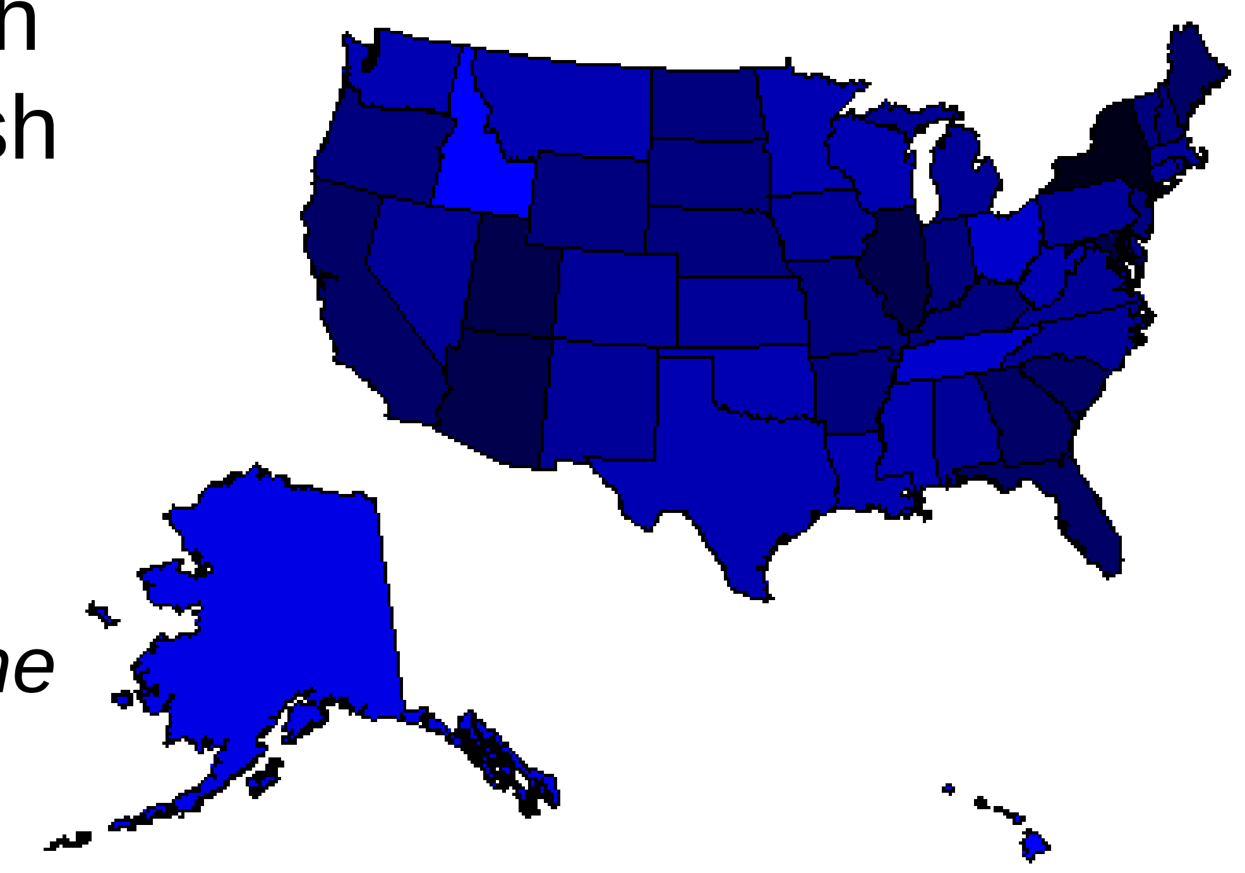
Stanford University

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.



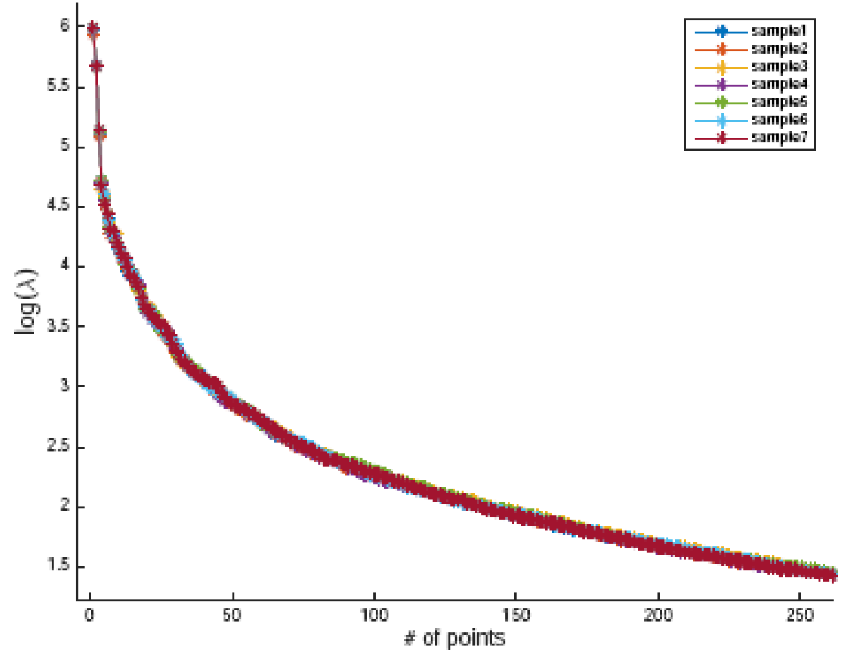
The US map shows the average length of time for immigrants to become American citizens over 5,000,000 samples

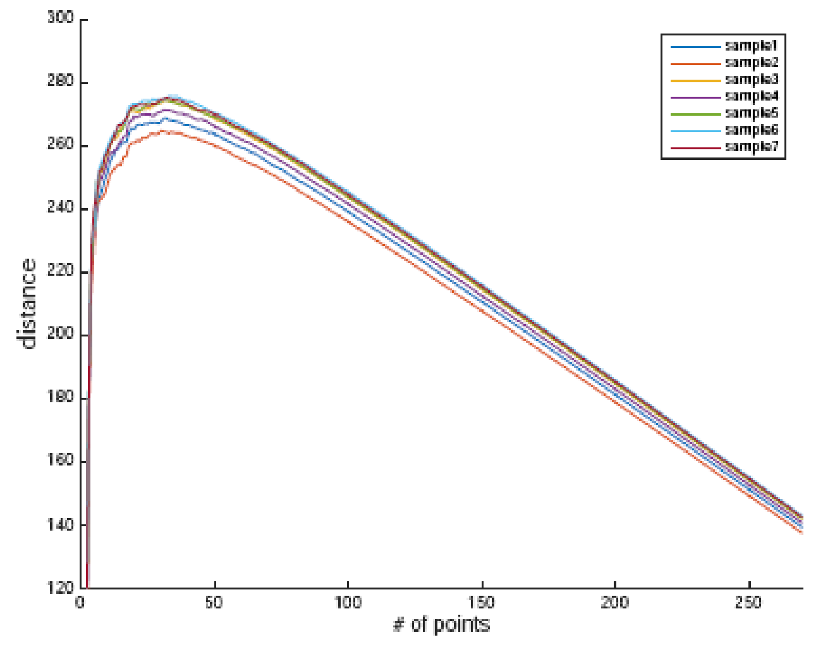
(Darkest: 13.5 years; lightest: 8.5 years)

Kernal Matrix Analysis & K-Means Cluster

$$c^{(i)} = \arg\min_{\theta} \left\| x^{(i)} - \mu_j \right\|^2$$

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$$





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

Group Number (Male)		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Wolrd Area of Birth	Latin America	v	v	v					v								
	Africa													v			
	Asia				v	v	v	v		v	v					v	v
Race	White			v								v			v		
	Black																
	Others		v														
Marital Status	Married		v	v			v	v	v		v	v		v	v	v	v
	Never Married	v				v				v							
	Separated				v								v				
English Speaking Ability	Good	v	v	v	v	v	v	v	v	v		v	v	v	v	v	v
	Not Good										v						
Class of Worker	Private Company	v	v	v		v			v	v	v					v	
	Government																
Education Attainment	High School or Lower		v	v					v	v	v	v					v
	Bachelor Degree						v									v	
	Master or Higher							v							v		
Age at time of Entry	Young	v				v				v					v		
Income	Low									v							
	High							v						v	v		
Naturalization	Long	v	v	v				v	v				v		v		

Group Number (Female)		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Wolrd Area of Birth	Latin American					v			v						
	Asian	v	v	v			v	v		v	v			v	v
Race	White				v										
	Black														
	Others														
Marital Status	Married	v		v		v	v	v					v	v	v
	Never Married		v								v				
	Separated									v		v			
Female have children under 17	With Children					v	v	v					v		v
	No Children														
English Speaking Ability	Good	v	v		v		v	v		v	v	v	v	v	
	Not Good														
Class of Worker	Private Company		v			v		v							v
	Government						v								
Education Attainment	High School or Lower		v	v	v	v			v			v			
	Bachelor Degree							v						v	
	Master or Higher	v													
Age at time of Entry	Young		v		v						v		v		
Income	Low		v			v			v						
Naturalization	Long				v	v			v			v	v		

Dataset

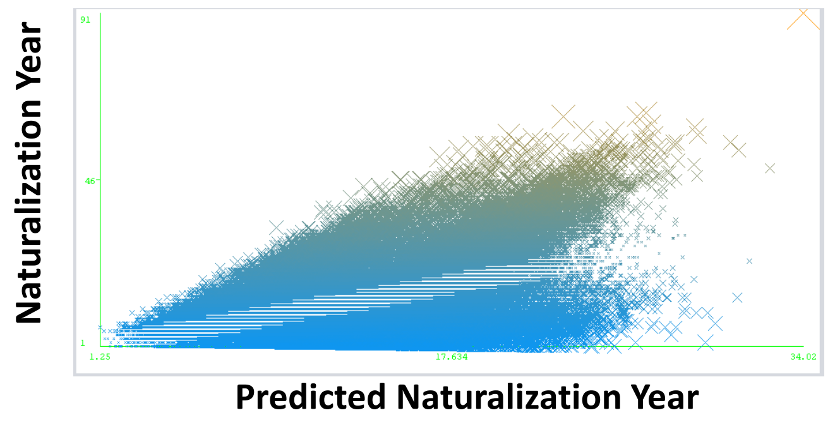
American Community Survey 2008-1013

Linear Regression

$$h(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x$$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} h_\theta(\mathrm{x}^{(i)}) - \mathrm{y}^{(i)})^2$$

$$\theta_j := \theta_j + \alpha \sum_{i=1}^{m} (\mathrm{y}^{(i)} - h_\theta(\mathrm{x}^{(i)})) x_j^{(i)}$$



The table highlights the features that contribute to faster naturalization.

Features	Weights
Year of entry	-30.3
Age	-4.2
Wage income	-0.4
Disability	0.4
Gender	0.2

World Area of Birth	Weights
Born in Latin America	1.4
Born in Asia	-2.9
Born in Europe	-2.9
Born in Africa	-2.2
Born in Northern America	-0.9
Oceania and at Sea	-1.6

Educational attainment	Weights
Below 12th grade - no diploma	-1.1
Below colloge	-2.5
Associate's degree	-3.3
Bachelor's degree	-3.0
Master's degree	-2.4
Professional degree beyond a bachelor's degree	-3.4
Doctorate degree	-1.6

Ability to speak English	Weights
Very well	0.1
Well	0.8
Not at all	2.3

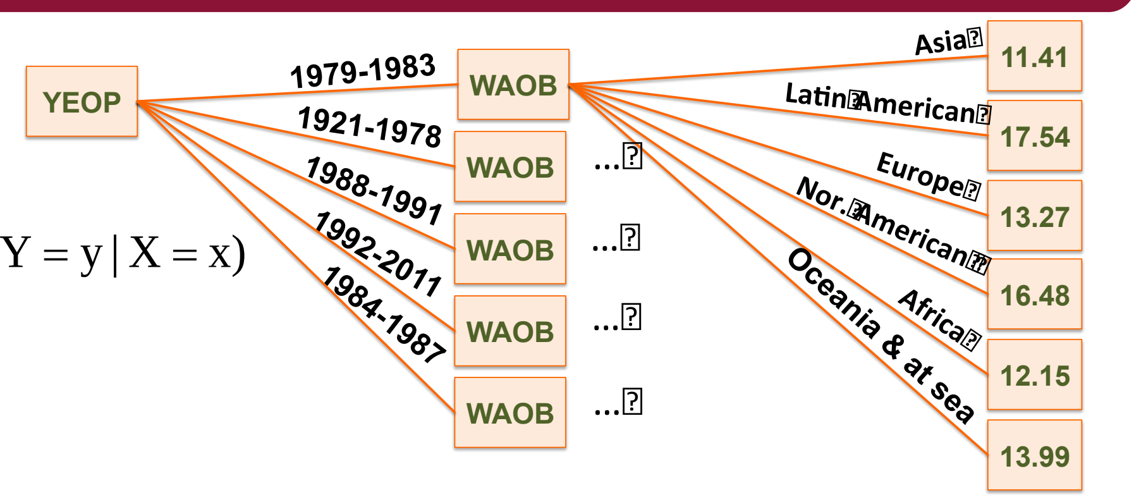
Decision Regression

$$H(Y|X) = \sum_{x} P(X = x) H(Y|X = x)$$

$$= -\sum_{x} P(X = x) \sum_{y} P(Y = y|X = x) \log_2 P(Y = y|X = x)$$

$$= -\sum_{x,y} P(X = x, Y = y) P(Y = y|X = x)$$

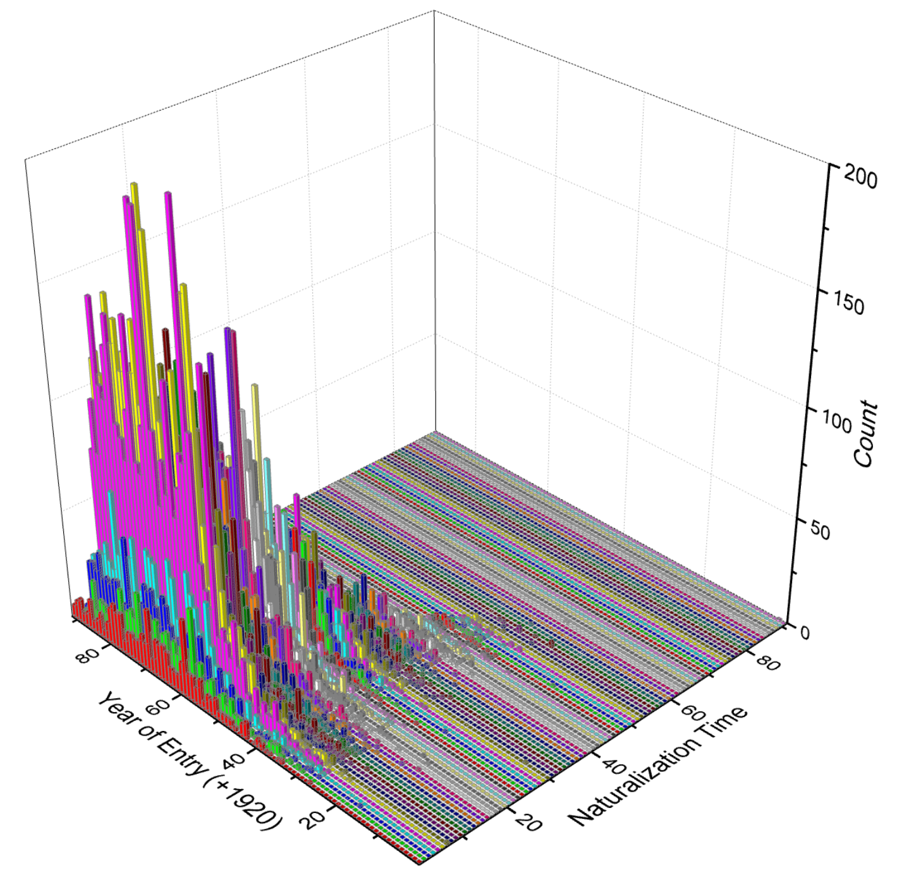
$$IG(X) = H(Y) - H(Y|X)$$



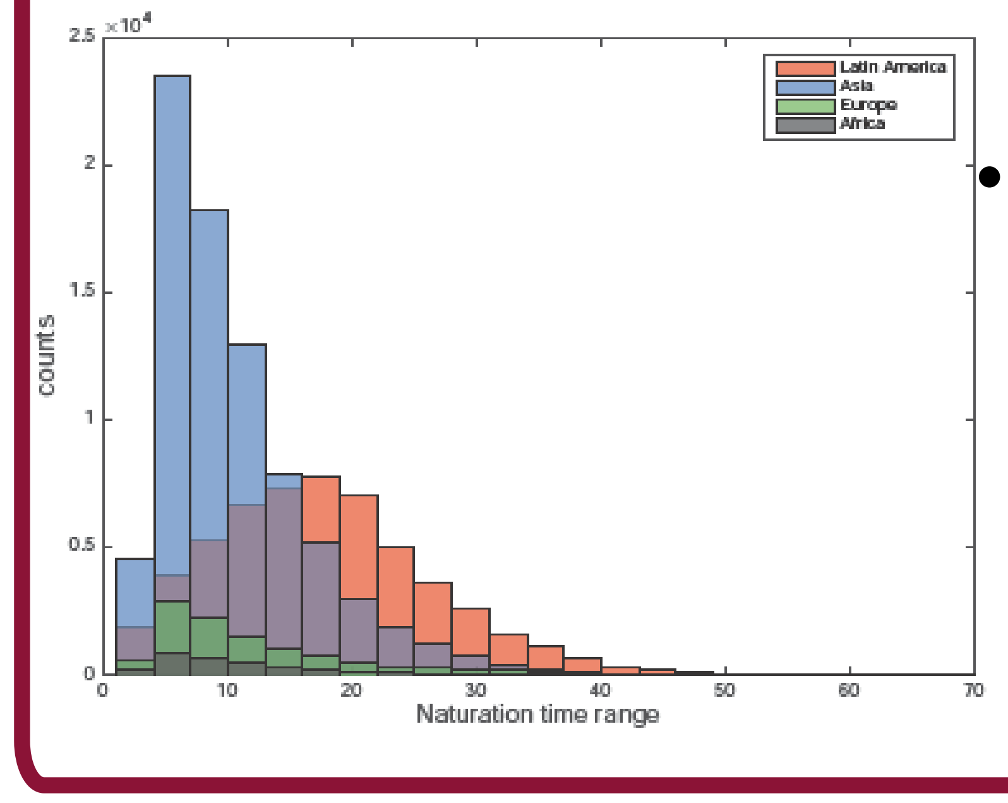
- Features selection optimization was implemented to achieve correlation 0.595.
- The important features match linear regression result.

Number of Features	Feature	Correlation Coefficient	Size of Tree
1	Year of Entry	0.4252	6
	Arrive Age	0.2652	8
	Education Level	0.2091	4
	World Area of Birth	0.4141	7
	English Level	0.0924	5
	Race	0.3513	7
2	Year of Entry World Area of Birth	0.5749	36
3	Year of Entry WAOB Education Level	0.5817	105
4	Year of Entry World Area of Birth Arrive Age Education Level	0.591	475
5	Year of Entry World Area of Birth Arrive Age Education Level English Ability	0.596	958

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 roll, 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_1yr/pums/csv_pus.zip
2. https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set
3. https://alliance.seas.upenn.edu/~cis520/wiki/index.php?n=Lectures.DecisionTrees