Predicting How New Yorkers Seek Medical Advice

Michelle Ho and Shay Lehmann

CUSP-GX-5006
Machine Learning Assignment # 2

Abstract

In this assignment, the authors explore how New Yorkers seek medical advice and how this can be predicted by other variables found in the NYC Community Health Survey. The exercise demonstrates that the use of Naive-Bayes and Support Vector Machine (SVM) classification techniques and their in performance in a real life scenario.

1 Introduction

The main goal of this assignment is to demonstrate the use of two classification techniques, Naive-Bayes and Support Vector Machine (SVM), on predicting how New Yorkers seek medical advice.

The authors examined the results of the Community Health Survey (CHS) conducted by the New York City Department of Health and Mental Hygiene. This rich dataset includes measurements on multiple aspects of health– including access, nutrition, demographic information, diet and exercise habits, medical history, and lifestyle choices.

The motivation is to better understand how New Yorkers seek and receive advice for their medical needs. Outreach programs hoping to improve access to health care can make use of these results.

2 Methods and Data Sets

The raw dataset of the 2014 NYC Community Health Survey contains 8562 observations across 188 variables. For this assignment, a subset of the raw dataset was used for analysis, and the variables are described below:

- Sick Advice: A categorical variable about the respondent's usual resource for health advice. This is the dependent variable in the analyses for this assignment. The options are “A private doctor”, “Community health center”, “A hospital outpatient clinic”, “ED/urgent care center”, “Alternative health care provider”, “Family/friend/self/resources”, “Non-hospital clinic”, “Other”, “No usual place”, or “Clinic, unknown type”.

- Education: A categorical variable on educational attainment. Categories are “Less than HS”, “High school grad”, “Some college”, or “College graduate”

- Marital status: A categorical variable on marital status. Categories are “Married”, “Divorced”, “Widowed”, “Separated”, “Never married”, or “Member of unmarried couple living together”

- US born: A binary variable answering the yes or no question ”Are you US or foreign born?”

- Sexual ID: A categorical variable on sexual orientation. Categories are “Heterosexual”, “Gay/Lesbian”, or “Bisexual”

- At Home Language: A categorical variable answering the question “What language do you speak most often at home?” Options are “English”, “Spanish”, “Russian”, “Chinese”, “Indian”, or “Other”

- Insured: A binary variable answering the yes or no question “Do you have any kind of health insurance coverage, including private health insurance, prepaid plans such as H-M-Os, or government plans such as Medicare or Medicaid?”

- The dependent variable being classified and predicted is ‘Sick Advice’ in this assignment’s analyses. The authors refer to this variable as ‘Y’. The goal is to attempt to predict how people will seek their medical advice based on other characteristics.

- The independent variables are all the others.

- For all variables, the options “do not know” and “refuse” were treated as null values and dropped. After these drops, the number of observations in the dataset is 7827.

- Finally, independent variables are binarized so that all samples are represented by boolean feature vectors.

The steps taken for this assignment:

1. A classification model is fitted for the binarized 6 independent variables with a Naive Bayes algorithm assuming Bernoulli distributions.

2. A second classification model is fitted on the same variables with a SVM using a linear kernel.

3. Finally, a third classification model is fitted with a SVM using a radial basis function (RBF) kernel.

4. The authors adjust the RBF kernel parameters to assess how these parameters affects the overall performance of the classification model and possible overfitting.

5. All models are assessed via cross validation on training and test sets split from the original dataset. The training and test set sizes are adjusted to assess the effect of training size on quality of classification.

6. A confusion matrix is created for all categories of the dependent variable “Sick Advice” to assess the performance of the classifications.

3 Results

It appears

4 Conclusions

In summary, the classification models performed poorly in predicting how people tend to seek out medical advice based on other characteristics.

Further steps for this assignment would include earlier years of data from the NYC Community Health Survey to expand the dataset, since we have seen that training size of the dataset can have an affect on classification models.

Python code used to generate the results, tables, and figures for this assignment can be found at `https://github.com/michellemho/machine_learning_for_cities`.

Table 1: Results of Classification Models

	Normalized raw data (LM1)		PCA-featured data (LM2)
constant	0.40926343	constant	0.4101
Neighborhood	-0.12254491	x1	-0.0379
BldClassif	0.02690248	x2	-0.0113
YearBuilt	0.02695934	x3	-0.0441
GrossSqFt	0.13819696	x4	-0.1773
GrossIncomeSqFt	0.01101943	x5	-0.0273
R-squared	0.001	R-squared	0.001

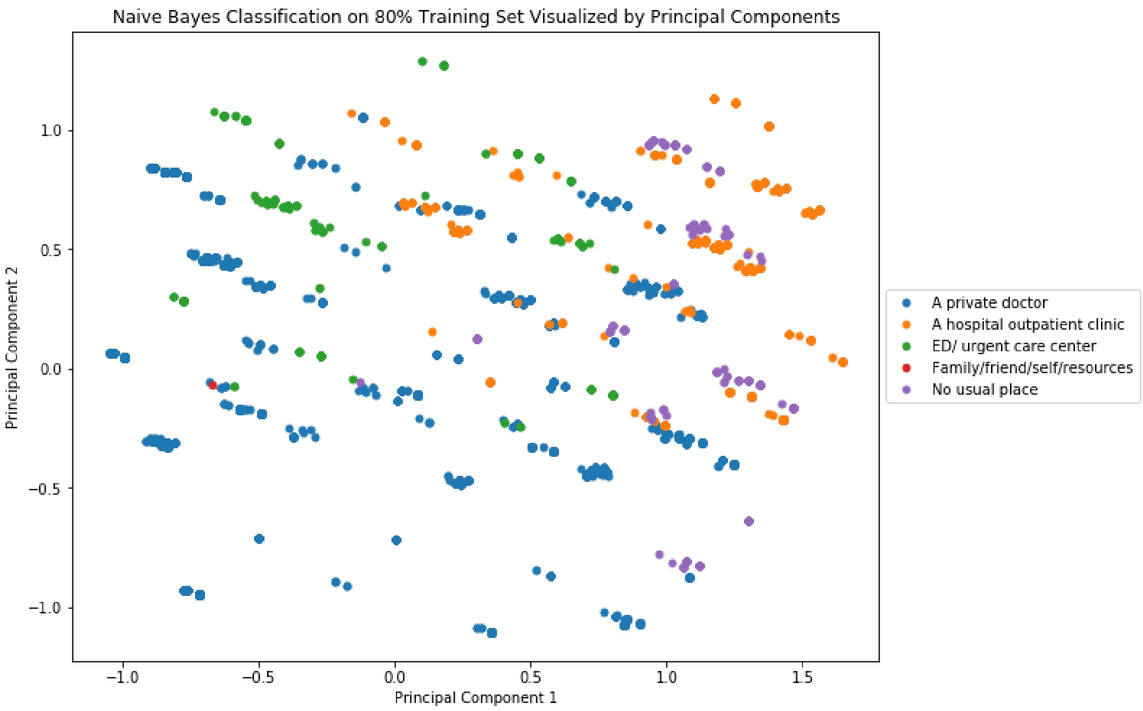

Figure 1: The results of Naive Bayes Classifier