Summarizing the Course in Seven Points

36-462/662, Spring 2022

28 April 2022 (Lecture 27)

1. Statistical learning is about trying to find decision rules that will work well, on average, on new data.
    - Implications: We need to choose what kind of rules we can work with (model class), and we need to choose what counts as "working well" (loss function).
    - Even the best decision rule might not work very well.
2. Our models all use the principle of "what would have worked well, on average, on similar cases in the past?"
    - We assume that the future will be like the past, or at least that we understand how it will differ.
    - The differences between models are in how they decide what to count as "similar cases", and, less importantly, how to do the averaging.
3. What counts as a "similar case" is determined by what features we use to represent cases, so building good features is usually more important than the choice of a specific model class.
    - This requires knowledge of the data-generating process, plus trial and error, plus building on the tradition of what's worked well on similar problems in the past.
    - Methods like dimension reduction can help but there are no guarantees.
4. We usually tune our models by optimizing on our training data.
    - Statistical theory can tell us a lot about how well we can expect to optimize a noisy function.
    - Sometimes we impose constraints, or penalties, on the optimization, in the hope that this will guide the search in good directions.
5. **Never** trust in-sample performance as an estimate of true risk.
    - **Always** cross-validate, or use a test set, or apply a well-designed penalty from optimization theory.
6. There is always a bias-variance trade-off.
    - Flexible model classes which are capable of predicting accurate in many different situations and with high-dimensional data need lots of information to figure out which prediction rule to implement; they converge slowly. Simple, rigid model classes that converge rapidly will make systematic mistakes, even with unlimited training data, unless the prejudices built into the model class happen to match reality.
    - Implication: Try to bias your models to what you know about the process you're dealing with.
    - Implication: Fantasies about throwing unlimited data into an AI and having it learn everything will remain fantasies.
7. You get, at best, what you optimize for.
    - It doesn't make a lot of sense to optimize some criterion and then complain that the result isn't also good in some other way. When we select our decision rules to minimize risk, we should not expect them to do anything other than have low loss, on average, on new data.
    - Implication: If we want them to do anything else, like be fair, we need to change what we optimize for. That means we need to be clear enough about what we want that we can program it into the optimization. It also means that there will be trade-offs, on the frontier, between our criteria.