Naive Bayes and SVM: A Comparative Study

CUSP-GX-5006 Machine Learning for Cities
Dror Ayalon, Xinge Zhong
March 3rd, 2017

Abstract

Using three different classification algorithms: Naive Bayes, SVM Linear, SVM Kernel, we choose five building features to predict its neighborhood and compare the corresponding classification results by both checking the cross-validation results and error rates.

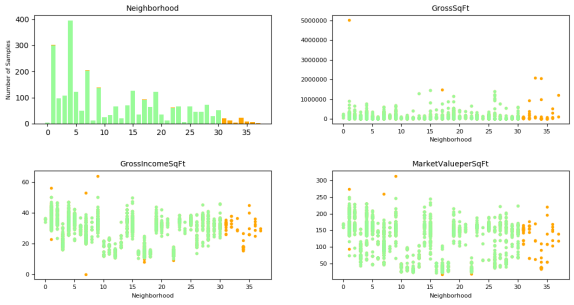
Dataset

We used the "manhattan-dof.csv" as our dataset.

Remove the outliers

For each five independent variables, we remove their outliers correspondingly, both extreme values and also especially, the zero value. One consideration is to avoid the "Zero Frequency" in Naive Bayes, the other consideration is based on Tobler's first law, near thing is more related than distant thing. When a feature shows itself distinctively, it would disturb the model even though it is belonged to that neighborhood.

For dependent variable, the neighborhoods, we also remove those neighborhoods whose number is greater than 30, because they got too small sample.



Whiten the data

We used scipy whiten method to standardize the data, which numerically is to (x-mean())/standard deviation.

This step is to make our predictors seem like come from a Gaussian distribution, so as to better fit the acquirements of these classification algorithms.

PCA the data

After the whiten step, we tried to use PCA method to reduce the dimension of data. After observing the explanation ratio rate, we conclude that PCA can't significantly filter the noise, it may be better to leave the raw data for classification.

Cross Validation

We used sklearn train_test_split method to divide the data into two parts: training data(80%), test data(20%).

Classification Algorithm

Three different classification algorithms were used for making classification to this real estate dataset. They are: Naive Bayes, SVM Linear Kernel, SVM RBF Kernel.

Besides this, in order to find the best cluster pattern of the neighborhoods, we also explore the classification algorithms by attempting to classify the predictor into 30, 10, 3 classes. Then they all got different results.

Naive Bayes

Naive Bayes, an easy-performed multi-class algorithms, works better in categorical variables, asks for the independence among the predictors, performs better at larger-scale dataset.

Our dataset is a mixed combination of categorical and numerical variables, besides, these predictors are not independent from each other. Thus we guess that the Naive Bayes' results would be the comparatively worse.

SVM Linear Kernel & SVM RBF Kernel

Support Vetor Machine classify things by creating a hyperplane which tries to maximize the distance from the data to the separate line.

This calculation method can be messed in our dataset, because we know that the neighborhoods number's

closeness also stands for its spatial closeness, which means that when two number is closer, the two corresponding neighborhoods are also closer.

This is also why when the classification groups become smaller, for example, from 30 groups to 10 groups, the SVM results have much smaller error rates.

Results

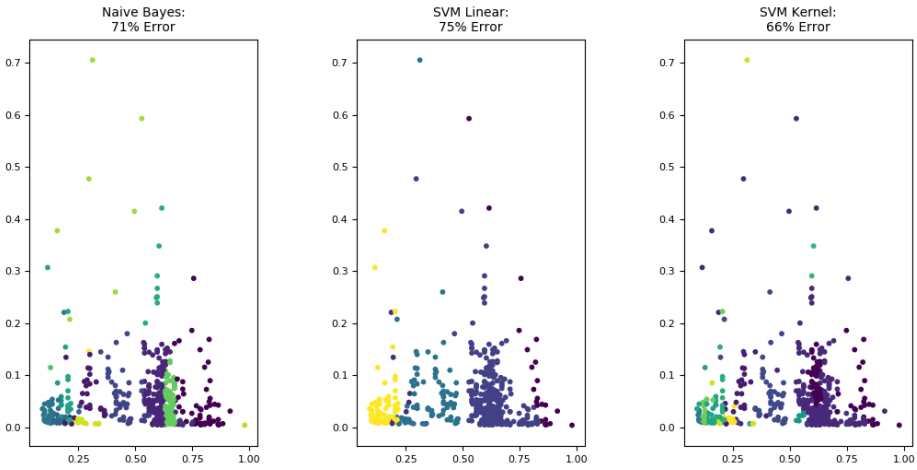
This is the cross-validation error results of three classification algorithms with three cluster numbers.

	30 Groups	10 Groups	3 Groups
Naive Bayes	71%	55%	14%
SVM Linear	75%	58%	13%
SVM RBF	66%	42%	15%

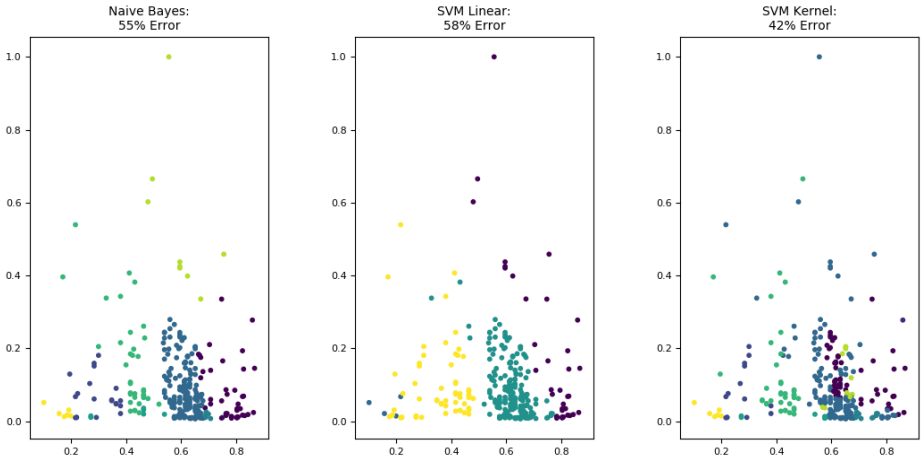
Figure 1: Error Rates of three classification

The following three plots are classification results of 3 groups, 10 groups and 30 groups.

Classification into 30 groups



Classification into 10 groups



Classification into 3 groups