Inductive bias

Daniel Hsu (COMS 4771)

Theorem (informal). Assume $(\vec{X}_1, Y_1), \ldots, (\vec{X}_n, Y_n)$ (training data) and $(\vec{X}, Y)$ are IID random examples, each taking values in $\mathbb{R}^d \times \{0,1\}$. Let $\mathcal{F}$ be a family of binary classifiers (each mapping $\mathbb{R}^d$ to $\{0,1\}$), and assume it can be ordered as $\mathcal{F} = \{f_1, f_2, \ldots\}$.[1] With high probability (over the realization of the training data), for every $k \in \mathbb{N}$,

$$\mathrm{err}(f_k) \;\leq\; \min_{f \in \mathcal{F}} \mathrm{err}(f) \;+\; \widehat{\mathrm{err}}(f_k) - \min_{f \in \mathcal{F}} \widehat{\mathrm{err}}(f) \;+\; \left[\text{small number when } n \text{ is large compared to } \log(k)\right],$$

where $\mathrm{err}(f)$ is the error rate of a classifier f, and $\widehat{\mathrm{err}}(f)$ is the training error rate of a classifier f.

Remark 1. Think of the ordering of classifiers in $\mathcal{F}$ as expressing some external "preferences"—i.e., an inductive bias. Classifiers earlier in the ordering are "preferred" over classifiers later in the ordering. Suppose $\mathcal{F}$ contains a classifier with low error rate, i.e.,

$$\min_{f \in \mathcal{F}} \mathrm{err}(f)$$

is small. Moreover, suppose your learning algorithm picks a classifier f_k such that

$$\widehat{\mathrm{err}}(f_k) - \min_{f \in \mathcal{F}} \widehat{\mathrm{err}}(f)$$

is small. (This is zero if f_k happens to be a minimizer of the training error rate.) Then, the theorem implies that, in the IID model, with high probability, the error rate of f_k will also be small as long as the number of training data is large compared to $\log(k)$. So, what qualifies as a large-enough number of training examples depends on the inductive bias!

Remark 2. There are many variants/extensions of this theorem that are applicable in a wide variety of settings, e.g., different families of classifiers, different ways of expressing preferences or encoding inductive biases, different types of prediction problems. But the high-level message in each of them is the same as the one given here.

[1] There is a version of this theorem that works even for uncountable families of classifiers.