Machine Learning: Context

36-290 – Statistical Research Methodology

Week 7 Tuesday – Fall 2021

What is Machine Learning?

The short version:

- Machine learning (ML) is a subset of statistical learning that focuses on prediction.

The longer version:

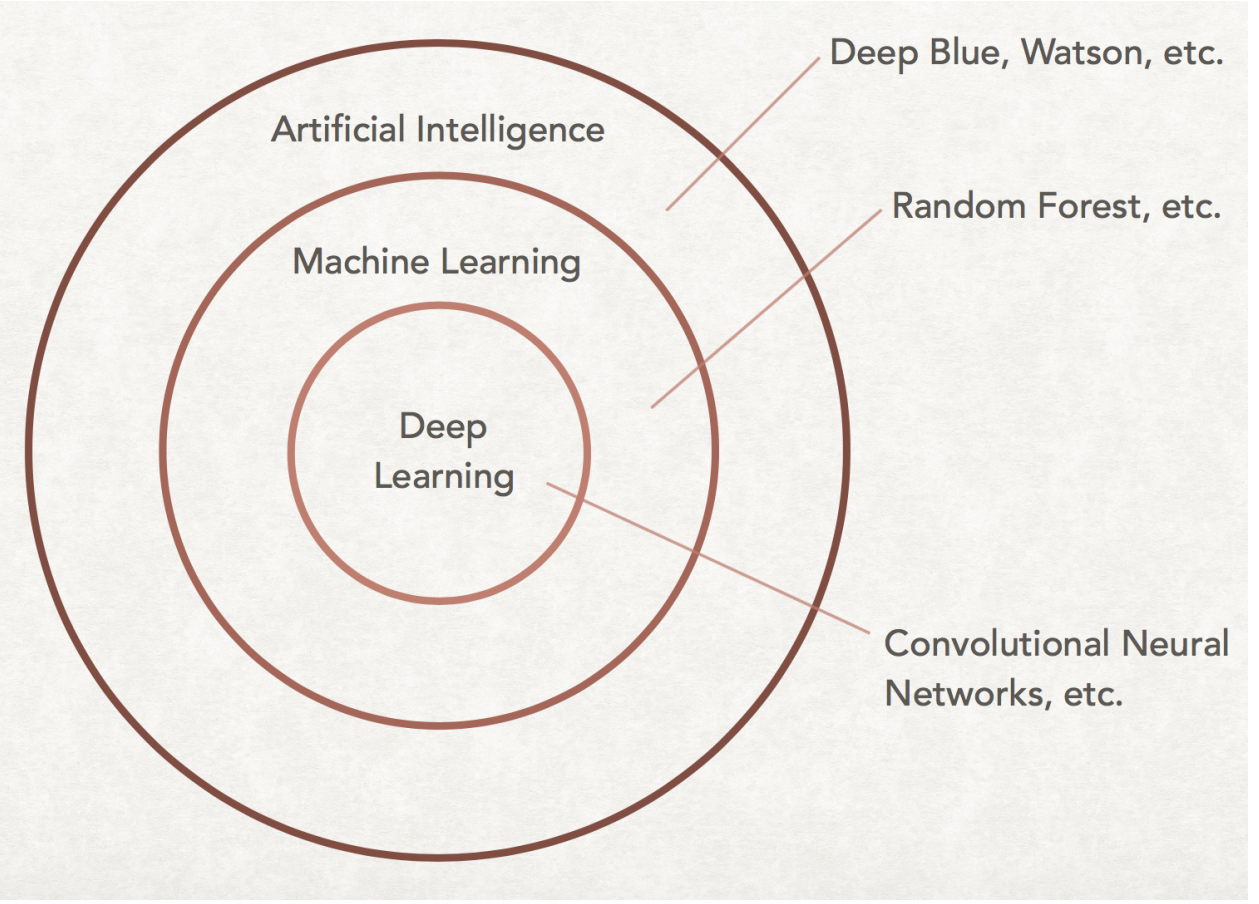
- ML is the idea of constructing data-driven algorithms that *learn* the mapping between predictor variables and response variable(s). Specifically, we suppose no parametric form for the mapping *a priori*, even if technically one can write one down *a posteriori* (for example, by translating a tree model to a indicator-variable-filled mathematical expression).

- Linear regression, for instance, is not a ML algorithm since we can write down the linear equation ahead of time, but random forest is a ML algorithm since we've no idea what the number of splits will end up being in each of its individual trees.

However, note that despite the statement that ML "focuses on prediction," there are those (e.g., Cynthia Rudin) who are leading an effort towards developing "interpretable ML," for instance through the construction of globally optimal logical models (models akin to trees). (See, e.g., the `corels` package and its documentation.)

- Note: do not confuse "interpretable ML" (which focuses on constructing interpretable models) with "explainable ML" (which are post-hoc attempts at explaining the results of black-box algorithms).

Machine Learning: the Broader Context



Machine Learning: Which Algorithm is Best?

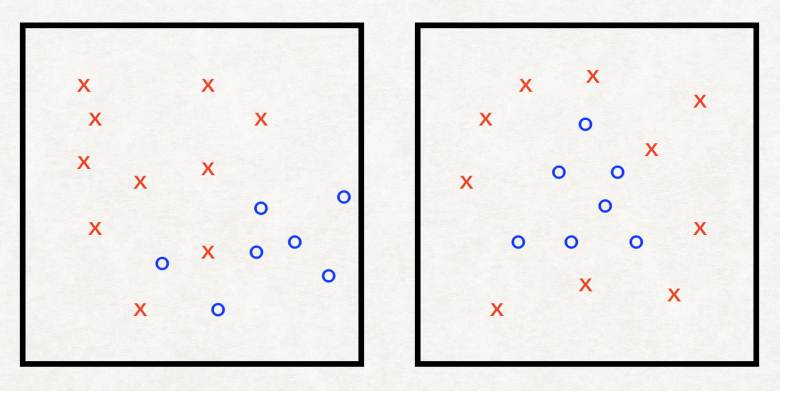
That's not actually the right question to ask.

(And the answer is *not* deep learning. Because if the underlying relationship between your predictors and your response is truly linear, *you do not need to apply deep learning*! Just do linear regression. Really. It's OK.)

The right question is ask is: why should I try different algorithms?

The answer to that is that without superhuman powers, you cannot visualize the distribution of predictor variables in their native space. (Of course, you can visualize these data *in projection*, for instance when we perform exploratory data analysis.) And the performance of different algorithms will be predicated on how predictor data are distributed...

Data Geometry



The picture above shows data for which there are two predictor variables (along the x-axis and the y-axis) and for which the response variable is binary: x's and o's. An algorithm that utilizes linear boundaries or segments the plane into rectangles will do well given the data to the left, whereas an algorithm that utilizes circular boundaries will fare better given the data to the right.

"do well/fare better": will do a better job at predicting whether a new datum is actually an x or an o.