GLRM Technical References

- Paper
  - arxiv.org/abs/1410.0342

- Other Resources
  - H2O World Video
  - Tutorials

Generalized Low Rank Models

Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd

May 6, 2015. (Original version posted September 2014.)

Abstract

Principal components analysis (PCA) is a well-known technique for approximating a tabular data set by a low rank matrix. Here, we extend the idea of PCA to handle arbitrary data sets consisting of numerical, Boolean, categorical, ordinal, and other data types. This framework encompasses many well known techniques in data analysis, such as nonnegative matrix factorization, matrix completion, sparse and robust PCA, k-means, k-SVD, and maximum margin matrix factorization. The method handles heterogeneous data sets, and leads to coherent schemes for compressing, denoising, and imputing missing entries across all data types simultaneously. It also admits a number of interesting interpretations of the low rank factors, which allow clustering of examples or of features. We propose several parallel algorithms for fitting generalized low rank models, and describe implementations and numerical results.

Online Resources

- H2O Tutorials and Training Material
  - https://github.com/h2oai/h2o-tutorials
  - https://github.com/h2oai/h2o-tutorials/tree/master/training/lending_club_exercise
  - https://github.com/h2oai/app-consumer-loan
- Datasets
  - https://s3.amazonaws.com/h2o-public-test-data/bigdata/laptop
  - https://s3.amazonaws.com/h2o-public-test-data/bigdata/laptop/lending-club/loan.csv

H₂O.ai

GLRM Technical References

- Paper
  - arxiv.org/abs/1410.0342

- Other Resources
  - H2O World Video
  - Tutorials

Generalized Low Rank Models

Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd

May 6, 2015. (Original version posted September 2014.)

Abstract

Principal components analysis (PCA) is a well-known technique for approximating a tabular data set by a low rank matrix. Here, we extend the idea of PCA to handle arbitrary data sets consisting of numerical, Boolean, categorical, ordinal, and other data types. This framework encompasses many well known techniques in data analysis, such as nonnegative matrix factorization, matrix completion, sparse and robust PCA, k-means, k-SVD, and maximum margin matrix factorization. The method handles heterogeneous data sets, and leads to coherent schemes for compressing, denoising, and imputing missing entries across all data types simultaneously. It also admits a number of interesting interpretations of the low rank factors, which allow clustering of examples or of features. We propose several parallel algorithms for fitting generalized low rank models, and describe implementations and numerical results.