Statistics > Machine Learning

Hashing Algorithms for Large-Scale Learning

Ping Li, Anshumali Shrivastava, Joshua Moore, Arnd Christian Konig

(Submitted on 6 Jun 2011)

In this paper, we first demonstrate that b-bit minwise hashing, whose estimators are positive definite kernels, can be naturally integrated with learning algorithms such as SVM and logistic regression. We adopt a simple scheme to transform the nonlinear (resemblance) kernel into linear (inner product) kernel; and hence large-scale problems can be solved extremely efficiently. Our method provides a simple effective solution to large-scale learning in massive and extremely high-dimensional datasets, especially when data do not fit in memory.

We then compare b-bit minwise hashing with the Vowpal Wabbit (VW) algorithm (which is related the Count-Min (CM) sketch). Interestingly, VW has the same variances as random projections. Our theoretical and empirical comparisons illustrate that usually b-bit minwise hashing is significantly more accurate (at the same storage) than VW (and random projections) in binary data. Furthermore, b-bit minwise hashing can be combined with VW to achieve further improvements in terms of training speed, especially when b is large.

Subjects: **Machine Learning (stat.ML)**; Learning (cs.LG)

Cite as: **arXiv:1106.0967v1 [stat.ML]**

Submission history

From: Ping Li [view email]

[v1] Mon, 6 Jun 2011 06:38:20 GMT (237kb)

Which authors of this paper are endorsers?