Abstract

We design a new, fast algorithm for agnostically learning univariate probability distributions whose densities are well-approximated by piecewise polynomial functions. Let f be the density function of an arbitrary univariate distribution, and suppose that f is OPT-close in L_1-distance to an unknown piecewise polynomial function with t interval pieces and degree d. For any $\gamma > 0$, our algorithm draws $n = \widetilde{O}_\gamma(t(d+1)/\epsilon^2)$ samples from f, runs in time $\widetilde{O}(n)$, and with probability at least $9/10$ outputs an $O_\gamma(t)$-piecewise degree-d hypothesis h that is $(3+\gamma)\cdot\mathrm{OPT}+\epsilon$ close to f. Our approximation factor almost matches the best known information–theoretic (but computationally inefficient) upper bound of 3. Our general algorithm yields (nearly) sample-optimal and *nearly-linear time* estimators for a wide range of structured distribution families over both continuous and discrete domains in a unified way. For most of our applications, these are the *first* sample-optimal and nearly-linear time estimators in the literature. As a consequence, our work resolves the sample and computational complexities of a broad class of inference tasks via a single "meta-algorithm". Moreover, we demonstrate that our algorithm performs well in experiments. Our algorithm consists of three levels: (i) At the top level, we employ an iterative greedy algorithm for finding a good partition of the real line into the pieces of a piecewise polynomial. (ii) For each piece, we show that the sub-problem of finding a good polynomial fit on the current interval can be solved efficiently with a separation oracle method. (iii) We reduce the task of finding a separating hyperplane to a combinatorial problem and design a nearly-linear algorithm for this problem. Combining these three procedures gives a density estimation algorithm with the claimed guarantees.