American Immigrants Classification and Naturalization Time Prediction of Different Groups

Yixiao Sheng, Yu-Chung Lien, Ching-Hua Wang

Problem

How many years it takes, for people with different race, education, gender, English speaking ability etc., to be granted their naturalizations? Our project focus on people who lives in California.

The US map shows the average length of time for immigrants to become American citizens over 5,000,000 samples

(Darkest: 13.5 years; lightest: 8.5 years)

Dataset

American Community Survey 2008-1013

Linear Regression

\[ h(x) = \sum \theta x = \theta^t x \]

\[ f(\theta) = \frac{1}{2} \sum \theta(x^{(i)} - y^{(i)})^2 \]

\[ \theta_j > \theta_l + \sum (y^{(i)} - \hat{h}(x^{(i)}))x^{(j)} \]

Features | Weights
--- | ---
Year of entry | 1.53
Age | 1.34
Wage income | 0.72
Disability | 0.64
Gender | 0.62

World Area of Birth | Weights
--- | ---
Born in Latin America | 0.5
Born in Asia | 0.4
Born in Europe | 0.3
Born in Northern America | 0.1
Born in the World Sea | 0.01

Educational attainment | Weights
--- | ---
Below college | -2.5
Associate's degree | -3.0
Bachelor's degree | -3.5
Master's degree | -4.0
Professional degree beyond a bachelor's degree | -4.4

Ability to speak English | Weights
--- | ---
Very well | 2.0
Well | 1.7
No at all | 2.3

Decision Regression

\[ R(Y | X) = \sum \beta X + \alpha | Y = x \]

\[ R(Y | X) = - \sum P(x | y) \log P(x | y) + y | x \]

\[ R(Y | X) = \alpha | Y = x \]

Features | Correlation Coefficient | Size of Tree
--- | --- | ---
Year of entry | 0.9417 | 2
Age | 0.977 | 2
Wage income | 0.977 | 2
Disability | 0.5871 | 2
Gender | 0.5607 | 2

World Area of Birth | Size of Tree
--- | ---
Born in Latin America | 1016
Born in Asia | 573
Born in Europe | 958
Born in Northern America | 36
Born in the World Sea | 36

Feature | Correlation Coefficient | Size of Tree
--- | --- | ---
Year of entry | 0.9417 | 2
Age | 0.977 | 2
Wage income | 0.977 | 2
Disability | 0.5871 | 2
Gender | 0.5607 | 2

The table highlights the features that contribute to faster naturalization.

Kernal Matrix Analysis & K-Means Cluster

Use sorted eigenvalue from Gaussian kernel matrix versus data points and distance versus data points to determine the number of groups.

Statistic Distribution

- The statistical distribution of Year of entry v.s. naturalization time.
- World area of birth versus naturalization time.

Conclusion

1. The clustering results indicate that people from Asia with higher degree need longer naturalization time. However, the linear regression shows generally, higher degree actually contribute to faster process. As world area of birth also plays a major role, we use regression tree to reveal more details.
2. The large weight of year of entry and world area of birth match well with the statistical distribution.
3. After using decision regression, the correlation coefficient improves from 0.56 to 0.59.

Reference

1. http://www2.census.gov/acs2013_3y5/puma/csv_pus.zip