Thao Bach

Divorce Prediction using Machine Learning

1. Introduction and motivation

Divorce is a termination of marriage between two people that is a widespread social issue. Given that the "divorce rates have doubled over the past two decades among persons over age 35" (Kennedy), understanding the trends and characteristics that constitute divorce is important. Using machine learning models for classification, this project aims to predict the number of times a person gets married in a lifetime, given information about that person such as race, income, nativity, educational attainment, etc. The hope is by selecting features to predict the outcome of a relationship, one can understand how those features have an impact on relationships and provide further insight on the matter of divorce.

2. Method

i. Overview

Four separate supervised learning models were used to fit two different data sets: Logistic Regression, Multinomial Naive Bayes, and K-Nearest Neighbors were used to fit the canned data, and Naive Bayes was used to fit the canned data.

ii. Data

Table 1 is uncanned data, taken from the US Census Bureau, and it describes the marital statuses of adults of ages 15 and over had ever married, varied by socioeconomic factors and demographics (Lewis). Each cell in the table is the percentage of people of that background (i.e. white alone, native, less than high school, etc.) who have never married, been married once, etc.

Characteristics of People 15 Years Old and Over by Times Married: 2008–2012

(For information on confidentiality protection, sampling error, nonsampling error, and definitions, see *www.census.gov/acs/www/Downloads/data_documentation/Accuracy/MultiyearACSAccuracyofData2012.pdf*)

Characteristic	Total	Never married	Ever married		
			Married once	Married twice	Married three or more times
Total	**240,099,612**	**73,648,554**	**125,502,358**	**32,347,199**	**8,601,501**
Percent	**100.0**	**30.7**	**52.3**	**13.5**	**3.6**
RACE AND HISPANIC ORIGIN					
White alone	182,506,899	27.1	54.0	14.8	4.1
White alone, non-Hispanic	159,791,948	25.5	54.4	15.6	4.5
Black alone	28,326,972	46.6	40.5	10.8	2.1
American Indian and Alaska Native alone	1,843,308	40.4	42.6	12.6	4.4
Asian alone	11,870,498	29.8	63.0	6.5	0.7
Native Hawaiian and Other Pacific Islander alone	378,521	37.0	51.4	9.8	1.8
Some other race alone	10,532,222	42.2	49.3	7.4	1.1
Two or more races	4,641,192	47.0	39.4	10.4	3.2
Hispanic (of any race)	35,205,139	39.8	50.0	8.7	1.4
NATIVITY					
Native	202,873,086	31.8	50.0	14.2	4.0
Foreign born	37,226,526	24.7	64.5	9.5	1.3
EDUCATIONAL ATTAINMENT					
Less than high school	44,674,288	48.9	39.0	9.2	2.8
High school graduate	64,690,924	27.0	53.2	15.4	4.4
Some college or associate's degree	70,109,281	30.2	50.1	15.3	4.3
Bachelor's degree or more	60,625,119	21.7	63.5	12.4	2.4
EMPLOYMENT STATUS					
Employed	141,721,827	29.0	54.4	13.4	3.1
Unemployed	14,286,225	49.7	37.2	10.2	2.9
Not in labor force[1]	84,091,560	30.3	51.2	14.1	4.5
INCOME[2]					
Less than $25,000	128,034,730	41.4	44.2	11.1	3.3
$25,000 to $49,999	57,177,276	22.4	57.7	15.7	4.1
$50,000 to $74,999	27,829,688	16.6	62.6	16.8	4.0
$75,000 to $99,999	12,100,162	13.5	65.9	16.9	3.7
$100,000 and over	14,957,756	10.0	70.0	16.6	3.4
POVERTY STATUS					
Below poverty level	29,437,844	48.6	38.7	9.6	3.0
100–199 percent of poverty level	42,507,005	35.4	48.8	12.2	3.6
200–299 percent of poverty level	40,214,714	31.4	51.6	13.3	3.7
300+ percent of poverty level	127,940,049	24.8	56.7	14.8	3.7
PUBLIC ASSISTANCE					
Household receives public assistance[3]	38,251,401	42.3	42.1	11.7	3.9
TENURE					
Owns home	165,220,307	24.0	57.0	15.0	3.9
Rents home[4]	74,879,305	45.3	41.7	10.1	2.8
PRESENCE OF OWN CHILDREN UNDER 18 YEARS					
With own children[5]	62,437,130	11.0	72.6	14.0	2.3

Table 1

Table 2 is canned data, taken from UCI Machine Learning Repository (UCI Machine Learning Repository). It has individual-level data on people's age, type of employer, education, occupation, and other variables.

	A	B	C	D	E	F	G	H	I	J
1	age	workclass	education_num	marital_status	occupation	race	sex	hours_per_week	native_country	income
2	39	State-gov	13	Never-married	Adm-clerical	White	Male	40	United-States	<=50K
3	50	elf-emp-not-i	13	Married-civ-spouse	Exec-managerial	White	Male	13	United-States	<=50K
4	38	Private	9	Divorced	Handlers-cleaners	White	Male	40	United-States	<=50K
5	53	Private	7	Married-civ-spouse	Handlers-cleaners	Black	Male	40	United-States	<=50K
6	28	Private	13	Married-civ-spouse	Prof-specialty	Black	Female	40	Cuba	<=50K
7	37	Private	14	Married-civ-spouse	Exec-managerial	White	Female	40	United-States	<=50K
8	49	Private	5	Married-spouse-abse	Other-service	Black	Female	16	Jamaica	<=50K
9	52	elf-emp-not-i	9	Married-civ-spouse	Exec-managerial	White	Male	45	United-States	>50K
10	31	Private	14	Never-married	Prof-specialty	White	Female	50	United-States	>50K
11	42	Private	13	Married-civ-spouse	Exec-managerial	White	Male	40	United-States	>50K
12	37	Private	10	Married-civ-spouse	Exec-managerial	Black	Male	80	United-States	>50K
13	30	State-gov	13	Married-civ-spouse	Prof-specialty	ian-Pac-Islan	Male	40	India	>50K
14	23	Private	13	Never-married	Adm-clerical	White	Female	30	United-States	<=50K
15	32	Private	12	Never-married	Sales	Black	Male	50	United-States	<=50K
16	40	Private	11	Married-civ-spouse	Craft-repair	ian-Pac-Islan	Male	40	?	>50K
17	34	Private	4	Married-civ-spouse	Transport-moving	er-Indian-Esk	Male	45	Mexico	<=50K
18	25	elf-emp-not-i	9	Never-married	Farming-fishing	White	Male	35	United-States	<=50K
19	32	Private	9	Never-married	Machine-op-inspct	White	Male	40	United-States	<=50K
20	38	Private	7	Married-civ-spouse	Sales	White	Male	50	United-States	<=50K
21	43	elf-emp-not-i	14	Divorced	Exec-managerial	White	Female	45	United-States	>50K
22	40	Private	16	Married-civ-spouse	Prof-specialty	White	Male	60	United-States	>50K
23	54	Private	9	Separated	Other-service	Black	Female	20	United-States	<=50K
24	35	Federal-gov	5	Married-civ-spouse	Farming-fishing	Black	Male	40	United-States	<=50K
25	43	Private	7	Married-civ-spouse	Transport-moving	White	Male	40	United-States	<=50K
26	59	Private	9	Divorced	Tech-support	White	Female	40	United-States	<=50K
27	56	Local-gov	13	Married-civ-spouse	Tech-support	White	Male	40	United-States	>50K
28	19	Private	9	Never-married	Craft-repair	White	Male	40	United-States	<=50K
29	54	?	10	Married-civ-spouse	?	ian-Pac-Islan	Male	60	South	>50K
30	39	Private	9	Divorced	Exec-managerial	White	Male	80	United-States	<=50K
31	49	Private	9	Married-civ-spouse	Craft-repair	White	Male	40	United-States	<=50K
32	23	Local-gov	12	Never-married	Protective-serv	White	Male	52	United-States	<=50K
33	20	Private	10	Never-married	Sales	Black	Male	44	United-States	<=50K
34	45	Private	13	Divorced	Exec-managerial	White	Male	40	United-States	<=50K
35	30	Federal-gov	10	Married-civ-spouse	Adm-clerical	White	Male	40	United-States	<=50K
36	22	State-gov	10	Married-civ-spouse	Other-service	Black	Male	15	United-States	<=50K
37	48	Private	7	Never-married	Machine-op-inspct	White	Male	40	Puerto-Rico	<=50K

adult.data

Table 2

iii. Preprocessing and Feature Selection

Uncanned data

Because the US Census Bureau report did provide the table in csv form, I copied and pasted all the values in the table onto a separate spreadsheet and then further partitioned the table into separate spreadsheets based on selected feature to parse and store them in feature vectors more conveniently.

Each feature is stored into a pandas dataframe, where each row corresponds with a marital status: row 0 denotes never married, row 1 denotes married once, row 2 denotes married twice, and row 3 denotes married three more more times. Each cell is a joint probability between a feature and a marital status with respect to the row number. To query for joint probabilities, a dictionary is created to map abbreviations of a feature to the feature name that is stored in the dataframe.

In terms of feature selection, I chose all, except “poverty status,” “public assistance,” and “presence of own children under 18 years” as features because they are less relevant in determining the marital status of people.

Canned data:

For canned data from UCI , the following changes to the original data:

1. Removal of the following columns: fnlwgt, education, relationship, capital gain, and capital loss. They were removed because they are either already represented by another column or extraneous.
2. Added labels to each column to make querying and interpreting data more understandable.
3. Removal of rows that contain cells with incomplete information, represented by ‘?’
4. One-hot encoding was applied on features to turn categorical variables such as type of work (“workclass”), marital status, occupation, race, sex, native country, and income (originally represented as either >50k or <=50k)

v. Training

Uncanned data

Because the data came in the form of a contingency table instead of individual-level data, it was not possible to split them into training and test sets. Therefore, the entire table was used as training data.

<u>Canned data</u>

Using scikit-learn's *train_test_split* method, I split the data into a 50:50 ratio, where 50% was used for training and the other 50% for testing.

vi. Implementation of algorithm

<u>Uncanned data</u>

Instead of using scikit-learn, I implemented Naive Bayes algorithm from scratch because the data format was incompatible to pass to scikit-learn's methods. To do so, I wrote three methods to calculate the joint probability between two events, the conditional probability between two events, and a *predict* method that calls the previous two helper methods. In order to make a prediction, the *predict* method accepts string abbreviations of a person's characteristics and prints out four probabilities in ascending order, corresponding to the likelihood of each marriage outcome based on the input parameters.

<u>Canned data</u>

I used scikit-learn's library to fit Logistic Regression, Multinomial Naive Bayes, and K-Nearest Neighbors models to the canned data and evaluate their accuracy.

3. Results

<u>Uncanned data</u>

Because the contingency table was used entirely for training and not testing, manual construction of full feature vectors was required to evaluate whether the model makes sense. Four tests were written to do so, and each test represents people with distinct features, one of whom is based on a real person. For each test, percentages were given to the likelihood of each marital status.

Below are descriptions of people represented in the results:

1. Person 1: 2 or more race, native, less than high school, employed, income over $100k, and rents home
2. Person 2: white alone, native, bachelor's degree, employed, income over $100k, and owns home (Hugh Hefner)
3. Person 3: hispanic (of any race), foreign born, less than high school, unemployed, income less than $25k, and rents home
4. Person 4: American Indian and Alaska Native alone, native, high school graduate, and income ranges from $75,000 to $99,999

Person	Never Married	Married Once	Married Twice	Married 3+
1	**46.98%**	41.6%	9.08%	2.33%
2	3.96%	**78.84%**	14.99%	2.22%
3	**87.16%**	11.69%	1.03%	0.12%
4	17.41%	**56.16%**	20.36%	6.07%

Canned data

	Train Accuracy	Test Accuracy
Logistic Regression	69%	68%
Multinomial Naive Bayes	64%	64%
K-Nearest Neighbors	72%	63%

4. Discussion and conclusion

Uncanned data

The most likely marital outcomes were "never married" and "married once." Looking at the original data in Table 1, this makes sense because the majority of people are either married once or never married, so it's less likely to randomly create a person who is married twice or three or more times. Person 2, who represents Hugh Hefner, was predicted to have a 78.84% likelihood of being married once. In reality, he has been married *three* times, which was forecasted as 2.22% chance by the model, the *least likely of all the marital outcomes.*

Canned data

The highest train accuracy was achieved by K-Nearest Neighbors with 72%, while Logistic Regression gave the highest test accuracy with 68%. However, all accuracies are within 8% margin of each other. To improve accuracy, I tried to split the data into training and testing in different proportions, but accuracy for each model maintained relatively unchanged, within 1-2% between each tune.

Comparison of models used on uncanned and canned data

In terms of providing the best insight from predicting divorce, each model has its own advantages and disadvantages. The advantage of the model used on uncanned data is its ability to capture the nuances of divorce by calculating the likelihood of four marriage outcomes as opposed to whether a person gets divorced. The disadvantage of the model is the challenge in measuring its accuracy. Individual-level data with similar features to ones provided in the contingency would be required to evaluate how well the model predicts.

On the flip side, what the model on uncanned data lacks the models on canned data do better because accuracy was easily attained because the canned data was compatible to being split into training and testing data. The downside of these models is accuracy is fairly low at high 60s and low 70s.

In conclusion, predicting divorce is a seemingly straightforward process at first glance but requires more interpretation upon closer inspection because measuring divorce is, itself, not a straightforward task. A big part of this challenge is there are numerous ways to measure divorce. For instance, “not all states report divorce statistics,” and that they are counted “based on the total population, not the total married population” (U.S. Divorce Rates and Statistics). This can be problematic because populations fluctuate, and it may be the case that there were fewer people when number of divorces was collected, which creates the illusion that the divorce rate is higher than in reality.

5. References

Kennedy, Sheela, and Steven Ruggles. "Breaking Up Is Hard to Count: The Rise of Divorce in the United States, 1980–2010." *Demography* 51.2 (2014): 587-98. Web.

Lewis, Jamie M., and Rose M. Kreider. *Cohabitation, Marriage, Divorce, and Remarriage in the United States*. Hyattsville, MD: Dept. of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, 2002. *Remarriage in the United States*. American Community Survey Reports, Mar. 2015. Web.

"UCI Machine Learning Repository: Adult Data Set." *UCI Machine Learning Repository: Adult Data Set*. N.p., n.d. Web. 18 Dec. 2016.

"U.S. Divorce Rates and Statistics - Divorce Stats - Divorce Source." *U.S. Divorce Rates and Statistics - Divorce Stats - Divorce Source*. N.p., n.d. Web. 19 Dec. 2016. <http://www.divorcesource.com/ds/main/u-s-divorce-rates-and-statistics-1037.shtml>.