Where Should I Live?

Nicholas Choo (nchoo@stanford.edu)
Stanford University

Introduction

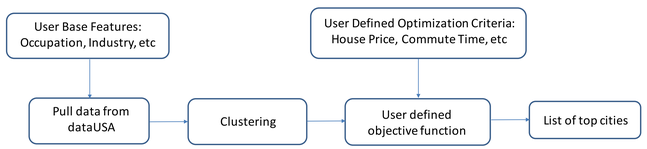
DataUSA is a collaboration designed to structure US public data into a more easily accessible format. The dataset is a union of many different types of career, income and personal demographic data that can be used as a numerical generalization of a geographic region. We want to study:

How can this data be generalized to identify career optimal regions for a young professional?

Predicting & Data

- DataUSA contains hundreds of metrics for thousands of geographical regions for thousands of occupations and industries.
- Set limits by specifying occupation and industry, limiting the geographic level to PUMAs – public use microdata areas with population > 100,000. n = 2079
- Aim to identify the city that best matches the 'feature vector' of an individual. The output of the algorithm will be a list of cities the algorithm 'believes' will be a best fit
- Cleaned by removing all non numerical values, margin of error features, and part time working data which was often incomplete.
- PUMAs that had under 100 individuals in target occupation were removed to remove outlying data.
- PUMAs that did not have a complete set of data since certain algorithms did not deal with uncertainties as well as other data

Data Pipeline

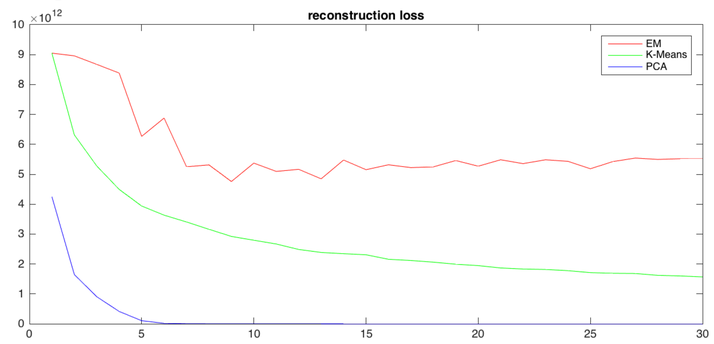


Features

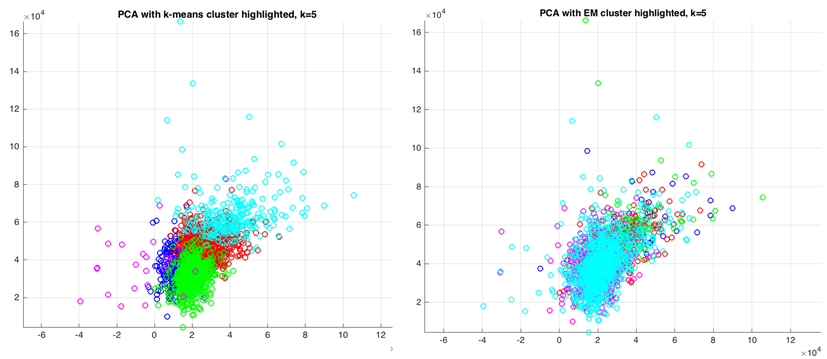
- Cleaned by removing all non numerical values, margin of error features, and part time working data which was often incomplete.
- PUMAs that had under 100 individuals in target occupation were removed to remove outlying data.
- PUMAs that did not have a complete set of data since certain algorithms did not deal with uncertainties as well as other data

Models & Results

For modeling, k-means and expectation maximization algorithms were investigated, PCA was used to visualize



Average reconstruction loss for k = 1 to 30 was calculated with 10 samples



PCA with 2 principle components was graphed to illustrate clustering at k = 5

Models & Results Continued

$$J(c,\mu) = \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$$

Reconstruction Loss Formula

- EM performance was sporadic, even after averaging
- K-means performed better than EM at a given k
- EM clusters were visually not as 'tight' as k-means clusters
- PCA had the least reconstruction loss, however it did not fit into clustering/prediction framework

Top Cities for Electronics Engineer with Priorities
Wage = 10, House Price = 9, Commute Time = 3

	K-means	EM
1	Austin, TX	King County (North East), WA
2	Greater Bellevue City, WA	Huntington Town, NY
3	San Diego, CA	Walnut Creek, CA
4	Alameda County, CA	Santa Clara County, CA
5	Huntington Town, NY	LA Calabasas, Malibu & Westlake, CA

Discussion

- Low 'elbow' of reconstruction loss for data suggests there are about 5-7 archetypes of cities, variation in city types is not high
- However, there are many outlier 'cities' with low population in the given occupation which skew the top cities list
- K-means had better performance than EM in most cases, however both often converge to local minima
- Very subjective results, difficult to interpret 'good or bad' decisions from the algorithm, only ways to quantify accuracy is through mathematical means via loss functions
- Clustering provides an expectation where your person 'vector' should work, it may be different than where you would like to work

Future Work

- Use models to predict census data from cities in other countries
- Add on additional features such as immigration data, hiring data to model chances of finding work in certain region
- Examine hierarchal clustering and density based clustering methods