Contrastive Learning vs Regularization

Learning is generalization

One of the key features of intelligence is the ability to learn.

Learning is often done from past data (large datasets), or through experience (interacting with environment and recording outcomes; learning which interactions work well and which do not).

Either way, we get observations, potentially with a label to indicate what kind of observation it is (e.g. positive or negative, etc.). We need to learn the rules (or weights) that will mimic or recall or generate future data that resembles those training observations.

The learning method needs to learn to distinguish things—either between different labels or in case of unsupervised learning, between plausible observations and unlikely ones).

Without this "ability to distinguish" there is no learning (e.g. a machine or human that consumes all data and always return *true* on all future inputs has not learned anything).

Recall

One "easy" learning method is to simply record past observations, and recall them when needed. If we have not seen something in the past, then it does not exist.

This is the domain of databases. Unless a customer performed some action, we will not find that action in the database—and a pure recall function of the database cannot tell us how likely that action is for that customer.

In other words, this learning method does not generalize at all.

Assumptions

To learn to generalize, we need to make assumptions. One key assumption is that what we observe is merely a small sample of what is possible to observe.

That means that between every positive and negative training instance, there is an infinite spectrum of unobserved positive and negative instances—and somewhere in between there's a boundary that separates positive and negative instances.

The learning task is to learn that boundary, so we can quickly check any future instance against that boundary.

Contrastive Learning

The basic idea behind contrastive methods is that we have labeled data, and we are learning a model to separate labeled instances by labels. Once we learn the model, we can apply that model on future instances.

For unsupervised tasks (when we have no labels), we can synthesize a lot of random instances to "fill in" the empty voids between instances we actually observe.

There are several issues with this approach. It works well when we have clean labeled data—which often we do not. In high dimensions, randomly generated instances do not fill up the space in-between observed instances.

For example, if we are labeling images, then there are more "random" images than there are atoms in the universe.

Regularization

Regularization is the idea that instead of building the boundary by using (potentially generated) negative instances, what if we learn the boundary by shrink-wrapping the positive (or observed) instances—and treating everything outside the shrink-wrap as negative.

The tightness of the 'shrink-wrap' can be adjusted via a regularization parameter that can be minimized.