An Application of K-means Clustering to Professor Recommendation

Mackenzie Leake
Scripps College
West Coast Conference for Undergraduate Women in Physics
January 19, 2014
Outline

• Our professor recommendation system
  • Data
  • Clustering
  • Interface

• Can these techniques be applied to other datasets?
Recommendation Systems

- Common systems
  - Yelp
  - Amazon
- Our goal
  - Generate professor recommendations for students that provide information beyond what advisors or course catalogs can offer
Preparing the data

1. Professor reviews from Rate My Professors
   • Numerical scores for clarity, easiness, helpfulness, and overall quality
   • Written student reviews
2. Scrape written reviews for 229 Pomona College professors (Raman 2012)
3. Filtering
4. Create a matrix of all words (8170) from all reviews
5. Term frequency-inverse document frequency (tf-idf) (Cai 2012)
K-means Clustering

- Unsupervised machine learning algorithm
- Finds structure within unlabeled data
- Partitions data into $k$ groups based on Euclidean distance from randomly initialized centroids

A two-dimensional clustering of separated random data (Mathworks 2013)
Clustering

• Choose the number of clusters (k)

Elbow plot to choose k-value to minimize intra-cluster variance
User Input

- User selects 1-3 professors from the drop down menu (EasyGui 2013)

Our interface to allow users to select professors from database
Professor Selection

1. Locate cluster label of each input professor.
2. Find 4 nearest professors to input.
3. Select 5 most highly rated professors from candidates.
4. Return recommendation.

Diagram of recommendation generation
Cluster Validation

- Subjectively

<table>
<thead>
<tr>
<th>Cluster 0 (120)</th>
<th>Cluster 1 (27)</th>
<th>Cluster 2 (40)</th>
<th>Cluster 3 (11)</th>
<th>Cluster 4 (31)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard</td>
<td>Nice</td>
<td>Awesome</td>
<td>Psych</td>
<td>Amazing</td>
</tr>
<tr>
<td>Fun</td>
<td>Relatively Easy</td>
<td>Is awesome</td>
<td>Social</td>
<td>Passionate</td>
</tr>
<tr>
<td>Work</td>
<td>Fairly Easy</td>
<td>Reading</td>
<td>Social psych</td>
<td>The subject</td>
</tr>
<tr>
<td>Material</td>
<td>Super nice</td>
<td>Guy</td>
<td>Political</td>
<td>Really lovable</td>
</tr>
</tbody>
</table>

Cluster labels, sizes, and most common words

- Objectively

Average easiness score for each cluster

[Graph]

Average clarity score for each cluster

[Graph]
Is K-means clustering a helpful tool for other applications?

• Advantages
  • Requires no labeled data
  • Can provide relevant structure within data
  • Runs relatively quickly

• Disadvantages
  • Must specify number of clusters (k)
  • Hard clustering
Other Applications of K-means Clustering

- Taxonomic classification of asteroids and clustering chemical species on a Mars hyperspectral image (Galluccio et al. 2008)
- Three classes of gamma-ray bursts (Chattopadhyay et al. 2008)
- Stable isotope ratios in methylamphetamine (Salouros et al. 2013)
Acknowledgements

• I would like to thank my fellow project team members, Arianna Perkins and Chelsea Fried, for their hard work and friendship throughout this project.

• I would also like to thank Professor America Chambers for her guidance this semester.
References