Multiview LSA: Representation Learning Via Generalized CCA

Abstract

Multiview LSA is a way of utilizing hundreds of data sources to learn representations for millions of words/phrases that outperform baselines like Word2Vec and Glove [4, 5].

Max-Var GCCA

LSA is an application of PCA to a single term-document cooccurrence matrix. CCA learns linear projections that are maximally correlated to each other from two views. Generalized CCA is a family of extensions of CCA to maximize correlation across multiple views. One variant of GCCA called MAX-VAR GCCA induces an auxiliary representation $G$ that is maximally correlated to linear projections of the views in terms of sum of squared correlations [1, 2].

$$G = \text{eig} \left( \sum_{i=1}^{J} P_i \right)$$

Where, $P_i = X_j(X_j^\top X_j)^{-1}X_j^\top$

Handling Missing Values

Sparse cooccurrence matrices contain plenty of missing values that cripple the performance of methods that rely on spectral decompositions. We address this sparsity by optimizing our representations only on the observed rows using a variant of MAX-VAR GCCA presented by [3].

$$G = \text{eig} \left( \sum_{j=1}^{J} K_j \right)^{-1} \left( \sum_{j=1}^{J} P_j \right) \left( \sum_{j=1}^{J} K_j \right)^{-2}$$

(1)

where $[K_j]_{ii} = 1$ if row $i$ of view $j$ is observed and zero otherwise.

Further Information

- Visit: www.cs.jhu.edu/~prastog3/mvlsa
- Email: pushpendre@jhu.edu

References