

Title:

Conditional Random Sampling: A Sketch-based Sampling Technique for Sparse Data

Author(s):

Ping Li, Kenneth W. Church, and Trevor J. Hastie

Technical Report number (Dept. of Statistics, Stanford Univ.):

2006-8

Date:

June 2006

Abstract:

We develop a sketch-based sampling algorithm, called *Conditional Random Sampling (CRS)*, particularly suitable for sparse data. In many important applications such as information retrieval, the datasets are often very large and highly sparse. Our technique is a combination of sketching and sampling in that it converts sketches of the data into *conditional random samples* online in the estimation stage, from which we estimate the original space using well-understood statistical methods. In addition, we can take advantage of the marginal information to (often considerably) enhance the estimation accuracy at little incremental cost.

Our method can efficiently approximate pairwise distances (inner product, l_1 , l_2 , or l_p distances) and multi-way associations, as well as many other summary statistics. This study focuses on pairwise l_2 and l_1 distances (and inner products) for which *random projections* are popular. Our method is provably better than random projections in boolean (0/1) data. We show using real-valued text and image data that our algorithm often outperforms random projections in approximating inner product and l_2 distance. In many learning and data mining tasks including association rules, estimating joins, distance-based clustering, nearest neighbor searching, and kernels for (e.g.,) support vector machines (SVM), computing pairwise (or multi-way) distances is usually the vital step. Therefore, *Conditional Random Sampling* will be useful for these applications.