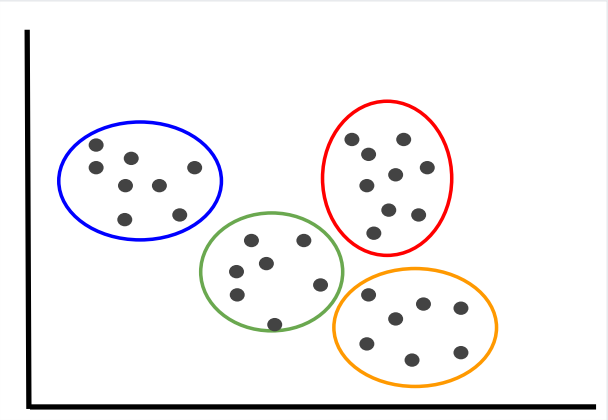
Hard ML Problems

This section explicitly calls out particularly challenging problems in ML.

Clustering

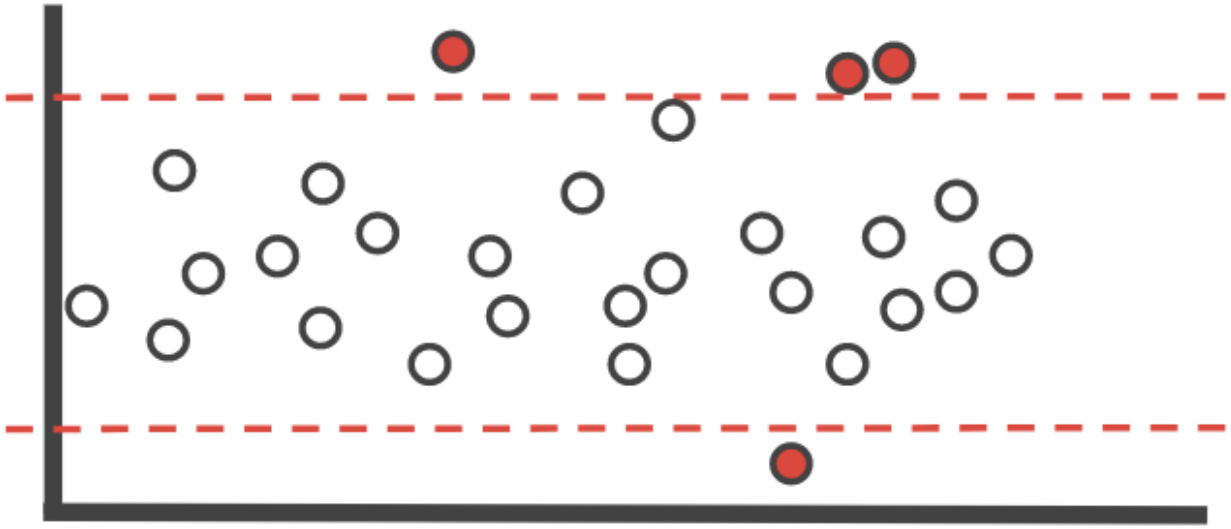
What does each cluster mean in an unsupervised learning problem? For example, if your model indicates that the user is in the blue cluster, you'll have to determine what the blue cluster represents.



Sometimes, you can take actions based on clustering. For example, Google Photos uses clustering to group pictures of the same person together. Other times, it is challenging to determine what action to take based on the cluster. You can try to assign a meaning to a cluster, but this can be tricky because the model might not group by criteria that you find intuitive.

One alternative approach is to label some items *before* you cluster, and then try to propagate those labels across the entire cluster. For instance, if all items with label X end up in one cluster, maybe you can spread label X to other examples.

Anomaly Detection



Sometimes, people want to use ML to identify anomalies. The trick is, how do you decide what constitutes an anomaly to get labeled data? One option is to define a heuristic and use it to label anomalies. However, once you've defined this heuristic, you might as well use the heuristic in your production system, since an ML model can't beat the heuristic used to train it.

You can sometimes craft a high-precision low-recall heuristic, and then use a semi-supervised approach t
el to grow from a "seed" set of predictions to also classify a larger set of unlabeled data.

If your heuristics are sufficiently complicated, then it may be worth considering ML to replace that system. However, be careful going forward since you won't be able to refine the model as easily as you refined your heuristics.

Causation

ML can identify correlations—mutual relationships or connections between two or more things. Determining causation (one event or factor causing another) is much harder. In other words, it is easy to see that something happened, but much harder to understand *why* it happened.

Example

Did consumers buy a particular book because they saw a positive review the week before, or would they have bought it even without that review?

You can't determine causation from only observational data. As in the example above, you can't determine whether the review caused the purchase just by looking at past events. You would need to run an experiment, comparing users who didn't see the review with similar users who did. In general, you need to intervene in the world—run an experiment—to determine causation; you can't see it in purely observational data.

No Existing Data

As previously mentioned, if you have no data to train a model, then machine learning cannot help you. Without data, use a simple, heuristic, rule-based system. Many new products with no training data start with a heuristic rule system, and obtain training data only after users interact with it. Once you have training data, try to find patterns in it. If there are no patterns or only trivial patterns, then machine learning probably will not provide value. If there are many patterns and it is important to make accurate predictions, then using machine learning might be the right approach.

Previous
← Identifying Good Problems for ML (/machine-learning/problem-framing/good)

Next
Deciding on ML (/machine-learning/problem-framing/framing) →

Last updated 2020-02-04 UTC.