Identifying Labels and Sources

Direct vs. Derived Labels

Machine learning is easier when your labels are well-defined. The best label is a **direct label** of what you want to predict. For example, if you want to predict whether a user is a Taylor Swift fan, a direct label would be "User is a Taylor Swift fan."

A simpler test of fanhood might be whether the user has watched a Taylor Swift video on YouTube. The label "user has watched a Taylor Swift video on YouTube" is a **derived label** because it does not directly measure what you want to predict. Is this derived label a reliable indicator that the user likes Taylor Swift? Your model will only be as good as the connection between your derived label and your desired prediction.

Label Sources

The output of your model could be either an Event or an Attribute. This results in the following two types of labels:

- **Direct label for Events**, such as “Did the user click the top search result?”
- **Direct label for Attributes**, such as “Will the advertiser spend more than $X in the next week?”

Direct Labels for Events

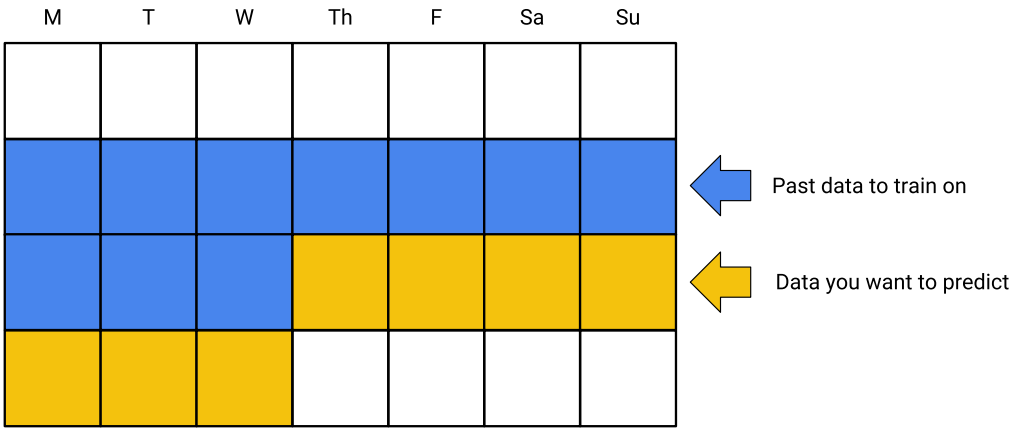
For events, direct labels are typically straightforward, because you can log the user behavior during the event for use as the label. When labeling events, ask yourself the following questions:

- How are your logs structured?
- What is considered an “event” in your logs?

For example, does the system log a user clicking on a search result or when a user makes a search? If you have click logs, realize that you'll never see an impression without a click. You would need logs where the events are impressions, so you cover all cases in which a user sees a top search result.

Direct Labels for Attributes

Let's say your label is, “The advertiser will spend more than $X in the next week.” Typically, you'd use the previous days of data to predict what will happen in the subsequent days. For example, the following illustration shows the ten days of training data that predict the subsequent seven days:



Remember to consider seasonality or cyclical effects; for example, advertisers might spend more on weekends. For that reason, you may prefer to use a 14-day window instead, or to use the date as a feature so the model can learn yearly effects.

e event data carefully to avoid cyclical or seasonal effects or to take those effects into account.

Direct Labels Need Logs of Past Behavior

In the preceding cases, notice that we needed data about the true result. Whether it was how much advertisers spent or which users watched Taylor Swift videos, we needed historical data to use supervised machine learning. Machine learning makes predictions based on what has happened in the past, so if you don't have logs for the past, you need to get them.

What if You Don’t Have Data to Log?

Perhaps your product doesn’t exist yet, so you don't have any data to log. In that case, you could take one or more of the following actions:

- Use a heuristic for a first launch, then train a system based on logged data.
- Use logs from a similar problem to bootstrap your system.

- Use human raters to generate data by completing tasks.

Why Use Human Labeled Data?

There are advantages and disadvantages to using human-labeled data.

Pros

- Human raters can perform a wide range of tasks.
- The data forces you to have a clear problem definition.

Cons

- The data is expensive for certain domains.
- Good data typically requires multiple iterations.

Improving Quality

Always check the work of your human raters. For example, label 1000 examples yourself, and see how your results match the raters'. (Labeling data yourself is also a great exercise to get to know your data.) If discrepancies surface, don't assume your ratings are the correct ones, especially if a value judgment is involved. If human raters have introduced errors, consider adding instructions to help them and try again.

Looking at your data by hand is a good exercise regardless of how you obtained your data. Andrej Karpathy did this on ImageNet and wrote about the experience (http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/).

Previous
← Joining Logs (/machine-learning/data-prep/construct/collect/joining-logs)

Next
Check Your Understanding →
(/machine-learning/data-prep/construct/collect/check-your-understanding)

Last updated 2019-08-20 UTC.