Machine Learning Engineer Nanodegree
Capstone Proposal

Isaac Robson
January 13, 2018

Proposal

Domain Background

Keywords:

Economics, Demographics, Unsupervised Machine Learning, Clustering, Preprocessing, PCA, Sparsity, Logistic Regression

Description:

One common form of unsupervised learning is clustering. As R. Tibshirani and G. Walther discuss in their 2005 paper, "Cluster Validation by Prediction Strength", 'predictionstrength.pdf' (original web address hyperlinked) there is no clear nor perfect way to evaluate the validity of a particular clustering. Methods range from silhouette score to the 'prediction strength' criterion described by Tibshirani and Walther, which reinterprets clustering as a quasi-supervised learning problem. In this project, we compare clusterability via silhouette score and 'prediction strength' but focus on an alternative metric via supervised learning of the clusters with an external dataset. The datasets analyzed relate to French demography, where INSEE (a government agency) has a 'baseline' classification of geo codes into one of four tiers (clusters for our purposes) on the urban-rural spectrum. This data is classified mostly via population count, a traditional metric as described by the US Census Bureau, the US Department of Agriculture, and others: 'what-is-rural.pdf' (original web address hyperlinked).

Problem Statement

City officials and urban planners frequently attempt to plan the landscape and policy of a region in conjunction with business interests in the area. For instance, officials submitting bids to Amazon's famous HQ2 would have to consider the effects of Amazon's presence on the city's revenue forecasts, infrastructure planning, school system, and many more dynamical challenges that are frequently subject to feedback loops. While solving these challenges is beyond the scope of this project, we do explore the interface between business and government in a very limited sense. By focusing on the relationship of company distribution and demographics, we can begin to understand - with new levels of granularity thanks to machine learning - how the workforces of companies shape the demography of a region. Our particular emphasis is the validity of INSEE's 4-tier urban-rural classification, which we believe may not best cluster the business and demographic meanings of urban and rural.

From an engineering perspective, we are primarily concerned with evaluation of clusterings, e.g. how can we determine if unsupervised clusters are valid? By examining the predictability of our clusterings with a disjoint (but correlated) dataset, we can train a supervised learning algorithm (i.e. logistic regression) to predict these clusters and validate our unsupervised results. This takes the quasi-supervised approach of Tibshirani and Walther (2005) a step further, but not beyond comparison. Doing this ties heavily to using unsupervised learning to label training data, but our primary concern here is to examine the efficacy of our clusters. Finally, we can explore the potency of data preprocessing for sparse data in the digital humanities. By simple A/B testing at each stage, we can look at success and failure in transformations and model selection techniques when dealing with sparsity and skew in data.

Datasets and Inputs

We rely on three datasets,

- 'base-cc-evol-struct-pop-2014.xls' - aka '**pop**'
- 'base_etablissement_par_tranche_effectif.csv' - aka '**firm**'
- 'AU2010 au 01-01-2017.xls' -aka **'tiers**'

All of the datasets relate to approximately n = 36000 geo codes (like zip codes in the US) across France, although some are excluded in each dataset, so we use SQL to select only shared geo codes.

'**firm**' has 10 data columns that each contain counts of businesses for each size, with columns that translate (roughly) to 'Unknown Size', 'Firms with 1-5 employees', 'Firms with 6-20 employees', 'Firms with 500+ employees' etc, corresponding to each geo code in France, and is available on Kaggle, as part of a larger demography effort (primarily focused on income/wealth): **https://www.kaggle.com/etiennelq/french-employment-by-town**

'**pop**' has 103 data columns, that range from population count to number of artisans and day laborers, although we are primarily concerned with just 8 population columns, e.g. 'Population between 0-14', 'Population between 30-44', and 'Population 90+', listed by geo code. This is taken directly from the INSEE site (where '**firm**' originates): https://www.insee.fr/fr/statistiques/2862200#consulter

'**tiers**' is a reference document that has the 'urban areas' each geo code belongs to and the urban-rural 4-tier classification given by INSEE for each 'urban area.' Note that each tier has city/suburb/periphery subtiers in the larger categories, but we ignore this. (we use a Vlookup-like technique to match the 'CATAEU2010' column in the second sheet, '*Composition_communale*' to each geo code, 100s being major city, 200s medium cities, 300s smaller cities, and 400s being rural): https://www.insee.fr/fr/information/2115011

Solution Statement

A solution to this project will preprocess the '**firm**' data with a combination of PCA, MaxAbsScaler, and the Box-Cox transformation, adjusted to sparsity as needed to create a number of clusterable datasets. These datasets should then be clustered with both Gaussian Mixture Models and K-means. This should create geo code clusters by frequencies of business size (e.g. geo codes of giant cities with lots of businesses vs small cities with few). We will next preprocess the '**pop**' data (with techniques like '**firm**') to create a supervised learning metric with logistic regression. Then, we will select a few contender clusterings with our supervised learning metric, which should serve as a proxy for silhouette score. Finally, we will use traditional metrics such as silhouette score and Tibshirani's and Walther's (2005) 'prediction strength' metric to evaluate whether our metric helped us find a 'superior' clustering compared to the benchmark. Thus we examine the efficacy of the supervised learning metric while still having more traditional metrics to rely on.

Benchmark Model

For a benchmark, we use INSEE's 4-tier classification on the urban-rural spectrum. These follow hard-coded traditional population count tiering for the most part, although some employment factors and shared borders with other communes do affect the classification. This means a supervised learning algorithm may easily pick up the hard-coded rules for classification and approach 100% accuracy. One can peruse the 'chap 2.pdf' file (an excerpt from the full report) to see the meanings of the tiers and

their subtiers, if any (just focus on the numbers if you don't speak French). Note that the silhouette score et al depends on the preprocessing, so the benchmark score will slightly vary for each variant dataset (e.g. unprocessed data has one benchmark score while PCA'd data will have another).

Evaluation Metrics

We should select 2-4 promising clusterings with our supervised learning prediction accuracy. At least one of these clusterings should have supervised prediction accuracies within 15% of the benchmark's supervised learning prediction accuracy with the un-preprocessed '**pop**' data (90-10 splits). We should then compare these 2-4 clusterings to the benchmark and at least one should achieve a higher silhouette score or 'prediction strength' score (as defined by Tibshira and Walther (2005)).

The 'prediction strength' score is a two-fold cross-validated 'training cluster' and 'testing cluster' procedure. Basically, we find centroids on 50% of the data (training) and use those centroids to assign test labels to the other 50% of the data. We then measure if the test labels are the same cluster as the clustering from just clustering the test data, i.e. do the test clusters match the clusters predicted by the training clusters. We can then compute the 'prediction strength' score as the minimum proportion of correctly assigned points for each test cluster (see equation (2.1) in the paper).

Project Design

The project pipeline should more or less follow:

1. Import
    a. Import necessary modules
    b. Import and display '**firm**' data
2. Explore '**firm**'
    a. Examine '**firm**' data for sparsity, outliers, etc
    b. Visualize '**firm**' data
3. Preprocess '**firm**'
    a. Select preprocessing techniques (sparsity in mind)
    b. Visualize, justify, and compare results of techniques
4. Cluster '**firm**'
    a. Experiment with different cluster sizes and algorithms
    b. Visualize the clusterings and discuss the effects of preprocessing
5. Explore/Preprocess '**dat**'
    a. Similar to steps 1-3 for '**firm**', in less detail
    b. Use slicing/SQL to ensure matching training data/labels
6. Supervised learning on '**dat**'
    a. Train/test split on '**dat**' with labels from '**firm**' clusters
    b. Use our supervised learning metric to select best clusterings
    c. Examine results and implications on validity of '**firm**' clusters
7. Conclusion/future work
    a. Test the selected clusterings from step 6 with silhouette score and 'prediction strength' vs the benchmark
    b. Discuss findings and any expansions to the project/new project ideas