CS 584 Machine Learning

Homework Project: Crime Prediction

You have been hired by the FBI to develop predictive models for crime, to help the Bureau and police departments around the country to use machine learning to better focus their resources on locations where crimes are more likely to be committed. You are expected to implement your solution and submit it along with your answers to the following questions.

You should use Python 3.6 as interpreter with Scikit-learn 0.18.1. Develop your code in a Jupyter notebook and give your answers and explanations in the same notebook, along with your code (That is, code snippet, showing results followed by explanation, then another code snippet etc.). Please avoid unnecessary discussion, redundant or irrelevant results, and repeated code. Note that a Jupyter notebook can be very neat and helpful, but it also can get easily messy. You are encouraged to use latex for mathematical formulas. Also please get familiar with mark down and use it properly.

Submit a single notebook file named "*yourlastname-yourCWID*-homework1.ipynb".

You may do this homework in teams of 2 or 3 – if you are working in a team, the base requirement for this assignment is higher. See the end of the assignment for explanations.

Demographics and Crime Rate

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 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?

5. Teams

a. If you are working in a team of two people:
   i. Experiment with two learning methods other than those described above (one can be a non-linear kernel for SVM) for the classification problem, explaining clearly what you did. Show CV results for both the clean and full datasets.
   ii. What method gives the best results?
   iii. What feature(s) seem to be most consistently predictive of high crime rates? How reliable is this conclusion?

b. If you are working in a team of three people:
   i. Do the requirement for two person teams.
   ii. Devise a method to find the most useful threshold for dividing high crime areas from low crime areas (i.e., discretizing XXX to compute highCrime). Define clearly what you mean by "useful".
   iii. Show CV results for both the clean and full datasets for at least three different classification methods above.
   iv. How are these results similar and different from the previous results (with a fixed threshold of 0.1). What does this say about how to approach such a problem.

6. Extra Credit

a. Do a team requirement (above) that your team is not already required to do.
b. Experiment with other learning methods such as polynomial or other kernels in SVM, decision forests, boosting, etc. and show your results. Make sure to explain clearly what you did.
    i. What method gives the best results?
    ii. What feature(s) seem to be most consistently predictive of high crime rates? How reliable is this conclusion?
c. Find other data sets to combine with this data set that might improve results. This may include weather data (from NOAA), other demographics data from the US Census, etc. One good source for data sets is http://data.gov.
    i. Explain precisely what you did to combine the datasets.
    ii. Give accuracy, precision, and recall results with and without the new data.
    iii. Does the added data help? What features help? Does it matter what learning method you use?

Make sure to note clearly what portion of your notebook is meant to be extra credit material.