Predicting Future Employment, Productivity and Income (Finding Patterns in Economic Data)

Jake McKinnon [#1]

[#] *Computer Science Department, Stanford University*

[1] jakemck@stanford.edu

***Abstract*—We implement and evaluate several methods for making predictions using, or finding patterns in, heterogeneous economic data. Though both supervised and unsupervised algorithms are used, the primary focus is on k-means clustering and logical extensions for use with non-continuous data. By altering the dissimilarity function and centroid update procedure, k-means can be made to perform reasonably well on mixed or categorical data while retaining the other characteristics of k-mean clustering—namely, relatively good speed on larger data sets. Such a dissimilarity function can be defined either using domain-specific knowledge, or in a generalizable manner.**

I. Introduction

Whether making decisions about individual career paths or about overarching government policy, people often want to know what causes there are for a number of different economic outcomes. People commonly choose their majors and careers with the belief that they will be economically secure on such a path. Further, our society promotes education, marriage, home ownership and more on the basis that we believe such factors to contribute to or cause wellbeing for individuals and society. But are we correct in these assumptions? And moreover, on the margin, where we are forced to trade off a dollar spent on one program for a dollar spent on another, which program makes best use of the additional dollar? These are questions that ought to be answered thoroughly, rather than assumed or ignored.

Better mathematical evidence using economic data may be one way to help robustly answer these questions. And from technology companies to baseball teams [1], statistical methods have helped many organizations be more effective in recent years. If certain factors can be shown to be strong predictors or causes of desirable outcomes, then perhaps better policy decisions can be made, grounded in hard evidence rather than intuition.

To help find such factors, I attempt to use census data to predict individual outcomes and to look at how different observations are associated with each other. For instance, if there is a cluster of healthy, happy, wealthy people who share certain unique characteristics or history, we would like to know about it. Though this is far from perfect justification for policies—pushing on these characteristics may break the association with good outcomes for any number of reasons—at a minimum, it provides a way to make predictions about existing people, which is itself valuable. Perhaps further work could prove which relationships are robust to external pressure, such that policies might move push both the characteristic and the outcome in the desired directions.

II. Related Work

There are two dimensions to this project: one within economics and one within machine learning. From the economics perspective, it is hard to point to specific papers as a starting point because this is such a broad target. There is a good deal of discussion and confusion about the causes of worker productivity, which seems to be a major point of both confusion and research [7][8]. Further, there are a number of papers about applying clustering methods to the problem of market segmentation [2].

From the machine learning perspective, though we briefly tried a number of prediction methods, the primary focus of this paper is on k-means and extensions into settings where traditional k-means breaks down. Therefore, most of the machine learning papers we referenced have to do with discussions of k-means and modifications.

Huang (1998) discusses a method for extending k-means to use on categorical or mixed data, called k-modes [4]. Many of our approaches follow from his work or related work [5], or draw ideas from it. The main idea is to find a way to measure category dissimilarity, and use this metric to cluster samples. This method retains the speed of k-means.

Another method called k-medoids updates using sample data points in place of centroids, and can be made to work with categorical data and arbitrary distance functions [3]. Though it is more expensive.

III. Data Set and Features

Many countries, particularly in the first world, collect a good deal of information about their citizens through some form of census. This was the data we aimed to make inferences from, as it was most available and had very large numbers of samples to compensate for the relative imprecision of the data. The primary data set we used some partially-cleaned US census data from 1994 with ~50,000 samples. Though some of the later methods might have been more successful on a different data set, the focus of this application project not so much the particular results, but the process of applying machine learning to the overall category of problem. Extension to another data set was relatively low priority. With the significant exception of our application-specific distance metrics, our methods could be readily extending for use with other census data.

A. Features

The data set used had 15 features. Six were continuous features; the remaining nine were categorical, including binary and non-binary categories. The features included: education-level, gender, race, occupation, age, income-level, marital-status, capital-gain, and native-country. Not that the particular features are not important, except insofar as interpreting findings on this data set. An example from the data set is provided below, as is basic information about some of the more relevant features.

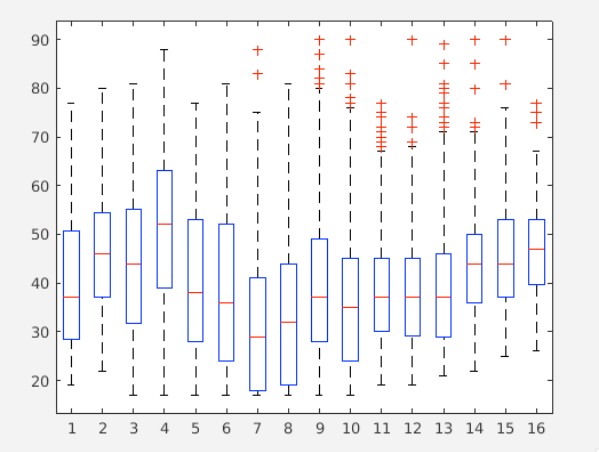

Fig. 1. Graph of Age vs. Educational Attainment (16 = PhD)

Feature	Mean	Standard Dev.
education-num	10	2.5
age	39	14
capital-gain	1078	7385
hours-worked	40	12

Fig. 2. Sample Features. Note that edu-num=9 is high school grad.

B. Preprocessing

Of the 48,841 samples in the data set, 3,620 had missing features. We considered dealing with this a number of ways, but ultimately decided to remove them from the data set as many samples with missing features seemed to be very anomalous in other ways as well—e.g. working very low hours or exceptional mismatch between education and income. This left 43,221 samples. We justify choice, which arguable skewed our data samples to be more average than reality, by noting that trying different methods for dealing with this missing data would be relatively straightforward and that having exceptional or unreliable data points may have made our clustering results unreliable or otherwise more difficult to assess, particularly when custom distance metrics were used.

Further, we normalized the data to have unit variance. When dealing with k-means and extensions, large difference in variance can dramatically overweight certain features depending on dissimilarity measure that is chosen.

If the data set had had more features, we may have used Principal Components Analysis to reduce dimensionality while retaining most of the useful variation in the data set. Though we did not require this for this particular data set, it is worth noting if one finds a much higher-feature data set on which to apply similar methods (which may be useful). Note that this would complicate the discussion of several distance functions later discussed.

Finally, from these processed points, we randomly selected ~2/3 as our training set, with ~1/3 remaining for testing. For clustering these distinctions were less important, though they helped provide a means of testing whether or not the clusters actually existed in all economic data, or was merely a by-product of clustering on a particular data set (e.g. by clustering on one set, then on the other, and comparing clusters).

IV. METHODS

With a project goal of finding interesting relationships in census data, our choice methodology was relatively open. Initially, we looked at simple relationships in the data and tried the classic task of predicting income, with reasonable success using Naïve Bayes. But the primary work was diving into how to apply k-means-like algorithms to census data, which tends to be large. With clustering, our goal to find what similar clusters of people existed in the country, and how this might be associated with desirable features such as income and family stability. The clusters also provide a reasonable partition of that data such that other less-efficient methods of evaluating data might be applied going forward [3].

A. Naïve Bayes Classifier

By defining one of the features collected in the census as the target label, and removing it from the feature set, we have the common setup for our supervised learning problem. We straightforwardly apply classic Naïve Bayes to the problem of predicting income. Initially using the Matlab implementation for continuous features and later our own. Tinkering with inputs, e.g. by removing features, provides some information about which features are most consequential in predicting income, or any other targeted category.

B. K-Means Clustering

We use K-means to see if we can discover any relationships based on the centroids or the clusters created, similar to how one can see relationships with simple graphs such as the age vs. education one above. Ignoring categorical features (as well as continuous features for which Euclidean distance makes less sense, such as education-number), we apply k-means to the data samples. We further apply k-means with different initialization and with differing numbers of clusters.

For the sake of the following discussion, recall how K-means operates: Imagine the (continuous) data samples as points in a high-dimensional Euclidean space, where axes correspond to features. We place a number of centroids either randomly or by some other more intelligent process in the space. Then we group all samples based on whichever centroids they are closest to as judged by the square of the L2 (or Euclidean) norm. Finally, we move the centroids to the mean of all points in its group. Then, we repeat this process with the new centroids. The algorithm terminates when no samples shift grouping.

Note a few things:

1) The algorithm effectively uses Euclidean distance/ the square of the L2 norm to group samples with centroids.
2) K-means is polythetic, so it may not be clear from looking at any 2-d plot why the clustered points are similar to each other (and indeed this held).
3) It has no objective measure of performance built in. At best, it says, "These samples are similar as judged by this particular measure of similarity, which may or may not make any sense."

C. K-Modes Clustering

The most glaring deficit, for our purposes, in classic K-Means is that it doesn't make use of the majority of our census data, which is mostly categorical. This is unfortunate because K-Means is exceptionally fast on very large data sets such as the US Census, as far as clustering algorithms go [4]. However, by looking more carefully at the two observations about k-means above, we can see how one might logically extend k-means into categorical data: by defining a new function by which to assign samples to centroids, one that makes sense in the categorical context, and change our centroids update rule correspondingly.

Specifically, we define the similarity of two objects as follows, where m is the number of samples:

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j)$$

$$\delta(x_j, y_j) = \begin{cases} 0 (x_j = y_j) \\ 1 (x_j \neq y_j) \end{cases}$$

Further, we change our update rule to a frequency based update rule: For each feature, the label of the new centroid becomes the mode of all samples in the group.

Note that this applies to purely categorical data now. To most greatly benefit from this extension, we must reintegrate it with k-means.

D. K-Prototypes Clustering

Combining k-means and k-modes back together, so that we can now deal with heterogeneous data, gives us the k-prototypes clustering algorithm. Strictly speaking, this was defined as flows, where features p through m are categorical and μ is a weight to avoid favouring either attribute type.

$$d_p(X,Y) = \sum_{j=1}^{p-1} (x_j - y_j)^2 + \mu \sum_{j=p}^{m} \delta(x_j, y_j)$$

In practice, we just choose μ it such that the standard deviation of the categorical features, as defined by the k-modes dissimilarity metric, is the same as the standard deviation for the continuous features (effectively, just normalizing the variance of categorical data).

E. Application-Specific Dissimilarity Functions

Further exploring different dissimilarity functions or update rules for k-means-like algorithms, we look towards application or domain-specific distance functions that provide more information about similarity or dissimilarity than the simple equivalence-based metric used in k-modes and k-prototypes. Note that defining a good custom dissimilarity metric may require domain-specific knowledge, and has a number of potential issues that the other methods discussed do not.

1) *Linear or Other Euclidean Metrics for Category Dissimilarity*: One suggestion for custom distance metrics was a distance function that functioned by placing all the categories on the line, and treating distance as the distance between the category points. Such a metric naturally fits for categories like education level, but not others. E.g. a stock analyst might be similar to a doctor (both wealthy and educated) and doctor similar to caretaker, but the stock analyst is not similar to the caretaker. In this example, categorical entries may be similar or dissimilar to each other along many dimensions, which cannot be effectively projected onto a single dimension. Putting categories in a higher-dimensional Euclidean space can help, as it does here, but even this may not be the case for less transitive categories. However, we still made use of this strategy for some categories. It corresponds to assigning each label in a category a higher-dimensional vector and calculating Euclidean distance between any two labels using their assigned vectors.

2) Changing Dissimilarity Measurements for Continuous Features: In certain settings, it may make sense to use a measurement other than the L2-norm to assign samples to centroids. Though this may break convergence guarantees, it can be made to work in certain settings. For example L1-norm, or Manhattan distance, may be used if you were clustering locations in a city based on travel time (this is effectively k-medians). Though we played with such measurements, we didn't explore them deeply.

3) Other Custom Dissimilarity Metrics: Given sufficient knowledge about the domain in which the data is being used, one might have a better metric for judging category differences. In certain cases, manually defining all the relationships between categories might work, but only for very low number of labels.

4) The Difficulty in Judging Dissimilarity Metrics: Note that, for such these dissimilarity measurements, the second problem noted in the K-Means Clustering section is particularly potent: we are defining a distance metric, and we cannot judge our success by this metric, as that would be circular.

Instead, the metrics must be defined and judged as being reasonable in the context of the application (e.g. returning clusters that make sense) or by other means. Experts might decide what is a good metric for dissimilarity as judged by their experience and prior research in that particular field, and on whether the clusters produced are sensible.

5) A Note about Convergence: Note that the procedures described here break the convergence guarantee that k-means typically has. While working on the application, we did not realize this at first because in many cases the algorithm kept working. However, we have no proof of this in the general case and there are likely ways one might use the methods covered here in a way that might break convergence. We acknowledge this, but note that empirically most of the methods discussed seem to converge. There seem to be such guarantees for k-medoids, though the algorithm is slower [3].

V. Results

Brief summary of the results generated by applying the above methods to the US census data set. Most of the discussion is above, although some of the problems discussed do return in results.

A. Naïve Bayes Classifier

When trying to predict income using the 14 other features of the data set, the Naïve Bayes Classifier produced correct labels reasonably well. It had ~17% training error and ~20% generalization error (on the 1/3 random untrained samples). Furthermore, removing categories revealed that marital status/ relationship was the most important feature in predicting higher income.

B. K-Means Clustering

The only continuous variables available on which to k-means cluster were age, education-number (with higher being more educated), capital-gains, capital-loss, and hours-worked. Despite the low number of features, k-means still revealed some interesting relationships. Specifically, it had the following clusters:

-Near-50, well educated, high cap-gains, high hours
-Mid-40s, college, medium cap-gains, medium hours
-Old, medium cap-gains, little college, low hours
-Young, no capital gains, lower education, modest hours

Rerunning the algorithm several times we find that the clusters are consistent even with substantially different initialization points. Increasing the number of clusters helped equalize the number of samples in each cluster by further differentiating between the large mid-age and

young groups. The same basic clusters still present. Note that, though we solved the question of how many centroids to use by trying a variety of numbers, there are more intelligent methods of deciding how many clusters are present [6].

C. K-Modes and K-Prototypes Clustering

Given the low number of categories, there were only a few relationships that showed clearly in the categorical-only k-modes clustering. Marital Statuses were clustered with the corresponding relationship statuses (e.g. wife was also Married-civ-Spouse) and genders exceptionally well, as and work-class and certain occupations (e.g. no self-employed armed-forces).

Extending to K-prototypes gave richer clusters, but they primarily confirmed expected trends. We saw similar clusters to K-means, with the additional information that: the high-cap-gains earners were likely married, in particular occupations, and white. The young people were least likely to clustered with married people. Other obvious clusters

D. Application-Specific Distance Function

Didn't finish this implementation, though with the partial-complete we got better clusters around doctorate and profession-school people with high income as well as divorced, widowed, and separated clustered with Low incomes. May have had difficulty due to the convergence concerns mentioned previously.

VI. Conclusion and Future Work

Though most of the results were nothing new, we still managed to replicate a number of well-known findings. It turns out that, indeed, stable marriages generally promote good life outcomes, as does being white, educated, male, and non-manual-labour occupations. A somewhat interesting finding, though not exactly counterintuitive, is that people who had high capital gains incomes generally worked quite long—over 50 hours on average. At least it wasn't something I was expecting to find as much as those other relationships. Overall, it was interesting to explore several approaches to clustering that we didn't cover in lecture.

E. Future Work

(Had I had more time or other team members, with is what I would have worked on going forward).

Though this data set was very friendly to work with, an obvious next step would be to apply these methods to a data set with more features and that is more recent. Many of the methods outlined generally perform better with a higher number of categorical features, as the means by which they differentiate categories is entirely binary and therefore grainy. This contrasts with continuous data, which can more easily produce intermediate measures of similarity. It would also be interesting to see any changes over time (e.g. women trending towards becoming more educated and higher earning) by comparing the results using this data set with a more recent one.

Another extension to this application that might have improved the observations would have been using the clusters generated by k-means or k-prototypes in further algorithms. By reducing the number of samples on which to run, k-prototypes might allow more computation-intensive learning algorithms to find more original relationships in certain subsets of the population.

Finally, one method of finding many patterns in census or other economic data that maps very straightforwardly onto this setting is Bayesian Networks. However, I avoided this topic because there appeared to have already been a good deal of work in Bayesian Nets that used or referencing such census data. In contrast, there seemed to be less work with K-means—one couldn't apply classic K-means effectively to most census data, which is primarily categorical. However, though Bayesian Networks have been used in this area often before, they would still be a good place to explore recreationally or in any setting where originality is not a priority.

References

[1] J.S. Armstrong, Predicting job performance: The moneyball factor. Foresight: The International Journal of Applied Forecasting, 25 (2012), pp. 31-34. Print.

[2] H. Rezankova. "Cluster Analysis of Economic Data", *Statistika,* vol.94, no. 3, pp. 73-86, 2014.

[3] A. K. Jain, R. C. Dubes, "Algorithms for Clustering Data." Prentice-Hall, 1988.

[4] Huang, Zhexue, "Extensions to the k-means Algorithm for Clustering Large Data Sets with Categorical Values." International Journal of Data Mining and Knowledge Discovery, Vol. 2, No. 3, pp. 283-304, 1997.

[5] Michael K. Ng, Mark Junjie Li, Joshua Zhexue Huang, Zengyou He, "On the Impact of Dissimilarity Measure in k-Modes Clustering Algorithm", *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol.29, no. 3, pp. 503-507, March 2007, doi:10.1109/TPAMI.2007.53

[6] Milligan, G.W. and Copper, M.C. 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2):159-179.

[7] Bewley, Truman F. "Why Wages Don't Fall during a Recession." Cambridge, Ma: Harvard UP, 1999. Print.

[8] Judge, Timothy A., Christine L. Jackson, John C. Shaw, Brent A. Scott, and Bruce L. Rich. "Self-efficacy and Work-related Performance: The Integral Role of Individual Differences." Journal of Applied Psychology 92.1 (2007): 107-127. Web.