Online Categorical Subspace Learning for Sketching Big Data with Misses

Abstract:

With the scale of data growing every day, reducing the dimensionality (a.k.a. sketching) of high-dimensional data has emerged as a task of paramount importance. Relevant issues to address in this context include the sheer volume of data that may consist of categorical observations, the typically streaming format of acquisition, and the possibly missing entries. To cope with these challenges, this paper develops a novel categorical subspace learning approach to unravel the latent structure for three prominent categorical (bilinear) models, namely, Probit, Tobit, and Logit. The deterministic Probit and Tobit models treat data as quantized values of an analog-valued process lying in a low-dimensional subspace, while the probabilistic Logit model relies on low dimensionality of the data log-likelihood ratios. Leveraging the low intrinsic dimensionality of the sought models, a rank regularized maximumlikelihood estimator is devised, which is then solved recursively via alternating majorization-minimization to sketch high-dimensional categorical data "on the fly." The resultant lightweight first-order algorithms entail highly parallelizable tasks per iteration. In addition, the quantization thresholds are also learned jointly with the subspace to enhance the predictive power of the soughtmodels. Performance of the subspace iterates is analyzed for both infinite and finite data streams, where for the former asymptotic convergence to the stationary point set of the batch estimator is established, while for the latter sublinear regret bounds are derived for

the empirical cost. Simulated tests with both synthetic and real-world datasets corroborate the merits of the novel schemes for real-time movie recommendation and chess-game classification.