, that means that the heavy 100 kg dog is probably an anomaly, which will mess up the machine learning accuracy.

3. Batch learning

With batch learning, the system is unable to learn incrementally. It takes all the data available, which takes a lot computing resources and time, which is why it is most of the time done offline. The system learns, then is thrown into production, unable to learn anything anymore. If you want the system to learn anything new, you'll have to start from scratch.

4. Online learning

Here we train the system incrementally by feed it data instances sequentially. This is great for systems that need to be continually updated. Think for example of stock prices, they are updated every few seconds to bring the latest prices to the people. This is done by online learning, it keeps training the system with new data.

Online learning can also be usefull with bigger datasets that cannot fit in 1 machines main memory. The algorithm chops the dataset in smaller parts and trains on that data, and keeps doing that until all the data has been used.

A big challenge with online learning is finding the right learning rate. This means how fast it should keep adapting to new data. If you have a high learning rate, the system will forget old data really quick, which in some cases is not wanted, and sometimes not important. For the stock prices, you can have a high learning rate, because the old data is irrelevant when there is new data. For spam filters, you want a slower learning rate, because you want the system to recognize the old spam aswell as the new spam.

5. Instance based learning

This is the most trivial forming of machine learning, which, in some cases, is enough. The system simply learns by heart. The system learns a certain dataset, and when new data comes in, it just compares that new data to the old one. It generalizes new cases using similarity measure.

6. Model based learning

Here you build a model from a set of examples, and then uses that mode to make predictions.

I don't understand how this machine learning method works entirely.