Demographics and Crime Rate

This project works with the Communities and Crime data sourced from the UCI Machine Learning Repository. Link:

http://archive.ics.uci.edu/ml/datasets/communities+and+crime

The approach of this project is to be a learning exercise to apply a limited set of Machine Learning algorithms and techniques and study their impact. The work showcased in the python notebook tries to answer the questions framed below.

You have been given some data for per-capita crime rates around the country. Your task is to build models to predict the crime rate based on demographic and economic information about the particular locality. The data is given in the files "communities-crime-clean.csv" and "communities-crime-full.csv"; a description of the data and data fields is given in "communities-crime.names". The "full" dataset includes data fields with missing values (indicated by "?"), while in the "clean" set these fields have been removed.

1. **Decision Trees**

In this problem, you will use the clean dataset to predict whether the crime rate in a locality is greater than 0.1 per capita or not.

a. Create a new field "highCrime" which is true if the crime rate per capita (ViolentCrimesPerPop) is greater than 0.1, and false otherwise. What are the percentage of positive and negative instances in the dataset?
b. Use **DecisionTreeClassifier** to learn a decision tree to predict highCrime on the entire dataset. Remember to exclude the crime rate feature (ViolentCrimesPerPop) from the input feature set so you are not cheating.
    i. What are the training accuracy, precision, and recall for this tree?
    ii. What are the main features used for classification? Can you explain why they make sense (or not)?
c. Now apply cross-validation (**cross_val_score**) to do 10-fold cross-validation to estimate the out-of-training accuracy of decision tree learning for this task.
    i. What are the 10-fold cross-validation accuracy, precision, and recall?
    ii. Why are they different from the results in the previous test?

2. Linear Classification

a. Use **GaussianNB** to learn a Naive Bayes classifier to predict highCrime.
    i. What is the 10-fold cross-validation accuracy, precision, and recall for this method?
    ii. What are the 10 most predictive features? This can be measured by the normalized absolute difference of means for the feature between the two classes:

$$\frac{|\mu_T - \mu_F|}{\sigma_T + \sigma_F}$$

The larger this different, the more predictive the feature. Why do these make sense (or not)?

iii. How do these results compare with your results from decision trees, above?

b. Use **LinearSVC** to learn a linear Support Vector Machine model to predict highCrime.

i. What is the 10-fold cross-validation accuracy, precision, and recall for this method?

ii. What are the 10 most predictive features? This can be measured by the absolute feature weights in the model. Why do these make sense (or not)?

iii. How do these results compare with your results from decision trees, above?

3. Regression

Now you will attempt to directly predict the crime rate from the given features.

a. Use **LinearRegression** to learn a linear model directly predicting the crime rate per capita (ViolentCrimesPerPop).

i. Using 10-fold cross-validation, what is the estimated mean-squared-error (MSE) of the model?

ii. What is the MSE on the training set (train on all the data then test on it all)?

iii. What features are most predictive of a high crime rate? A low crime rate?

b. Now use **Ridge** regression to reduce the amount of overfitting, using **RidgeCV** to pick the best alpha from among (10, 1, 0.1, 0.01, and 0.001).

i. What is the estimated MSE of the model under 10-fold CV?

ii. What is the MSE on the training set (train on all the data then test on it all)?

iii. What is the best alpha?

iv. What does this say about the amount of overfitting in linear regression for this problem?

c. Now use polynomial features to do quadratic (second-order) polynomial regression.

i. What is the estimated MSE of the model under 10-fold CV?

ii. What is the MSE on the training set (train on all the data then test on it all)?

iii. Does this mean the quadratic model is better than the linear model for this problem?

4. Dirty Data

Repeat the decision tree learning question for the full (non-clean) data set and present the results.

a. Are the CV results better or worse? What does this say about the effect of missing values?