Term Project: Crime forecasting

Assigned Date: Tuesday, March 1, 2011

Educational Goal

Apply data mining techniques to real-world crime data.

Team

Team 1	Team 2
Ward, Max W Yu, Chung-Hsien Das, Priyank	Zhu, Bingchun Phan, Dung Huy Le, Hie

Phase I: Prototype for crime prediction

(200 points)

Step 1 and Step 2 are due 5:30 PM Thursday, March10, 2011

Step 3 is due 5:30 PM Tuesday, March 22, 2011

Requirements

- **Dataset**: The datasets contains 6 crime types and each crime is presented in monthly base in CSV (Comma-Separated Values) file format. **The data can only be used for this course project. Please collect it from the TA after signing a non-spreading agreement** (send an email to yangmu@cs.umb.edu to make an appointment).
  Note that you should not open the CSV files directly using MS Excel Spreadsheet because the data would be crashed if you do so. Read the files using a programming language, for example, Java or Matlab.
  Dataset description:
  Training data is a collection of 6 crime incidents (Streetrobbery, ResidentialBurglary, mv_larceny, Foreclousure, CommercialBurglary, Arrest) from January to December in 2006. Please note that each training data file contains crime data for one particular month. Test data is for Residential Burglary crime in January 2007.
  Each CSV file is a matrix containing 24 rows and 20 columns. Those 24*20 grids cover the geographical region of city Boston. Each entry of this matrix reflects the number of a certain crime incidents.

The coordinate file coord.CSV has the latitude and longitude of those 24*20 grids. One line is a coordination pair. Suppose A is the 24*20 grid matrix, then the coordination of A(i,j) can be found at (j-1)*24+i line in coord.CSV file (i and j start from 1).

- ## Team1 classification approach

Team 1 will focus on finding crime hotspots (have more than 1 crime) in January 2007 (1 for hotspot and 0 for coldspot).

There are three steps:

1. Training and test data generation->2. Classification -> 3. Results visualization

Step 1 Training and Test Data Generation.

Several attributes can be used to describe a Burglary crime incident. These attributes are Streetrobbery, ResidentialBurglary, mv_larceny, Foreclousure, CommercialBurglary, Arrest.

A Burglary crime incident happened in a place at a certain time can be described using those 6 features. For example: if we use crime incidents in January and February as training data, March as test data, and in the training process, we assume a month ahead may make a reasonable prediction for the following month. Then for each area x_i (one of those 24*20 grid cells), 6 attributes in January will be used as training features and Burglary crime in February will be used as the training label y_i for the area x_i,. Align all the areas we can have training data $[x_1, x_2, \dots, x_n]$ and training labels $[y_1, y_2, \dots, y_n]$. Similarly, test data is constructed by the 6 features in February, test labels for external evaluation on classification are the Burglary crime happened in March. In the following example:

Training data

	Crime 1	Crime 2	Crime 3
January	4 3 2 0	2 1 0 0	7 8 4 1
February	4 3 2 0	2 1 0 0	7 8 2 1

Test data		Test label	
March	? ? ? ?		4 3 2 0

In the figure above, a training sample has 3 crimes as its features. For the top left area, red grids are its three features and green grid is its label. Thus we have $x_1 = [4,2,7]$, $y_1 = [4]$, similarly we can have training samples: $x_2 = [3,1,8]$, $x_3 = [2,0,4]$, $x_4 = [0,0,1]$, and their corresponding labels $y_2 = [3]$, $y_3 = [2]$, $y_4 = [0]$. Because a hotspot is defined as a grid cell that has more than 1 crime incident, the label can be changed to: $y_1 = y_2 = y_3 = 1$, $y_4 = 0$ for the crime hotspot predication (1 is for hotspot and 0 is for coldspot).

In this project, we use the one month based prediction. And 12-month data happened in 2006 are used as training. January 2007 data is used for testing.

Step 2. Classification process

1 Nearest Neighbor classifier is used in this stage. Find the closest vector to each test sample in the training set and apply its label to the test sample. Use Euclidean distance as the distance measure. Calculate the precision, recall, F1 measure and accuracy for the prediction of hotspots.

Step 3. Results visualization

1) According to the Longitude and latitude provided in the documents, visualize the predicted hotspot and the real hotspot in January 2007 in Google earth through Google Earth API.
2) Discuss the differences and similarities between the predicated and real crime hotspots based on your observation.

- ## Team 2 regression approach

Team 2 will focus on predicating the exact number of Burglary crimes in January 2007.

There are three steps:

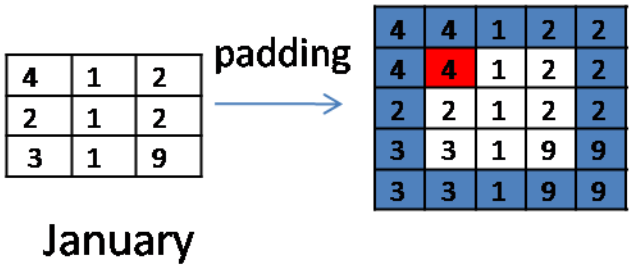
1. Training and test data generation-> 2. Regression -> 3. Results visualization

Step 1. Training and test data generation

Only 12-month Residential Burglary crime data in 2006 is used in this task.

The Burglary crimes happened in one area and its surrounding areas can be used as attributes to describe as a Burglary crime incident.

For example: In a one month based prediction, January and February data are used to predict March.



Red area in February can be described as[4, 4, 1, 4, 4, 1, 2, 2, 1] from January after padding. Training labels are the crime numbers in February. Test samples are constructed according to February. Test labels are the data in March.

In this project, we do monthly prediction. And 12-month data happened in 2006 are used as training. January 2007 data is used for testing.

Step 2. Regression process

Learn a linear regression model according to the training data and apply this model to regress the test data.

Suppose the training data is X which is a n*d matrix (n samples, d attributes). Training label is Y which is n*1 vector. We suppose Y = X*W+e, W can be computed by (TA will post a Java Library code in UMassOnline for matrix operations):

$$W = (X^T X)^{-1} X^T Y$$

Details of this equation please refer to ordinary lease squares. If $X^T X$ is not invertible, please add a diagonal matrix $I * eps$, where I is a d*d identity matrix and eps is a very close to 0 such as 10^{-10}.

e can be calculated by $Y - XW$.

For test data X', predicted label Y'can be calculated by $Y' = X'W + e$.

Part 3. Results visualization

1) According to the Longitude and latitude provided in the documents, visualize the predicted hotspot and the real hotspot in January 2007 in Google earth through Google Earth API.
2) Discuss the differences and similarities between the predicated and real crime hotspots based on your observation.

Submission Requirements

1. Write an experiment report to discuss your experimental results, including detailed parameter settings and a **KML file** for the Google earth visualization results. Submit the paper copy of the report, and source code of the scripts with the cover page in class. Paper copy should be bound firmly together as one pack (for example, staple, but not limited to, at the left corner). 5 points will be deducted for unbounded homework.
2. Submit the softcopy of the report and scripts through your UMassOnline account athttp://boston.umassonline.net/index.cfm.
3. Zip all the files. One submission per team. Save the file as sdm_ teamNumber. For example, Team 1 should name their file as*sdm_team1.zip.*
4. No hard copies or soft copies results in 0 points.