Tackling Big Data with Tensor Methods

Anima Anandkumar

U.C. Irvine
Learning with Big Data
Data vs. Information
Data vs. Information

I'm filling out a reader survey for chewing magazine.

See, they asked how much money I spend on gum each week, so I wrote, "$500." For my age, I put "43." And when they asked what my favorite flavor is, I wrote "garlic/curry."

This magazine should have some amusing ads soon. I love messing with data.
Data vs. Information

- Missing observations, gross corruptions, outliers.
Data vs. Information

- Missing observations, gross corruptions, outliers.
- High dimensional regime: as data grows, more variables!
Data vs. Information

- Missing observations, gross corruptions, outliers.
- High dimensional regime: as data grows, more variables!

Data deluge also a data desert!
Learning in High Dimensional Regime

- Useful information: low-dimensional structures.
- Learning with big data: ill-posed problem.
Learning in High Dimensional Regime

- Useful information: low-dimensional structures.
- Learning with big data: ill-posed problem.

Learning is finding needle in a haystack
Learning in High Dimensional Regime

- Useful information: low-dimensional structures.
- Learning with big data: ill-posed problem.

Learning is finding needle in a haystack

- Learning with big data: computationally challenging!

Principled approaches for finding low dimensional structures?
How to model information structures?

Latent variable models

- Incorporate hidden or latent variables.
- Information structures: Relationships between latent variables and observed data.
How to model information structures?

Latent variable models
- Incorporate hidden or latent variables.
- Information structures: Relationships between latent variables and observed data.

Basic Approach: mixtures/clusters
- Hidden variable is categorical.
How to model information structures?

Latent variable models

- Incorporate hidden or latent variables.
- Information structures: Relationships between latent variables and observed data.

Basic Approach: mixtures/clusters

- Hidden variable is categorical.

Advanced: Probabilistic models

- Hidden variables have more general distributions.
- Can model mixed membership/hierarchical groups.
Application 1: Clustering

- Basic operation of grouping data points.
- Hypothesis: each data point belongs to an unknown group.
Application 1: Clustering

- Basic operation of grouping data points.
- Hypothesis: each data point belongs to an unknown group.

Probabilistic/latent variable viewpoint

- The groups represent different distributions. (e.g. Gaussian).
- Each data point is drawn from one of the given distributions. (e.g. Gaussian mixtures).
Application 2: Topic Modeling

Document modeling

- **Observed**: words in document corpus.
- **Hidden**: topics.
- **Goal**: carry out document summarization.
Application 3: Understanding Human Communities

Social Networks

- **Observed:** network of social ties, e.g. friendships, co-authorships
- **Hidden:** groups/communities of actors.
Recommender System

- **Observed**: Ratings of users for various products, e.g. yelp reviews.
- **Goal**: Predict new recommendations.
- **Modeling**: Find groups/communities of users and products.
Application 5: Feature Learning

Feature Engineering

- Learn good features/representations for classification tasks, e.g. image and speech recognition.
- Sparse representations, low dimensional hidden structures.
Application 6: Computational Biology

- Observed: gene expression levels
- Goal: discover gene groups
- Hidden variables: regulators controlling gene groups
Learning Algorithms through Tensor Factorization

- Co-occurrence of three-words in a document, e.g. [apple, orange, banana].

**Tensor Eigenvectors**

- Can learn the hidden topics by finding tensor eigenvectors.
- Common friends (neighbors) of triplets of nodes in a social networks.
# Experimental Results on Yelp

Lowest error business categories & largest weight businesses

<table>
<thead>
<tr>
<th>Rank</th>
<th>Category</th>
<th>Business</th>
<th>Stars</th>
<th>Review Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Latin American</td>
<td>Salvadoreno Restaurant</td>
<td>4.0</td>
<td>36</td>
</tr>
<tr>
<td>2</td>
<td>Gluten Free</td>
<td>P.F. Chang's China Bistro</td>
<td>3.5</td>
<td>55</td>
</tr>
<tr>
<td>3</td>
<td>Hobby Shops</td>
<td>Make Meaning</td>
<td>4.5</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>Mass Media</td>
<td>KJZZ 91.5FM</td>
<td>4.0</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>Yoga</td>
<td>Sutra Midtown</td>
<td>4.5</td>
<td>31</td>
</tr>
</tbody>
</table>
Experimental Results on Yelp

Lowest error business categories & largest weight businesses

<table>
<thead>
<tr>
<th>Rank</th>
<th>Category</th>
<th>Business</th>
<th>Stars</th>
<th>Review Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Latin American</td>
<td>Salvadoreno Restaurant</td>
<td>4.0</td>
<td>36</td>
</tr>
<tr>
<td>2</td>
<td>Gluten Free</td>
<td>P.F. Chang’s China Bistro</td>
<td>3.5</td>
<td>55</td>
</tr>
<tr>
<td>3</td>
<td>Hobby Shops</td>
<td>Make Meaning</td>
<td>4.5</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>Mass Media</td>
<td>KJZZ 91.5FM</td>
<td>4.0</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>Yoga</td>
<td>Sutra Midtown</td>
<td>4.5</td>
<td>31</td>
</tr>
</tbody>
</table>

Bridgeness: Distance from vector $\left[1/\hat{k}, \ldots, 1/\hat{k}\right]^\top$

Top-5 bridging nodes (businesses)

<table>
<thead>
<tr>
<th>Business</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Four Peaks Brewing</td>
<td>Restaurants, Bars, American, Nightlife, Food, Pubs, Tempe</td>
</tr>
<tr>
<td>Pizzeria Bianco</td>
<td>Restaurants, Pizza, Phoenix</td>
</tr>
<tr>
<td>FEZ</td>
<td>Restaurants, Bars, American, Nightlife, Mediterranean, Lounges, Phoenix</td>
</tr>
<tr>
<td>Matt’s Big Breakfast</td>
<td>Restaurants, Phoenix, Breakfast&amp; Brunch</td>
</tr>
<tr>
<td>Cornish Pasty Co</td>
<td>Restaurants, Bars, Nightlife, Pubs, Tempe</td>
</tr>
</tbody>
</table>
My Research Group and Resources

Furong Huang  Majid Janzamin  Hanie Sedghi

Niranjan UN  Forough Arabshahi

ML summer school lectures available at
http://newport.eecs.uci.edu/anandkumar/MLSS.html