PROBABILITY MASS FUNCTION ESTIMATION
AND AN APPLICATION TO LANGUAGE MODELING

Abstract

We propose a new method for estimating the probability mass function (pmf) of a discrete and finite random variable from a small sample. We focus on the observed counts—the number of times each value appears in the sample—and define the Maximum Likelihood Set (MLS) as the set of pmfs that put more mass on the observed counts than on any other set of counts possible for the same sample size. We characterize the MLS in detail. We show that the MLS is a “diamond”-shaped subset of the probability simplex $[0, 1]^k$ bounded by at most $k(k - 1)$ hyperplanes, where $k$ is the number of possible values of the random variable. The MLS always contains the empirical distribution, as well as a family of Bayesian estimators based on a Dirichlet prior, particularly the well-known Laplace estimator. We propose to select from the MLS the pmf that is “closest” to a fixed pmf that encodes prior information. When using Kullback–Leibler distance for this selection, the optimization problem comprises finding the minimum of a convex function over a domain defined by linear inequalities, for which numerical procedures are presented. We apply this estimate to language modeling, where Zipf’s law plays the role of prior information. We show that this method permits obtaining state-of-the-art results while being conceptually simpler than competing methods.