Latent Class Analysis for Data Representation

COMP4441

March 9th , 2022

Hsing Yu Chen & Andrew J. Otis

Research Question

Can data from the Bertrand and Mullainathan (2004) study be broken down into distinct, unobservable classes through LCA?

Significance of Project

Proof of concept for the model of Latent Class Analysis

Demonstration of Probability based clustering

Modelling callback responses as latent classes

Data Set Description

Historical context:

The **data was collected in 2001 from employer ads** in Chicago and Boston as part of a larger study by Bertrand and Mullainthan in 2004; and is **regarding information of resume call-back**. A total of **4,780 fictious resumes were sent in response to a call from the employer**. These **resumes were generated with random traits** and **names** that either sounded Caucasian or African American that were **randomly assigned**.

Brief Summary:

- 4,780 employers responded to the fictitious resumes
- 27 variables (characteristics)

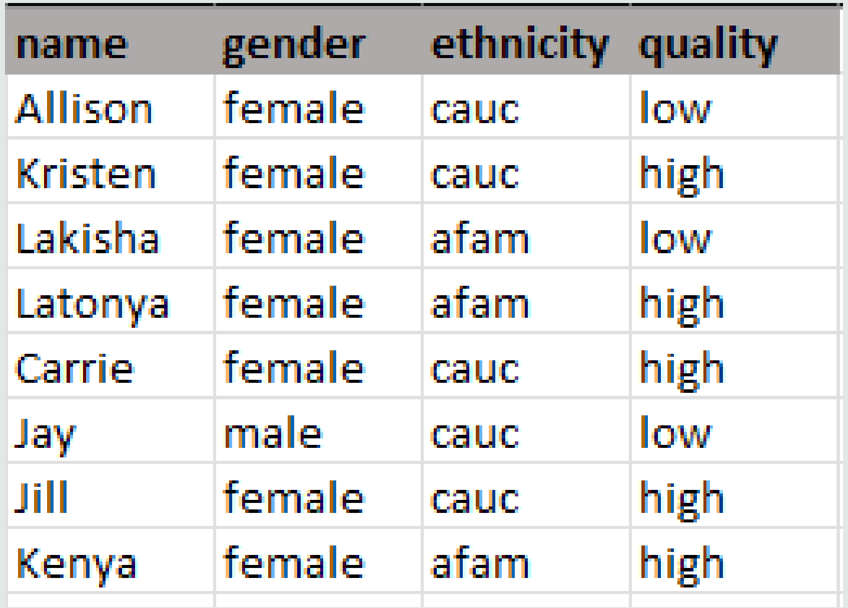
Data Prep

The data was already prepped and cleaned by researchers; however, it did require transforming column responses (items) into 1’s and 0’s for probabilistic algorithms to run.

Features desired of the data set are categorical variables where responses can be binomial (only 2 outcomes) or multimodal (more than 2 possible outcomes).

Data selection for our analysis was based on job requirements and the applicant’s gender and ethnicity, totaling in 7 items for Latent Class Analysis.

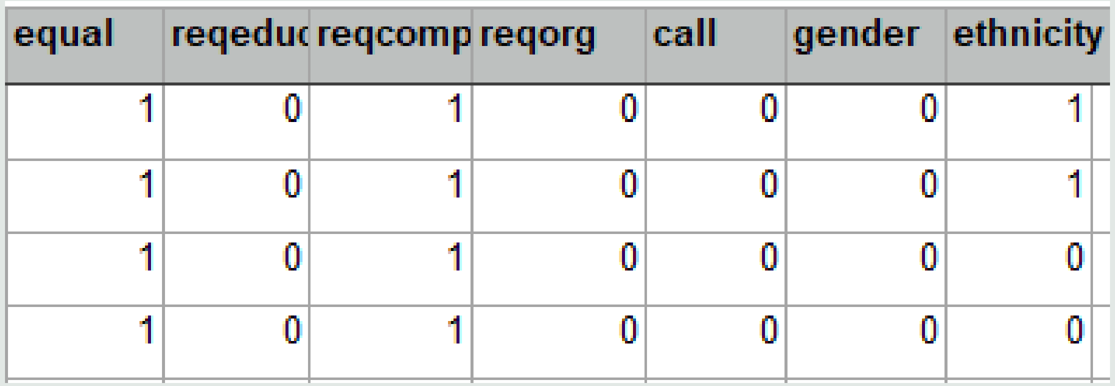
Data Visualization



name	gender	ethnicity	quality
Allison	female	cauc	low
Kristen	female	cauc	high
Lakisha	female	afam	low
Latonya	female	afam	high
Carrie	female	cauc	high
Jay	male	cauc	low
Jill	female	cauc	high
Kenya	female	afam	high

- 4,780 responses
- 27 variables (characteristics)
- Responses of items in original format

Data Transformation



equal	reqeduc	reqcomp	reqorg	call	gender	ethnicity
1	0	1	0	0	0	1
1	0	1	0	0	0	1
1	0	1	0	0	0	0
1	0	1	0	0	0	0

- 4,780 responses
- 7 variables (characteristics)
- Responses are binomial (1 or 0)

equal: *Ad has equal opportunity employment*
reqeduc: *Ad has education requirement*
reqcomp: *Ad has computer competency requirement*
reqorg: *Ad has organizational competency requirement*
call: *Was applicant called back?*
gender: *Gender of applicant*
ethnicity: *Ethnicity of applicant*

Latent Class Analysis

LCA is model in which **individuals** can be **classified into mutually exclusive and exhaustive events** known as **latent classes**, which are based on their pattern of response **measured with categorical variables**.

Model Assumptions

- Conditional Independence of variables
- True class membership is unknown

Maximum Likelihood using Expectation Maximization Algorithm

1. Begin with a random split of individuals into classes
2. Maximize the log-likelihood function
3. Reclassify individuals based on updated probabilities
4. Repeat steps 2 and 3 until the best Bayesian Interference Criterion is found

Latent Class Model Selection

Classes	G^2	X^2	df	AIC	BIC
2	390.6779	416.8358	112	32324.64	32422
3	162.2193	150.1037	104	32112.18	***32261.47***
4	109.3424	93.3258	96	32075.3	***32276.52***
5	78.7304	66.4020	88	32060.96	32313.83
6	57.4546	44.1142	80	32055.41	32360.48
7	42.92	33.2584	72	32056.88	32413.88

Notice that BIC begins increasing from 3 classes to 4 classes.

Therefore, choosing 2 or 3 classes would be best regarding model fit of the data set

Our group utilized the Bayesian Information Criterion (BIC) for estimating how well a given LCA model would fit our dataset.

$$BIC = kln(n) - 2\ln(\hat{L})$$

BIC prefers models where the number of samples (n), far outnumber the number of parameters (k) and in our case, a lower BIC value is indicative for the best fit model.

3-Class LCA Model Interpretation

	Class 1 (74.97%)	**Class 2 (5.91%)**	**Class 3 (19.12%)**
Item	**Probability of "Yes"**		
equal	73.61%	***0%***	76.86%
reqeduc	92.33%	***11.73%***	95.81%
reqcom	***37.02%***	55.62%	95.4%
reqorg	88.79%	100%	99.53%
call	92.04%	96.57%	91.02%
gender	**96.87%**	67.09%	***38.12%***
ethnicity	50.43%	50.5%	49.05%

Probability of an individual's membership in each class

Probability of a particular response given an individual's class

It appears that no matter what class an individual is in, they have higher than a 90% chance of progressing through the applicant process.

We can also observe that there exists other item differences that imply a distinction of classes

Class 1: Jobs that rarely require computer skills and consisting of mostly male applicants

Class 2: Jobs that don't offer equal opportunity, rarely have an education requirement

Class 3: Jobs whose applicants consisted of mostly female applicants

Method Validity

Due to a lack of company id being present in responses, we are unable to account for replicate samples. Responses for “call” do not meet the LCA requirement for conditional independence.

For example, one company having more than one ad or more than one resume being sent in response to the same ad.

Thus, LCA is NOT a valid model to be applied to the current data set.

Caveat: for the analysis to be carried out, our group assumed the responses were conditionally independent in order to demonstrate the process and interpretation of LCA

Conclusion

LCA is a model for finding unobservable classes of a sample using multivariate and categorical data. It differs from cluster analysis by being probability-based classification as opposed to distance-based grouping.

It’s a model that is widely used to represent data in sociological and mental health settings. Such as the classification of drinking groups, or types of voters, etc.

While we did not name our classes, doing so presents researchers with the unethical choice of purposefully giving names to classes that inaccurately represent class membership.

References

Research Papers

Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, **94**, 991–1013.

Journal Articles

Weller, B. E., Bowen, N. K., & Faubert, S. J. (2020). Latent Class Analysis: A Guide to Best Practice. *Journal of Black Psychology*, *46*(4), 287–311.

Websites

DataCamp. *poLCA function - RDocumentation*. RDocumentation. Retrieved March 3, 2022

P., B. (2014, October 31). *Latent Class Analysis vs. Cluster Analysis - differences in inferences?* Cross Validated. Retrieved March 7, 2022.

Petersen, K. J. (2019, May 29). *The Application of Latent Class Analysis for Investigating Population Child Mental Health: A Systematic Review*. Frontiers. Retrieved March 3, 2022.

Pratt, B. (2020, January 14). *Latent Class Analysis Using R*. Office of Population Research. Retrieved March 3, 2022.

Wikipedia contributors. (2022, January 18). *Bayesian information criterion*. Wikipedia. Retrieved March 5, 2022.