Using Big Data Analytics To Predict Crime Type

Abstract

With the development of the information technology and the computer science, big data is widely used in prediction analytics. Nowadays Crime prediction plays an important role in improving public security and reducing the financial loss of crimes. Studies of crime type predictions, combined with the current police-focused predictions of crime time and place, can make crime predictions more effective. In this paper, we use the crime data of Los Angeles in the last seven years to study the crime type prediction using k means and logistic regression algorithms respectively. 5-fold cross-validation is used to get a reliable and stable model. The results show that both algorithms can be used to predict the type of crime, and logistic regression algorithm is superior to k means algorithm. At last, combined with examples of real word, we analyze the impact of our results.

1. Introduction

1.1. Background introduction

With the development of the information technology and the computer science, big data is widely used in prediction analytics. Since the data generation regarding crime is also increased nowadays. Crime prediction is no longer strange for us. Crime prediction plays an important role in improving public security and reducing the financial loss of crimes. The figure 1 is a real crime prediction case in U.S. One commonly used approach in predictive policing seeks to forecast where and when crime will happen; another focuses on who will commit crime or become a victim.

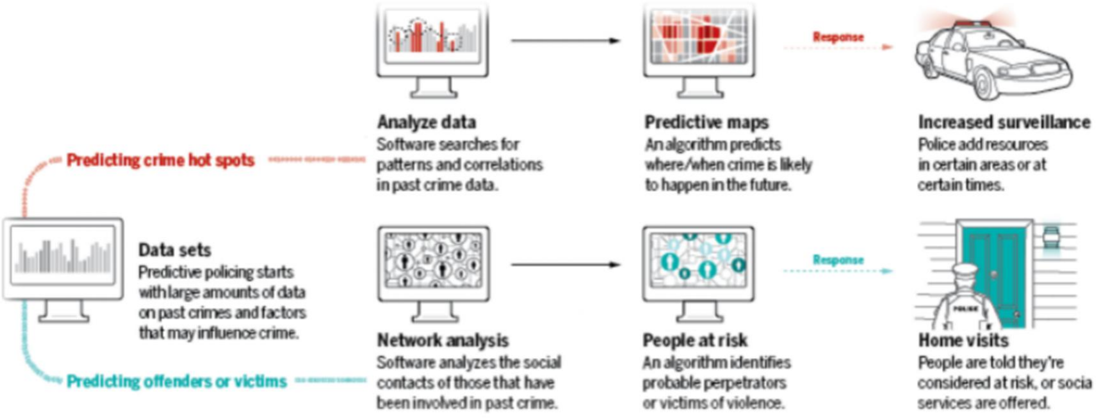

Predicting crime hot spots
Analyze data
Software searches for patterns and correlations in past crime data.
Predictive maps
An algorithm predicts where/when crime is likely to happen in the future.
Response
Increased surveillance
Police add resources in certain areas or at certain times.
Data sets
Predictive policing starts with large amounts of data on past crimes and factors that may influence crime.
Predicting offenders or victims
Network analysis
Software analyzes the social contacts of those that have been involved in past crime.
People at risk
An algorithm identifies probable perpetrators or victims of violence.
Response
POLICE
Home visits
People are told they're considered at risk, or socia services are offered.

Fig. 1 A real crime prediction case

1.2. Why this topic

Firstly, it is close to life. Crime is around us and impacts on our lives. As the background introduction introduced, crime prediction has already been effective. Secondly, it is interesting and cool. Only the cool person (e.g. FBI) could predict the crime, which has a really sense of justice. Meanwhile we want to know whether we can predict crime type with what we have learnt and different algorithms' performance. Research on what impacts the prediction of crime type may have in the real world is one thing we are very curious about.

2. Hypothesis

2.1. Inquiry question

Based on the interesting crime topic, the main idea of the hypothesis in the case study which is "Can the big data predict crime type?" This is precise topic found and can be tested in the real world related to crime information.

2.2. Scope

Crime is the general topic which cannot be tested or measured easily. Then narrowing the topic should be needed in first step of defining the hypothesis. If the crime information related to the whole world, it will be too big and not accurate from the result. Because there are different crime rates in different countries especially it may be higher in poor countries. Therefore, the dataset of one city should be the suitable scope to do the case study.

Moreover, crime type is more specific than crime. It is exactly the classification we can be recognized. Surely, the crime types are defined by the related the police department which is well-structure and expect to do that.

2.3. Algorithms

In the case study, two algorithms should be used like k-means and logistic regression which are famous clustering methods in the experiment.

2.4. Result testing

For measuring the result, the accuracy of the results is the main measurement to prove whether the big data can be predicted the crime type.

2.5. Assumption of implementation in the real world

If the result can be proved the hypothesis, it should be assisted in the real case such as the police using the data to predict the crime. The crime type prediction will assist police to allocate manpower, make specific prediction before the crime happened and do the actions efficiently after the crime happened.

3. Method:

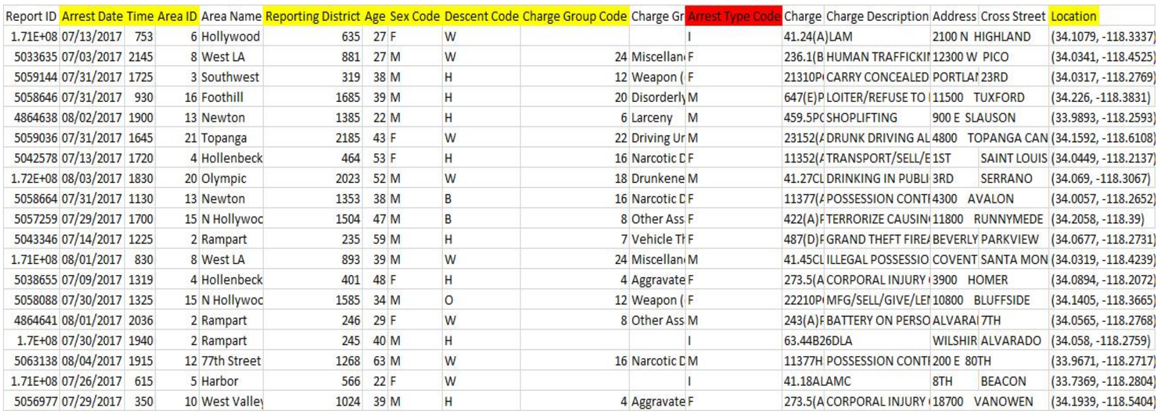
3.1. Dataset

In the case study, we have 1.11 million records in the database, there are 17 kinds of features like basically date, time, location, criminals' age and gender, and crime types etc. The data is about 7 years from 2010 and it is kept updating by Los Angeles Police Department (LAPD). The latest update date is November 22, 2017. LAPD is the governmental department, so the date is reliable and trusted.

Crime Type Code	D - Dependent	F - Felony	I - Infraction	M - Misdemeanor	O - Other

3.2. Feature Engineering

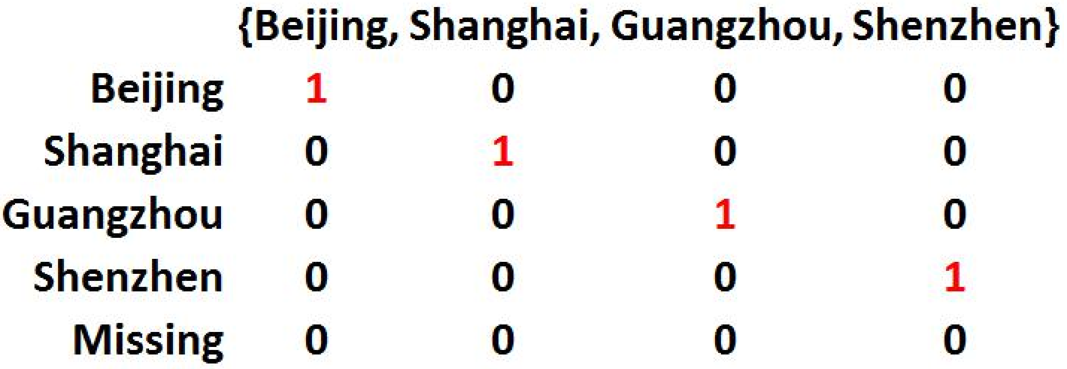
Obviously raw data cannot be used for our prediction task, so in the beginning we filtered out irrelevant features and selected some useful features at the same time. A sample of the raw data can be seen in Fig. 1. The red-colored column "Arrest Type Code" is the label that we used to predict, while the yellow marked columns are feature. It is clear to see Arrest Date, Area ID and so on from this figure.



Report ID	Arrest Date	Time	Area ID	Area Name	Reporting District	Age	Sex Code	Descent Code	Charge Group Code	Charge Gr	Arrest Type Code	Charge	Charge Description	Address	Cross Street	Location
1.71E+08	07/13/2017	753	6	Hollywood	635	27	F	W			I	41.24(A)LAM		2100 N HIGHLAND		(34.1079, -118.3337)
5033635	07/03/2017	2145	8	West LA	881	27	M	W	24	Miscellan	F	236.1(B	HUMAN TRAFFICKI	12300 W PICO		(34.0341, -118.4525)
5059144	07/31/2017	1725	3	Southwest	319	38	M	H	12	Weapon (	F	21310P	CARRY CONCEALED	PORTLAI	23RD	(34.0317, -118.2769)
5058646	07/31/2017	930	16	Foothill	1685	39	M	H	20	Disorderl	M	647(E)P	LOITER/REFUSE TO	11500 TUXFORD		(34.226, -118.3831)
4864638	08/02/2017	1900	13	Newton	1385	22	M	H	6	Larceny	M	459.5P(	SHOPLIFTING	900 E SLAUSON		(33.9893, -118.2593)
5059036	07/31/2017	1645	21	Topanga	2185	43	F	W	22	Driving Ur	M	23152(	DRUNK DRIVING AL	4800 TOPANGA CAN		(34.1592, -118.6108)
5042578	07/13/2017	1720	4	Hollenbeck	464	53	F	H	16	Narcotic D	F	11352(	TRANSPORT/SELL/E	1ST	SAINT LOUIS	(34.0449, -118.2137)
1.72E+08	08/03/2017	1830	20	Olympic	2023	52	M	W	18	Drunkene	M	41.27CL	DRINKING IN PUBLI	3RD	SERRANO	(34.069, -118.3067)
5058664	07/31/2017	1130	13	Newton	1353	38	M	B	16	Narcotic D	F	11377(	POSSESSION CONT	4300 AVALON		(34.0057, -118.2652)
5057259	07/29/2017	1700	15	N Hollywoc	1504	47	M	B	8	Other Ass	F	422(A)F	TERRORIZE CAUSIN	11800 RUNNYMEDE		(34.2058, -118.39)
5043346	07/14/2017	1225	2	Rampart	235	59	M	H	7	Vehicle Th	F	487(D)F	GRAND THEFT FIRE	BEVERLY	PARKVIEW	(34.0677, -118.2731)
1.71E+08	08/01/2017	830	8	West LA	893	39	M	W	24	Miscellan	M	41.45CL	ILLEGAL POSSESSIO	COVENT	SANTA MON	(34.0319, -118.4239)
5038655	07/09/2017	1319	4	Hollenbeck	401	48	F	H	4	Aggravate	F	273.5(A	CORPORAL INJURY	3900 HOMER		(34.0894, -118.2072)
5058088	07/30/2017	1325	15	N Hollywoc	1585	34	M	O	12	Weapon (	F	22210P	MFG/SELL/GIVE/LE	10800 BLUFFSIDE		(34.1405, -118.3665)
4864641	08/01/2017	2036	2	Rampart	246	29	F	W	8	Other Ass	M	243(A)F	BATTERY ON PERSO	ALVARAI	7TH	(34.0565, -118.2768)
1.7E+08	07/30/2017	1940	2	Rampart	245	40	M	H			I	63.44B26DLA		WILSHIR	ALVARADO	(34.058, -118.2759)
5063138	08/04/2017	1915	12	77th Street	1268	63	M	W	16	Narcotic D	M	11377H	POSSESSION CONT	200 E 80TH		(33.9671, -118.2717)
1.71E+08	07/26/2017	615	5	Harbor	566	22	F	W			I	41.18ALAMC		8TH	BEACON	(33.7369, -118.2804)
5056977	07/29/2017	350	10	West Valley	1024	39	M	H	4	Aggravate	F	273.5(A	CORPORAL INJURY	18700 VANOWEN		(34.1939, -118.5404)

Fig. 2 A sample of the raw data

Features in the raw data can be classified into two major types. The first one is the feature with enumerable values. One-hot Encoding was utilized to process this kind of features, using which we can expand one feature into many features. It is applied for Area ID, Descent Code and etc. For the details of One-hot Encoding, suppose we have a feature named city which only consists of Beijing, Shanghai, Guangzhou and Shenzhen. The relative relationship between these cities is hard to identify. If this feature is relevant to our prediction, we should not discard it for its rich information. If one instance has the value Beijing for feature city, we can represent this feature with a vector [1, 0, 0, 0]. Other cities can also be represented respectively by One-hot Encoding as shown in Fig. 2. If the value of city feature in an instance is missing, then it can be represented by 0s. After this process, if a city is more relevant to our prediction, e.g. Beijing, then it will gain higher weight than other cities.



	{Beijing,	Shanghai,	Guangzhou,	Shenzhen}
Beijing	1	0	0	0
Shanghai	0	1	0	0
Guangzhou	0	0	1	0
Shenzhen	0	0	0	1
Missing	0	0	0	0

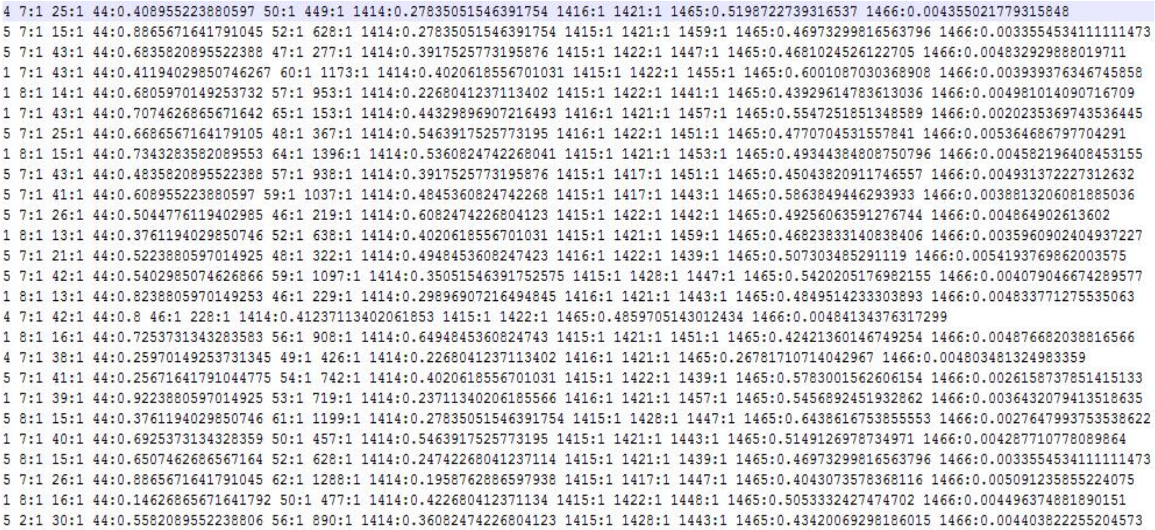
Fig. 3 An example of One-hot Encoding

Another type of features is the numerical feature such as a person's age. We used feature scaling technique [1] to prevent features with large values from dominating the prediction. For example, if we want to predict the percentage for a person going to get cancer and we only have two features, a person's age and the amount of this person's hair. Obviously, age feature is more relevant to this prediction task, but the hair feature will get higher weight since it is larger than age in terms of amount, which is unreasonable. In order to prevent this from happening, we use feature scaling, as shown in Fig. 3, to scale these features into the same range, e.g. in the scale of 0 to 1. This process applied for Longitude, Latitude and etc. For missing values, we used the average of existing values of the feature to fill up. These average values will also be scaled.

$$\bar{x}_{ij} = \frac{x_{ij} - x_j^{min}}{x_j^{max} - x_j^{min}}$$

Fig. 4 Formula of feature scaling

Fig. 4 shows the data after processing. The first column is the label, from which five classes can be seen. Other columns are features, and the integer in the front of colon is the index of a feature and the decimal behind the colon is the corresponding value. Those features with 0 values are discarded because storing them costs a lot of memory.



```
4 7:1 25:1 44:0.408955223880597 50:1 449:1 1414:0.27835051546391754 1416:1 1421:1 1465:0.5198722739316537 1466:0.004355021779315848
5 7:1 15:1 44:0.8865671641791045 52:1 628:1 1414:0.27835051546391754 1415:1 1421:1 1459:1 1465:0.46973299816563796 1466:0.0033554534111111473
5 7:1 43:1 44:0.6835820895522388 47:1 277:1 1414:0.3917525773195876 1415:1 1422:1 1447:1 1465:0.4681024526122705 1466:0.004832929888019711
1 7:1 43:1 44:0.41194029850746267 60:1 1173:1 1414:0.4020618556701031 1415:1 1422:1 1455:1 1465:0.6001087030368908 1466:0.003939376346745858
1 8:1 14:1 44:0.6805970149253732 57:1 953:1 1414:0.2268041237113402 1415:1 1422:1 1441:1 1465:0.43929614783613036 1466:0.004981014090716709
1 7:1 43:1 44:0.7074626865671642 65:1 153:1 1414:0.44329896907216493 1416:1 1421:1 1457:1 1465:0.5547251851348589 1466:0.0020235369743536445
5 7:1 25:1 44:0.6686567164179105 48:1 367:1 1414:0.5463917525773195 1416:1 1422:1 1451:1 1465:0.4770704531557841 1466:0.005364686797704291
1 8:1 15:1 44:0.7343283582089553 64:1 1396:1 1414:0.5360824742268041 1415:1 1421:1 1453:1 1465:0.49344384808750796 1466:0.004582196408453155
5 7:1 43:1 44:0.4835820895522388 57:1 938:1 1414:0.3917525773195876 1415:1 1417:1 1451:1 1465:0.45043820911746557 1466:0.004931372227312632
5 7:1 41:1 44:0.608955223880597 59:1 1037:1 1414:0.4845360824742268 1415:1 1417:1 1443:1 1465:0.5863849446293933 1466:0.0038813206081885036
5 7:1 26:1 44:0.5044776119402985 46:1 219:1 1414:0.6082474226804123 1415:1 1422:1 1442:1 1465:0.49256063591276744 1466:0.004864902613602
1 8:1 13:1 44:0.3761194029850746 52:1 638:1 1414:0.4020618556701031 1415:1 1421:1 1459:1 1465:0.46823833140838406 1466:0.0035960902404937227
5 7:1 21:1 44:0.5223880597014925 48:1 322:1 1414:0.4948453608247423 1416:1 1422:1 1439:1 1465:0.507303485291119 1466:0.0054193769862003575
5 7:1 42:1 44:0.5402985074626866 59:1 1097:1 1414:0.35051546391752575 1415:1 1428:1 1447:1 1465:0.5420205176982155 1466:0.004079046674289577
1 8:1 13:1 44:0.8238805970149253 46:1 229:1 1414:0.29896907216494845 1416:1 1421:1 1443:1 1465:0.4849514233303893 1466:0.004833771275535063
4 7:1 42:1 44:0.8 46:1 228:1 1414:0.41237113402061853 1415:1 1422:1 1465:0.4859705143012434 1466:0.00484134376317299
1 8:1 16:1 44:0.7253731343283583 56:1 908:1 1414:0.6494845360824743 1415:1 1421:1 1451:1 1465:0.42421360146749254 1466:0.004876682038816566
4 7:1 38:1 44:0.25970149253731345 49:1 426:1 1414:0.2268041237113402 1416:1 1421:1 1465:0.26781710714042967 1466:0.004803481324983359
5 7:1 41:1 44:0.25671641791044775 54:1 742:1 1414:0.4020618556701031 1415:1 1422:1 1439:1 1465:0.5783001562606154 1466:0.0026158737851415133
1 7:1 39:1 44:0.9223880597014925 53:1 719:1 1414:0.23711340206185566 1416:1 1421:1 1457:1 1465:0.5456892451932862 1466:0.0036432079413518635
5 8:1 15:1 44:0.3761194029850746 61:1 1199:1 1414:0.27835051546391754 1415:1 1428:1 1447:1 1465:0.6438616753855553 1466:0.0027647993753538622
1 7:1 40:1 44:0.6925373134328359 50:1 457:1 1414:0.5463917525773195 1415:1 1421:1 1443:1 1465:0.5149126978734971 1466:0.004287710778089864
5 8:1 15:1 44:0.6507462686567164 52:1 628:1 1414:0.24742268041237114 1415:1 1421:1 1439:1 1465:0.46973299816563796 1466:0.0033554534111111473
5 7:1 26:1 44:0.8865671641791045 62:1 1288:1 1414:0.1958762886597938 1415:1 1417:1 1447:1 1465:0.4043073578368116 1466:0.005091235855224075
1 8:1 16:1 44:0.14626865671641792 50:1 477:1 1414:0.422680412371134 1415:1 1422:1 1448:1 1465:0.5053332427474702 1466:0.004496374881890151
5 2:1 30:1 44:0.5582089552238806 56:1 890:1 1414:0.36082474226804123 1415:1 1428:1 1443:1 1465:0.43420069298186015 1466:0.004403822255204573
```

Fig. 5 A sample of processed data

3.2.1. n- fold cross validation

It is a method that devided the data sets into n parts, using n-1 parts to train, another part to test in turn, and computing the average value in the end.

In this way, can we get a reliable and stable model. In our trial, we have set n=5.

3.3. K means

3.3.1. Introduction

K means algorithm is a clustering algorithm based on the mean of clustering. Its basic idea is to find a partition scheme of K clusteringa by iteration, so that the global error by using the mean of K clustering to represent all kinds of samples is smallest.

The basis of the k-means algorithm is the Minimum Square Fitting Error Sum, and it's cost function is

$$J(\mathrm{c},\mu)=\sum_{i=1}^{k}\| \mathrm{x}^{(\mathrm{i})}-\mu_{c^{(\mathrm{i})}} \|^2$$

Fig. 6 Formula of cost function of K means

In the formula, $\mu_{\mathrm{c}^{(\mathrm{i})}}$ represents the mean of the i cluster. We hope the minimum cost function.

Intuitively speaking, the more similar the samples are, the smaller squared error between the data objects and the mean value. Then calculated the sum of the square-error for all clustering, as a result, we can verified that whether each cluster is optimal.

3.3.2. Algorithm

Input: clustering number k, and data containing n data objects.

Output: k clustering which satisfies the minimum standard of variance.

It has 4 steps.

Step 1. Selecting k objects from the n data objects as the initial cluster center randomly.

Step2. Assigning each data object to the cluster with the closest center.

Step3. Computing the mean of each clustering object, and assigning them to be the new cluster centre.

Step4. Repeat step2 and step3 until the convergence has been reached.

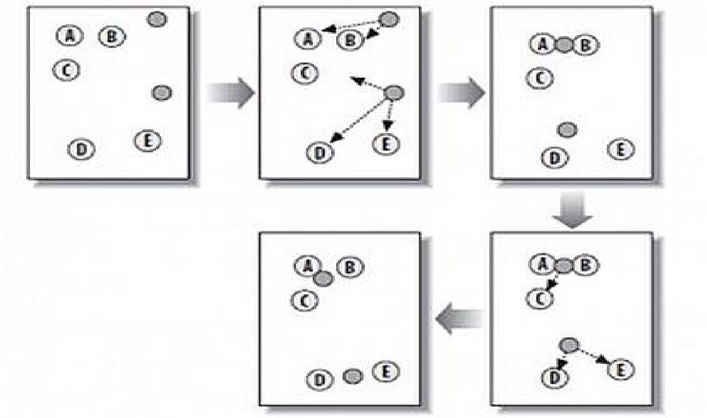


Fig. 7 Schematic of K means algorithm

In our trial, n= 1,113,552 and K is set to 5.

3.3.3. Process tool

Sklearn is a simple and efficient tool for data mining, the use of sklearn tools can facilitate feature engineering and model training.

It can realize k-means by paritioning the dataset into 5 clusters firstly. Then choosing the most origin labels as the cluster label as the dataset has already been labelled. After that, each test sample will be assigned to one of those 5 clusters. Finally, map to the origin label and calculate accuracy.

In our example, the cluster label is A.

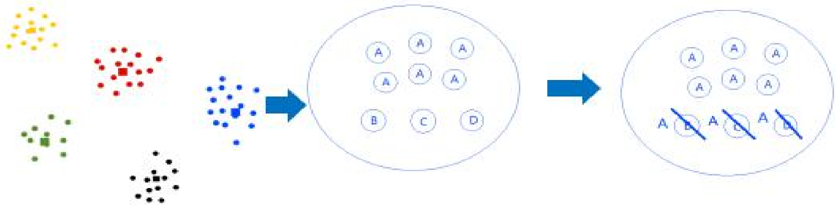


Fig. 8 process by sklearn

3.3.4. Scope of application and defects

The K menas algorithm tries to find a cluster that minimized the ordinary error criterion function. When the potential cluster shape is convex, the difference between clusters is obvious, and when the size of the cluster is close, the clustering result is ideal. The algorithm's time complexity is O (tKmn), which is linear correlation to sample size. Therefore, the algorithm is efficient and scalable to deal with large data sets. But in addition to determining the cluster number k and the initial cluster center, the algorithm always end with local optimal solution, sensitive to noise and outliers. What's more, it is not suitable for finding non convex shape clusters or vary the size of the cluster.

3.4. Logistic Regression

3.4.1. Introduction

Another method we use is logistic regression. Logistic regression, also known as logistic regression analysis, is a generalized linear regression analysis model. It is commonly used in data mining, disease diagnosis, economic forecasting and other fields. The use of maximum likelihood estimates ensures that the fit at each point is optimal. The process of logistic regression can be described like this: To figure out a regression or classification problem, firstly construct the cost function, and then the optimal model parameters are iteratively solved through the optimization method. Finally test the quality of the model that we construct.

3.4.2. Regression steps

(a) Find ħ function (i.e. prediction function)

Logistic function (or Sigmoid function), the function form is as follow:

$$g(z)=\frac{1}{1+e^{-z}}$$

So the prediction function can be constructed as follow:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

(b)Construct J function (i.e. loss function)

The Cost function and the J function are as follows, which are derived based on the maximum likelihood estimation. (m samples, each sample has n features)

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & if\ y = 1 \\ -\log(1 - h_\theta(x)) & if\ y = 0 \end{cases}$$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(h_\theta(x_i), y_i) = -\frac{1}{m}\left[\sum_{i=1}^{m} (y_i log h_\theta(x_i) + (1-y_i)\log(1-h_\theta(x_i)))\right]$$

(c) Find ways to minimize the J function and obtain the regression parameters (θ)

Use gradient descent method to get the minimum θ. The vectorization method and the regularization method can also be used to get the minimum θ.

$$\theta_j := \theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)x_i^j$$

3.4.3. multi-class classification

Logistic regression is commonly used for binary classification. For label represented by positive one and negative one, minimizing this loss function we can get weights for each feature, which are going to be used for prediction.

$$\min_{\omega} \frac{1}{2}\omega^T\omega - C\sum_{i=1}^{m}\log\left(\sigma(y_i\omega^T x_i)\right)$$

So how do we use logistic regression to do multi-class classification? We used the one-vs-rest technique. As shown in Figure 12. Suppose we are going to classify the data into three classes. We can choose one class as positive class and the others are treated as negative class. Then we do a binary classification. We can do this process for each class iteratively. After the training phase, when we do the prediction for a sample, it can get three possibilities for different classes. The class with the highest possibility is where this sample belongs to. As for implementing our task, we used this open source tool -- liblinear.

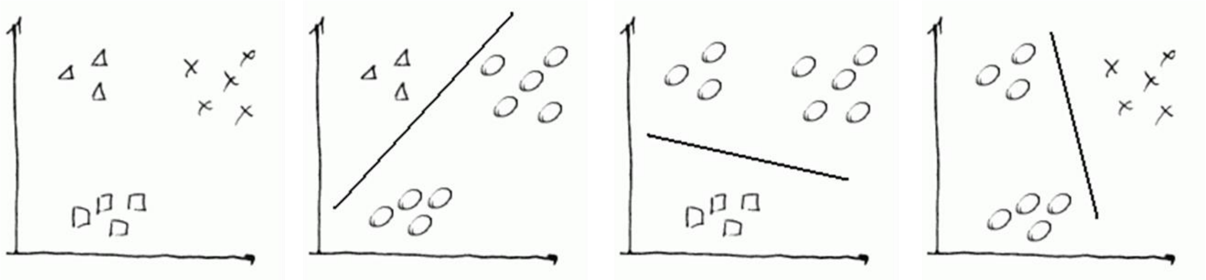


Fig. 8 Schematic of one-vs-rest technique

4. Impact in the real world

In the real world, nowadays, most police is using the data to predict crime by statistics method. Intuitionally, there is higher crime rate at night comparing with in the morning in general. So, the police should focus on some high risky places at night according to the data. Crime is actually very similar. Some crimes are caused by built-in features of the environment, like a bar that closes at 2 a.m. every night, unleashing rowdy drunks onto a neighborhood. Others, such as a series of gang murders or a rash of neighborhood burglaries, happen because criminals' success invites more crimes or incites retaliation. Criminologists call this "repeat victimization"—the criminal equivalent of aftershock. Thus, in most cases, data prediction and analysis are needed.

From the result of case study, obviously, it should be implemented in the real case. When the police predict crime type from the data sets like Felony, police should add resources in certain areas at certain times for surveillance. Otherwise, if the prediction of crime type is Misdemeanor, police should decrease surveillance and allocate the related resources to others efficiently.

Furthermore, the police department has taken predictive policing one step further—and made it personal. The department is using network analysis to generate a highly controversial Strategic Subject List of people deemed at risk of becoming either victims or perpetrators of violent crimes. Officers and community members then pay visits to people on the list to inform them that they are considered high-risk.

Finally, in other words, the prediction should be where the next recorded police observations are going to occur. Because the police will do the actions like increased surveillance or home visits according to the prediction. It is high risk that if the criminals get and data and do some similar prediction. They may be easier to escape for police chasing. Therefore, the data protection should be another thing the police should concern.

5. Conclusions

5.1. Result

	K means	logistic regression
1	0.621467	0.816432
2	0.622082	0.815907
3	0.619398	0.817385
4	0.621668	0.817014
5	0.622283	0.816632
Average	0.621379	0.816674

Table 1 Results

From the table, it is easy to conclude that the prediction accuracy rate for logistic regression is higher than K means, which means the accuracy rate of predicting crime type is almost 81 percent by using logistic regression, while K means only hold the accuracy rate of 62 percent.

5.2. Casual Analyse

(a) K means is not fit for the classification with labels, as it is a clustering algorithm.

(b) The data preprocess is more suitable for logistic regression than K means, so logistic regression holds the higher accuracy rate.

5.3. Conclusion

Big data can predict the crime type, which means our hypothesis has been confirmed.

Furthermore, it can be useful in real case as we expected. That is to say, helping arrest the criminal as a good assistant for police. Distinguishing possible crime type in various time and space with the different manpower allocating and equipment deployment before the crime, response swift and formulating countermeasures appropriately after the crime.

All the things above can reduce crime rate as well as the loss and hurt caused by the crime, therefore, making the earth a better place to live.

5.4. Lessons Learned

(1) how to collect useful data.
(2) Have a good comprehension of data preprocess.
(3) Know well about the k means and logistic regression, take a good practice with the combination of class understanding and self learning.
(4) Rational division and regular discussion to improve efficiency and create a favorable learning environment.

References

[1] Fan, Rong-En, et al. "LIBLINEAR: A library for large linear classification." *Journal of machine learning research* 9. Aug (2008): 1871-1874.

[2] Meng Jianliang, Shang Hai kun, Bian Ling. The application on intrusion detection based on K-means cluster algorithm [C]. International Forum on informationTcehnology and Applications. 2009:150-152.

[3] Inaba, M.; Katoh, N.; Imai, H. Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering. Proceedings of 10th ACM Symposium on Computational Geometry. 1994: 332−339.

[4] Daniel Antolos, Dahai Liu, Andrei Ludu, Dennis Vincenzi. Burglary Crime Analysis Using Logistic Regression. 2013: Human Interface and the Management of Information. Information and Interaction for Learning, Culture, Collaboration and Business,pp 549-558.

[5] Daniel Antolos. Investigating Factors Associated with Burglary Crime Analysis using Logistic Regression Modeling. (2011) Dissertations and Theses. Paper 15

https://data.lacity.org/A-Safe-City/Arrest-Data-from-2010-to-Present/yru6-6re4

https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/y8tr-7khq/data

http://www.sciencemag.org/news/2016/09/can-predictive-policing-prevent-crime-it-happens

Personal Statement：

STUDENT ID: 55287206

STUDENT NAME: Li Lei

We all together proposed the question, prepared the proposal, collected the data, analyzed and summarized the experimental results, presented the final slides and wrote the report.

As to separate parts for this project, I individually spent a lot of time on the feature engineering (preprocessing the data) and implementing our experiments. Basically, each one of us contributes equally to this project.

From this project, I learned how to process the data in a form that the open-source tools can employ, and particularly the techniques, One-hot Encoding and Feature Scaling, which are widely used in practice. Also, I got a chance to know the details of Logistic Regression and K-means algorithms and some other techniques we utilized in these two algorithms such as One-vs-all.

STUDENT ID: 54897176

STUDENT NAME: YAO Peijue

We all together proposed the question, prepared the proposal, collected the data, analyzed and summarized the experimental results, presented the final slides and wrote the report.

As to separate parts for this project, I spent a lot of time on learning logistic regression algorithm and implementing this algorithm with ZHOU Yanjun. Basically, each one of us contributes equally to this project.

From this project, I learned a new algorithm and find the open tool (libsvm) to realize it. Besides, it gave me a chance to learn the field of crime prediction. I found that many institutes recorded the crime data, and many institutes in the world devoted themselves to find kinds of ways to predict the crime.

STUDENT ID: 55041231

STUDENT NAME: ZHOU Yanjun

We all together proposed the question, prepared the proposal, collected the data, analyzed and summarized the experimental results, presented the final slides and wrote the report.

As to separate parts for this project, I learnt and implemented Logistic Regression algorithm with YAO Peijue. And I wrote the abstract, introduction and part of the logistic regression algorithm part of the report. Basically, each one of us contributes equally to this project.

From this project, I learned Logistic Regression and K means algorithms and how to implement them with open source tool. I also learnt a lot of techniques such as One-hot Encoding, Feature Scaling, One-vs-all and n- fold cross validation. Besides the techniques, what I got most from the project is the analytical method and the whole process of a case study. Group work is very important. I learnt a lot from my group members.

STUDENT ID: 54949626

STUDENT NAME: LI Qingru

Our group divide the work reasonable, and I am responsible for the data process by K-means with the help of python and things related to the conclusion, also write the related part in the report. What's more, our group members have finished most parts together, including data collection, problem set, summary analysis, PowerPoint completion, report integration and so on, each of us contributed a lot.

In this project, I have learned how to cooperate with others, as we always spend time discussing and exchanging our ideas, in this way, can we work very productively. Secondly, I now have a deep comprehension of the clustering method - K means, as it strengths in simple principle, easily implement, good clustering effect. On the other hand, it is sensitive to outliers, the complexity of the algorithm is not easy to control, local optimal solution rather than global optimum solution. So, when to choose it is important. Finally, my programming abilities in python is improved. I think that all things are helpful in my future studying, and I will work hard.

STUDENT ID: 54006280

STUDENT NAME: CHAN Kwun Lam

I attended and participated discussion in each team meeting. We all did the data collection, proposal presentation preparation, mentioning problems, summarizing conclusion and PPT preparation.

In presentations, I presented the whole introduction, hypothesis and method introduction in the proposal presentation. Moreover, I presented the hypothesis and method introduction in the final presentation.

In the report, we discussed together, and I mainly did the hypothesis, method introduction – database, and impact in the real world (assumption if the prediction is used in real case).

During the project, I learned how two algorithms (K-means and logistic regression) implementations in the real case. Also, I learned how to use big data to do analysis, how to set hypothesis, narrow the scope, and make it can be tested. In the discussion, I learned how to communicate with classmates who have different background especially knowledge areas.

Thanks for each teammate's contribution.