What is overfitting? (abstract)

ML literature has made lots of distinctions between different types of overfitting. Adaptive, benign, tempered, catastrophic and sequential overfitting are just some recent examples. It is often unclear to which foundational concept of overfitting these types refer. The most readily agreed upon foundational concept is generally stated in terms of difference between training and test error, or in comparison with the Bayes optimal error. It states that overfitting occurs if training error is very different from test or Bayes error. The acceptable difference is to be determined by the modeler and can depend among other considerations on the noise model. A second concept of overfitting relates a model's simplicity to its generalization performance. Here overfitting occurs if in the same context a simpler model has better generalization performance than a more complicated one. The measure of simplicity is usually in the eye of the beholder.

Often these concepts of overfitting are not clearly separated in the literature which can lead to apparent puzzles. For example the distinction between benign and catastrophic overfitting only runs counter to "[the] deeply ingrained statistical intuition[s], [that] fitting noisy training data exactly does not necessarily result in poor generalization" (Belkin 2021, p.16) if one expects simpler models to generalize better. A puzzle thus only appears if one employs the second concept of overfitting.

It is philosophically tempting to make such puzzles disappear by mere conceptual shifts, for example by enforcing the first concept. On the other hand our "deeply ingrained statistical intuitions" should not be discarded lightly. Which concept should we prefer in the light of these competing desiderata? I will discuss answers to this question by connecting the two concepts of overfitting to a well known problem in philosophy of science: accommodation vs. prediction and the null-support thesis. According to the null-support thesis a model that is (over)fitted to the data (i.e. a model that accommodates the data) is epistemically worse than one that predicts them independently. Arguably the null-support thesis is wrong in general (Howson 1990). It immediately follows that only very special cases of overfitting are problematic. These are cases where the overfitted model and a competing non-overfitted model have the same epistemic support. Only in these cases the null-support thesis is valid and can be used for model choice. For ML models this means one has to make their epistemic support explicit. This is obviously a hard problem and there can be no general inductive method to solve it. But it explains why there is a second notion of overfitting which relates to model simplicity. Model simplicity is seen as a proxy for an objective measure of epistemic support. And the null-support thesis then prescribes, everything else being equal, to chose the simplest, least overfitted model. The ML practitioner is now in the unenviable position to decide if everything else is equal. This choice I conclude has, perhaps prematurely, already been made in derived concepts like benign or catastrophic overfitting, thus hiding it from plain sight.

Belkin, Mikhail. “Fit without Fear: Remarkable Mathematical Phenomena of Deep Learning through the Prism of Interpolation.” arXiv, May 29, 2021. https://doi.org/10.48550/arXiv.2105.14368.

Howson, Colin. “Fitting Your Theory to the Facts: Probably Not Such a Bad Thing After All.” Minnesota Studies in the Philosophy of Science 14 (1990): 224–44.