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

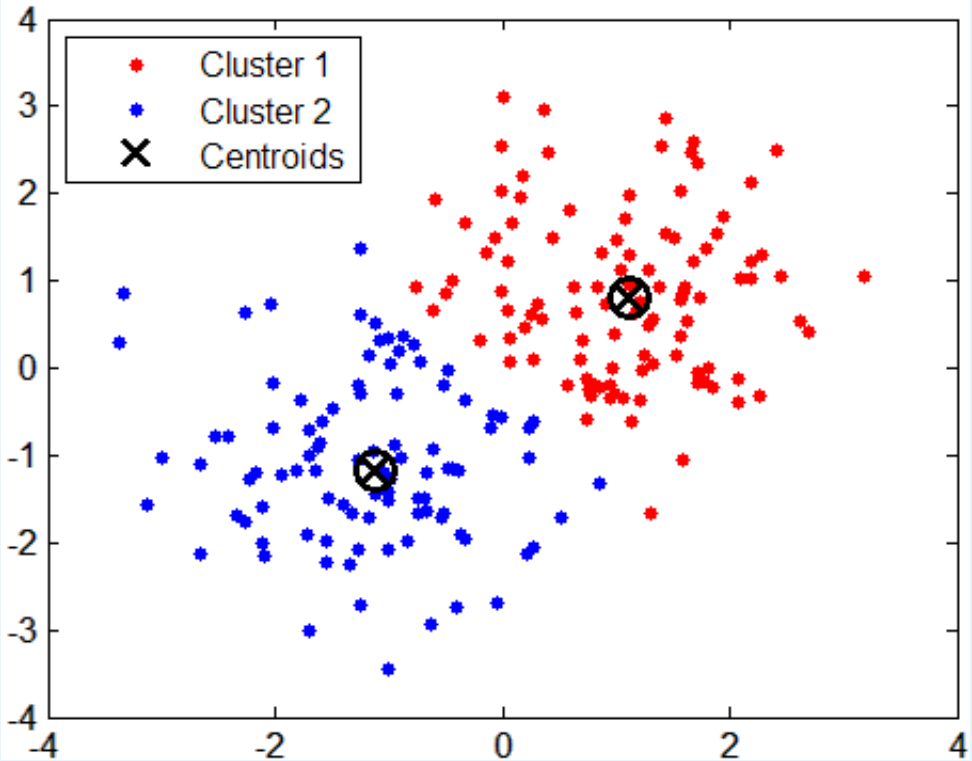
- Common systems
  - Yelp
  - Amazon
- Our goal
  - Generate professor recommendations for students that provide information beyond what advisors or course catalogs can offer

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

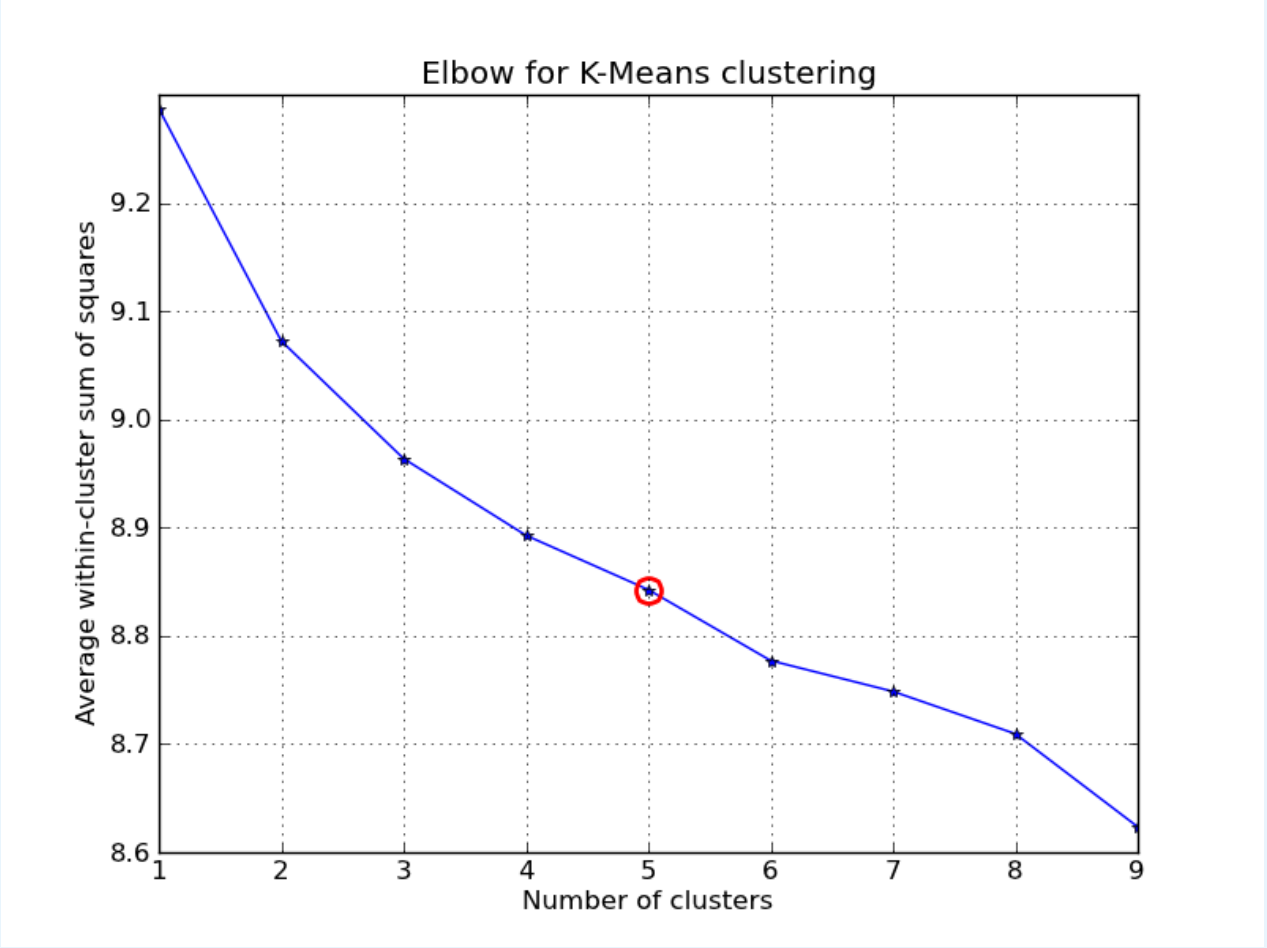
- Unsupervised machine learning algorithm
- Finds structure within unlabeled data
- Partitions data into k groups based on Euclidean distance from randomly initialized centroids



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

Clustering

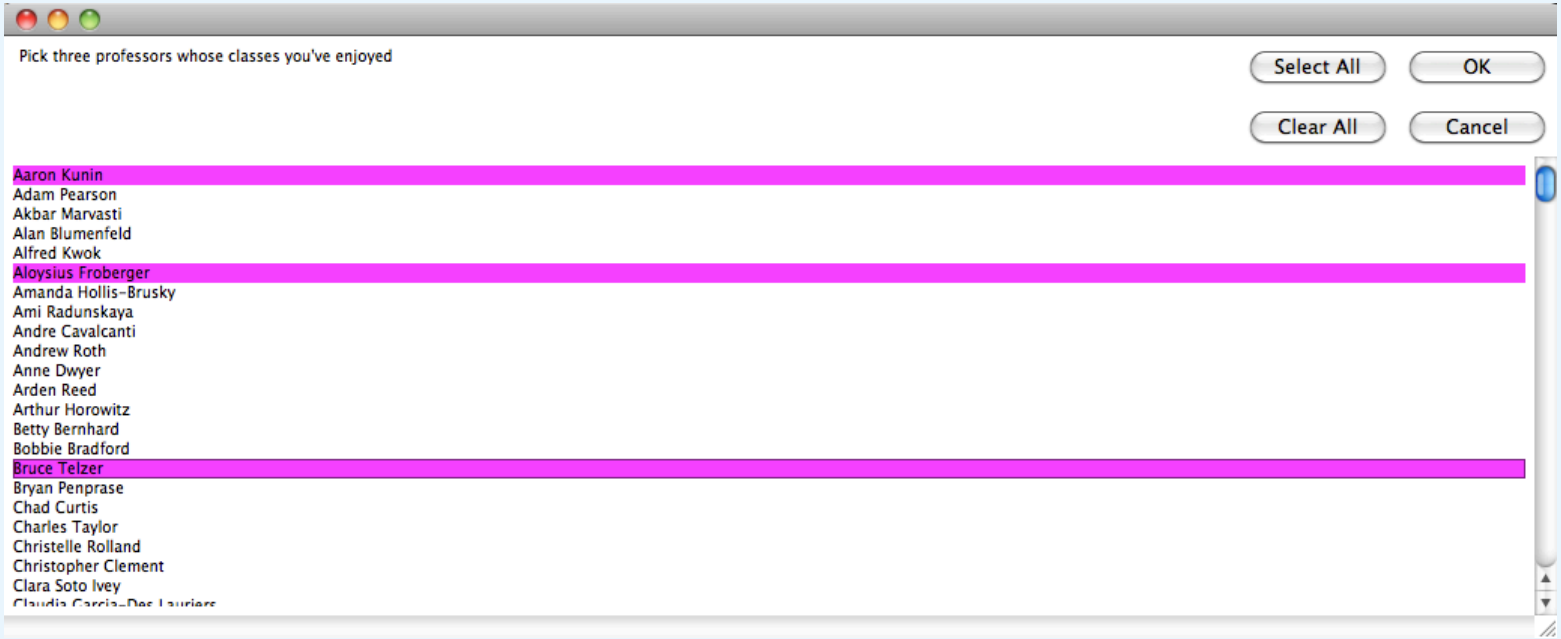
- Choose the number of clusters (k)



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

User Input

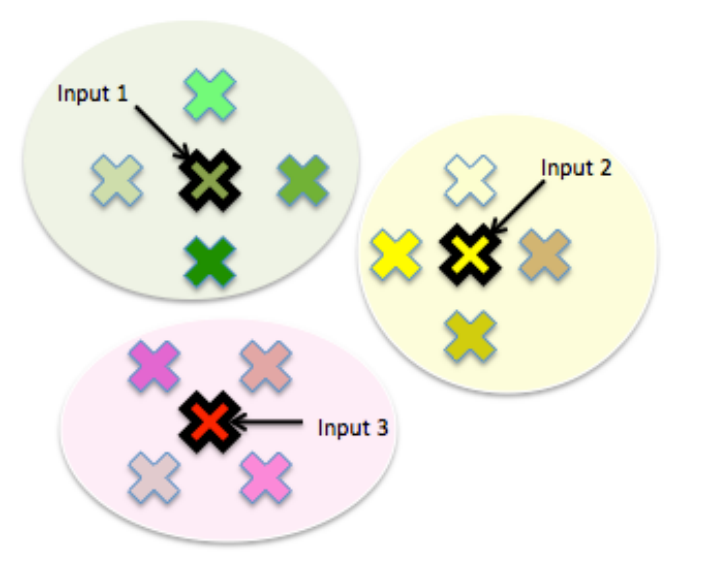
- User selects 1-3 professors from the drop down menu (EasyGui 2013)



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.



Return	Because you input

Diagram of recommendation generation

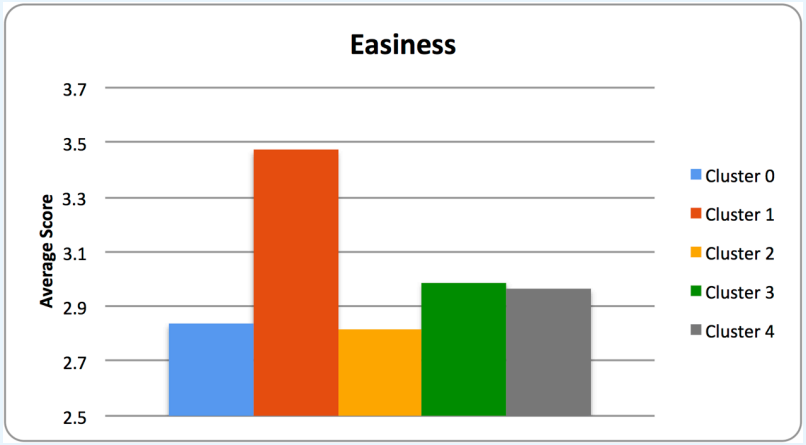
Cluster Validation

- Subjectively

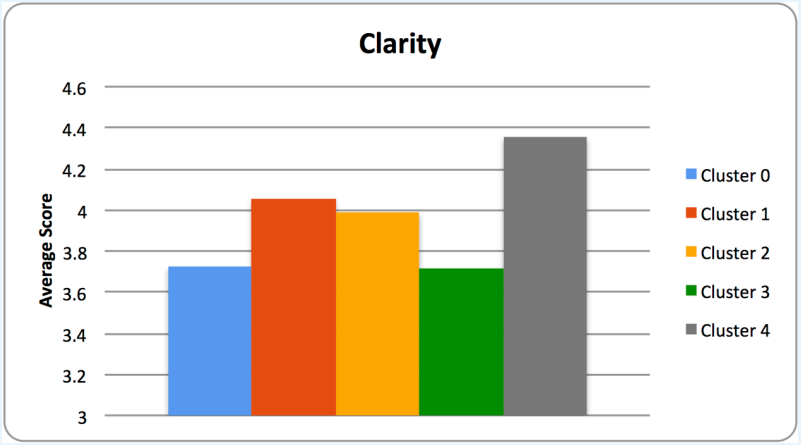
Cluster 0 (120)	Cluster 1 (27)	Cluster 2 (40)	Cluster 3 (11)	Cluster 4 (31)
Hard	Nice	Awesome	Psych	Amazing
Fun	Relatively Easy	Is awesome	Social	Passionate
Work	Fairly Easy	Reading	Social psych	The subject
Material	Super nice	Guy	Political	Really lovable

Cluster labels, sizes, and most common words

- Objectively



Average easiness score for each cluster



Average clarity score for each cluster

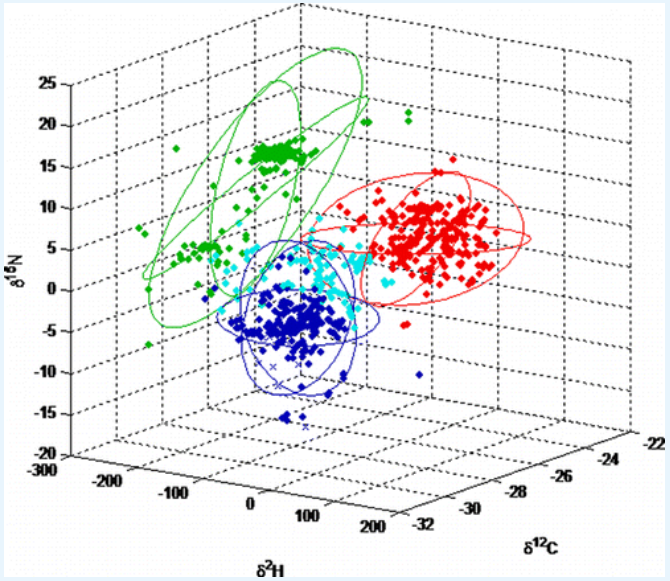
Is K-means clustering a helpful tool for other applications?

- Advantages
  - Requires no labeled data
  - Can provide relevant structure within data
  - Runs relatively quickly
- Disadvantages
  - Must specify number of clusters (k)
  - Hard clustering

(Naik)

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)



K-means clustering of isotope ratios of nitrogen, hydrogen, and carbon for three industrial routes to produce 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

1. Cai, Deng. *Tfidf.* 2012. http://cad.zju.edu.cn/home/ dengcai/Data/code/tfidf.m.
2. Chattopadhyay, Tanuka, et al. "Statistical Evidence for Three Classes of Gamma-Ray Bursts." Astrphys. J. 667:1017-1023, 2007.
3. EasyGui. Accessed December 9, 2013. http://easygui.sourceforge.net.
4. Galluccio, L., et al. "Unsupervised Clustering on Astrophysics Data: Asteroids Reflectance on Spectra Surveys and Hyperspectral Images." AIP Conf. Proc. 1082, 165, 2008.
5. Mathworks. Accessed January 15, 2014. http://www.mathworks.com/help/stats/kmeans.html.
6. Naik, Azad. Data Clustering Algorithms. Accessed January 15, 2014. https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm.
7. Raman, Karthik. Rate-Professors Scraping. October 7, 2012. http://cs.cornell.edu/~karthik/projects/rateprof-scrape.
8. Rate My Professors Review Professors and Teachers, School Reviews, College Campus Ratings. Accessed December 9, 2013. www.ratemyprofessors.com.
9. Salouros, Helen, et al. "Measurement of Stable Isotope Ratios in Methylamphetamine: A Link to its Precursor Source." Anal. Chem. 85(19): 9400-9408, 2013.
10. SciPy. Accessed December 9, 2013. http://www.scipy.org.