Machine Learning — Statistical Methods for Machine Learning

Consistency and nonparametric algorithms

Instructor: *Nicolò Cesa-Bianchi* version of March 26, 2019

Consistency is an asymptotical property certifying that the risk of the predictors generated by a learning algorithm converge to the Bayes risk in expectation as the size of the training set increases. Formally, a learning algorithm A is **consistent** if for any learning problem $(\mathcal{D}, \ell)$ it holds that

$$\lim_{m\to\infty} \mathbb{E}\Big[\ell_{\mathcal{D}}\big(A(S_m)\big)\Big] = \ell_{\mathcal{D}}(f^*)$$

where the expectation is with respect to the random draw of the training set S_m of size m from the distribution $\mathcal{D}$, and f^* is the Bayes optimal predictor for $(\mathcal{D}, \ell)$.

Since f^* can be any predictor, consistent algorithms cannot select their outputs from a restricted class $\mathcal{H}$ of predictors. In other words, given a big enough training set, a consistent algorithm must be able to output any arbitrary predictor f. Algorithms of this kind are called nonparametric because the predictors they generate cannot be described using a pre-determined number of "parameters" (i.e., variables). For example, k-NN is nonparametric because the description of the classifier typically requires a number of variables that scales with the number of training points. Most algorithms that output tree predictors are also nonparametric. This happens whenever the number of nodes in the tree does not have a pre-determined upper bound, but is free to grow with the size of the training set.

The standard k-NN algorithm is not known to be consistent for any fixed value of k. Indeed, one can only show that

$$\lim_{m\to\infty} \mathbb{E}\Big[\ell_{\mathcal{D}}\big(k\text{-NN}(S_m)\big)\Big] \le \ell_{\mathcal{D}}(f^*) + 2\sqrt{\frac{\ell_{\mathcal{D}}(f^*)}{k}} \tag{1}$$

for any data distribution $\mathcal{D}$. However, if we let k be chosen as a function k_m of the training set size, then the algorithm becomes consistent provided $k_m \to \infty$ and $k_m = o(m)$. Similarly, the greedy algorithm for building tree classifiers is consistent whenever the fraction of training examples routed to each leaf is not vanishing as the sample size grows.

In practice, a consistent algorithm may not be preferred over a nonconsistent one. This due to the fact that the rate of convergence to the Bayes risk of a consistent algorithm can be arbitrarily slow, as shown by the following result.

Theorem 1 (No Free Lunch)**.** *Let $a_1, a_2, \ldots$ be a sequence of positive numbers converging to zero and such that $\frac{1}{16} \ge a_1 \ge a_2 \ge \cdots$. For any learning algorithm A for binary classification, there exists a data distribution $\mathcal{D}$ such that $\ell_{\mathcal{D}}(f^*) = 0$ and $\mathbb{E}\big[\ell_{\mathcal{D}}\big(A(S_m)\big)\big] \ge a_m$ for all $m \ge 1$.*

Note that $\ell_{\mathcal{D}}(f^*) = 0$ implies $\eta(\boldsymbol{x}) \in \{0, 1\}$ for each $\boldsymbol{x}$, where $\eta(\boldsymbol{x}) = \mathbb{P}(Y = 1 \mid X = x)$. This means that $\eta : \mathcal{X} \to [0, 1]$ is not a Lipschitz function. Also, Theorem 1 does not prevent a consistent algorithm from converging fast to the Bayes risk for specific distributions $\mathcal{D}$. What the theorem

shows is that if A converges to the Bayes risk for any data distribution, than it will converge arbitrarily slow for some of these distributions.

For binary classification, we can summarize the situation as follows.

- Under no assumption on η, there is no guaranteed convergence rate to Bayes risk.
- Under Lipschitz assumptions on η, the typical nonparametric convergence rate to Bayes risk is $m^{-1/(d+1)}$ (exponentially slow in d).
- Under no assumptions on η, the typical parametric convergence rate (i.e., the convergence rate to risk of the best predictor in a parametric class $\mathcal{H}$) is $m^{-1/2}$, exponentially better than the nonparametric rate.