Machine Learning Method To Analyze Criminal Data

Ida Touray and Kiyatou Konate

Computer Information Technology ~ Marcos S. Pinto

Abstract

Machine Learning is a set of programmatic tools used to analyze huge amounts of data, so called big data. These tools enable systems to process, learn from, and draw actionable insights out of big data. Some of these tools are powerful algorithms used to spot patterns in data and make predictions about future events. One of the areas that machine learning has been employed is people's personal safety through the analysis of criminal data over a certain geographical region. Law Enforcement officials have turned to data mining and Machine Learning to aid in the fight of crime prevention and Law Enforcement. This project applies machine learning algorithm to a dataset of criminal activity to predict and take action against criminal activities and potential security risks. The project is done using Python programming language and its array of libraries properties for Machine Learning. The dataset is obtained from data.world/data-society/nyc-crime-data which contains information on NYPD complaint data from 2006 to 2015.

Introduction

Machine Learning is an application of Artificial Intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

Machine Learning focuses on the development of computer programs that can access data and use it to learn for themselves.

Python is the most used computer programming language in machine learning for data analysis and data science. Python offers concise and readable code; it also allows the developers to write reliable systems.

The New York police department keeps record of all of its crimes and for this research, we are only going to focus on New York City from 2006 to 2015. This data keeps record of the time, location, borough, description of the crime and many more across the five boroughs of New York City. Some cases were solved or added to the data in the same day as the complained was made, while other last up to years. This criminal data is used as an indication of how safe a borough is, and also as an aid to city authorities to implement policies/laws that provide more security to the people that live in that borough.

Data Cleansing

Steps used to understand data in machine learning are:

Step 1: Examining the Data Set

Loading The Data Into Pandas(a Python library).

Step 2: Narrowing Down Our Columns for Cleaning

First Group Of Columns

Second Group Of Columns

Third Group Of Columns

investigating FICO Score Columns

Decide On A Target Column

Visualizing the Target Column Outcomes

Remove Columns with only One Value

Step 3: Preparing the Features for Machine Learning

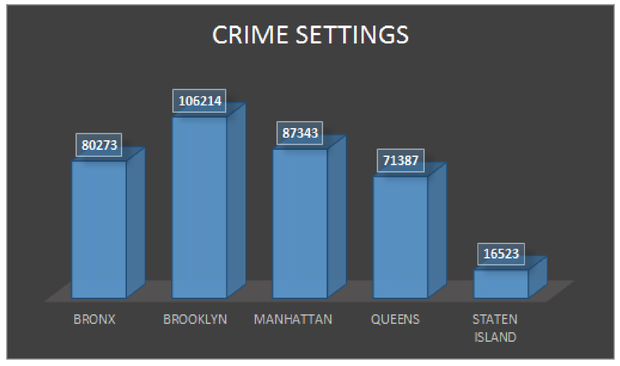
Handle Missing Values

Investigate Categorical Columns

Convert Categorical Columns to Numeric Features

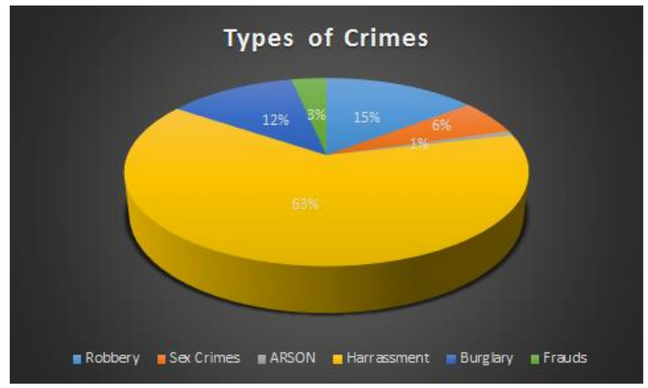
Save to CSV

Result/ Findings



Using the COUNTIF formula on MS Excel, Brooklyn has the highest crime rates in NYC and therefore can be see ass the least safe borough form 2006 to 2015.

After data cleaning, we will decide on which algorithm (model) is better suited to run a machine learning program that results in a better prediction of the likelihood of a complaint to occurs in a certain borough. Among the algorithms of machine learning such as Linear Regression and Logistic Regression. We will try use one of these two algorithms.



This chart demonstrates few of the most committed crimes across the five boroughs of New York City. Other crimes not showed in this chart includes: Murder, Car theft, drug abuse and many more.

Other categories of these data includes, suspects' and victims' age and race, time of the day and year a particular crime was committed, where the crime took place, either in a home, in the streets or other public places.

All these information is put together to predict the likelihood of crimes in NYC to help city dwellers stay alert of their environment.

Reference

DevlinData, Josh. "Data Cleaning and Preparation for Machine Learning." *Dataquest*, 23 Sept. 2019, https://www.dataquest.io/blog/machine-learning-preparing-data/.

(NYPD), Police Department. "NYPD Complaint Data Current (Year To Date): NYC Open Data." *NYPD Complaint Data Current (Year To Date) | NYC Open Data*, 1 Nov. 2019, https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243.

Computer Vision - Face detection

Sumya Raha, Prof: Marcos Pinto

Abstract

The primary aim of face detection algorithms is to determine whether there is any face in an image or not. It is a part of object detection and can be used in many areas such as security, bio-metrics, law enforcement, entertainment, personal safety, etc. Faces must be detected with all manner of orientations, angles, light levels, hairstyles, hats, glasses, facial hair, makeup, ages, and so on. The algorithms must be trained on huge data sets containing hundreds and thousands of face and non-face images. Once trained, the algorithms are able to answer two questions in response to input in the form of an image: are there any faces in this image? and, if yes, where are they? If a face or faces are present in an image, the algorithms will answer these questions by placing a bounding box around the detected face(s). The project will use the Viola-Jones algorithm, one of the most important algorithms for face detection, which is a fast detection method that slowly detects a face through a computation of matching faces proportions in an image.

Background Information

- Face detection is a type of computer vision technology that is able to identify faces within digital images. Face detection is pre step to facial recognition, facial analysis, and facial tracking.
- Facial recognition involves identifying the face in the image as belonging to person X and not person Y.
- Facial analysis tries to understand something about people from their facial features, like determining their age, gender, or the emotion they are displaying.
- Facial tracking is mostly present in video analysis and tries to follow a face and its features (eyes, nose, and lips) from frame to frame.
- Computer can't see colors however colors can be converted to numbers which can be read by computers. To convert colors to numbers, the computer uses various color models.
- The smallest element of an image is called a pixel. It is basically a dot in the picture. An image contains multiple pixels arranged in rows and columns.
- A feature is a piece of information in an image that is relevant to solving a certain problem. Computer vision and image processing have a large collection of useful features.

Viola-Jones Face detection Algorithms

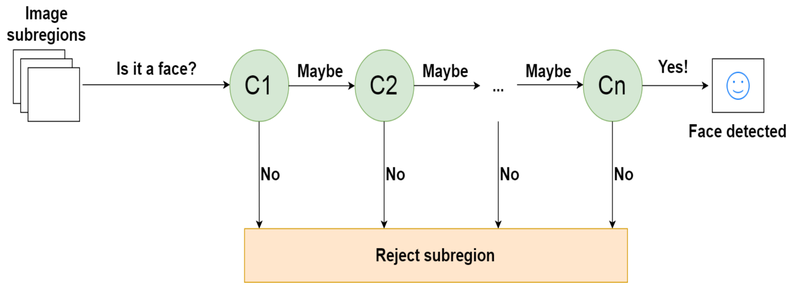
Viola-Jones algorithm, created by Paul Viola and Michael Jones in 2001, is the first form of face detection. According to "Traditional Face detection", by Ivancic, "Viola and Jones developed a general object detection framework that was able to provide competitive object detection rates in real time." . This algorithm has 4 main steps:

1. Selecting Haar Like features: this step is used to determine if an image contains a human face or not by using rectangular part of an image and dividing that rectangular part into multiple parts.

2. Creating an Integral Image: This step basically makes the Haar like features more efficient by calculating using the summed area table.

3. Running AdaBoost Training: This step is used to make a strong classifier by using many weak classifiers. Boosting is used to combine the weak classifiers to make a strong classifier.

4. Creating classifier cascades: This process eliminates any images that does not contain faces to avoid wasting time and computations. This is evaluated by the first stage to determine if it gets a no or a maybe. If it gets a no then the image enters the reject subregion, but if it gets a maybe then it is sent to the next stage. This is repeated until the image passes through all of the cascades.



References

Real Python. "Traditional Face Detection With Python." *Real Python*, Real Python, 31 July 2019, https://realpython.com/traditional-face-detection-python/.

Bettilyon, Tyler Elliot. "How to Classify MNIST Digits with Different Neural Network Architectures." Medium, Teb's Lab, 9 May 2019, https://medium.com/tebs-lab/how-to-classify-mnist-digits-with-different-neural-network-architectures-39c75a0f03e3

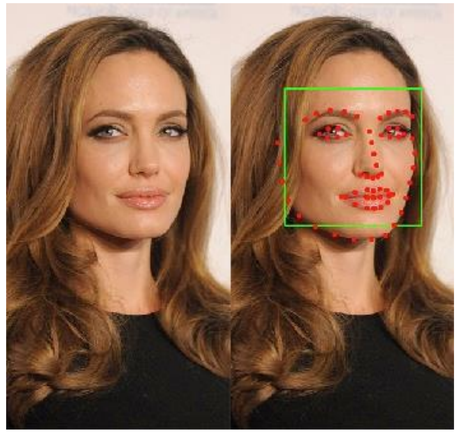
Arreola, et al. "Object Recognition and Tracking Using Haar-like Features Cascade Classifiers: Application to a Quad-Rotor UAV." ArXiv.org, 10 Mar. 2019, https://arxiv.org/abs/1903.03947.

Data Set

Training a Viola-Jones classifier from scratch can take a long time. For convenience , a pre-trained Viola-Jones classifier comes out-of-the-box with Open CV! You can use that one to see the algorithm in action. First you need to import OpenCv and load the image into the memory. Next, you need to load the Viola-Jones classifier. If you installed Open CV from source, it will be in the folder where you installed the Open CV library. Depending on the version, the exact path might vary, but the folder name will be Haarcascades, and it will contain multiple files. The one you need is called haarcascade_frontalface_alt.xml. If for some reason, your installation of Open CV did not get the pre-trained classifier, you can obtain it from the Open CV GitHub repo

Expectations

The output of the data set after everything is set right should look like this.



Conclusion

The Viola-Jones algorithm computes a lot of features including facial detection, which in reality helps facial tracking, facial recognition, and facial analysis. These features are widely used by many applications such as convenience stores, driverless car testing, daily medical diagnostics, human safety and security, and in monitoring the health of crops and livestock individuals. As with any data science analysis, this algorithm is computationally expensive.

Network Centrality Measures and Network Capacity

Luc Telemaque, Dr. Nadia Benakli (Mentor)
New York City College of Technology/City University of New York
Department of Mathematics

Abstract

Many real-life applications, including networks, are represented by graphs. Some examples of networks include: The World Wide Web, The Internet, Citation Networks, Biological Networks, Social Networks, and Transportation Networks. We are interested in the following two questions:

1. When and how to determine if a node of a (social) network important?
2. How can we manage the flow in a (transportation) network to maximize efficiency?

To answer the first question, we learn about centrality measures. To answer the second question, we study network capacity.

Introduction

Graph theory is the study of graphs to replicate the relationship between numerous objects. In mathematics, networks are often referred to as graphs, not to be confused with graphs of functions. A network is a set of objects called nodes that are connected together, and the connections between the nodes are called edges.

Examples of networks are friendship networks, food webs, and airline networks.

People studying social networks introduced centrality measures in order to find which individuals are best connected to others, and are the most important in the network. The answer to this question depends on what we mean by the word "important". According to Scott Adams, the creator of Dilbert, the power a person holds in the organization is inversely proportional to the number of keys on his keyring.

- A janitor has keys to every office, and yet has no power.
- The CEO does not need a key: people always open the door for him.

Centrality Measures

The degree centrality $C_D(v) = \deg(v)$

$\deg(v)$ is the degree of node which is the number of connections (edges) the node has.

Using this measure, a node is important if it has many connections.

The betweeness centrality $C_B(v) = \sum_{x \neq v \neq y \in V} \frac{g_{x,y}(v)}{g_{x,y}}$

$g_{x,y}$ the number of shortest paths between the nodes x and y, $g_{x,y}(v)$ the number of those paths that pass through v.

Using this measure, a node is important if it lies on a high portion of paths between other nodes in the network

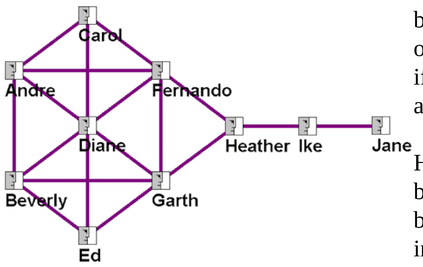
The closeness centrality $C_C(v) = \frac{1}{\sum_{u \neq v} d(u,v)}$

d(u,v) is the length of shortest path between the nodes u and v.

Using this measure, a node is important if its total distance from all other nodes is small, and therefore the node can communicate quickly with the other nodes

Centrality Measures in Social Networks

Social network analysis is the procedure of examining social connections utilizing networks in graph theory. It displays a visual representation of human connections and a mathematical analysis of human relationships. When a graph is plotted, we can use it to assess the person's importance and roles in the network such as finding the leaders, connectors, and isolates.



Kite Network of Friendships in a Classroom

Degree Centrality: Diane (0.667)

Betweenness Centrality: Heather (0.389)

Closeness Centrality: Fernando & Garth (0.643)

Diane has the highest degree centrality because she has the most direct links to her other classmates. Diane would be important if we were trying to spread information to as many people as possible the fastest.

Heather has the highest in-betweenness because she is essential to Ike and Jane being in the network. She would be important if we were trying to spread information to as many people as possible.

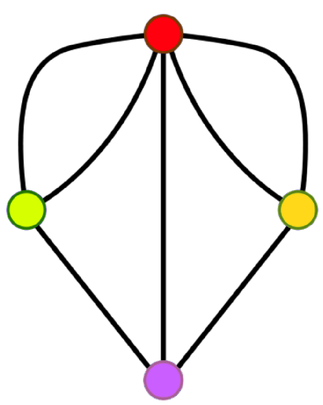
Fernando and Garth have the highest closeness centrality because they have the average shortest connection to every other person in the network. This would be important to monitor the flow of information through the network.

Seven Bridges of Königsberg

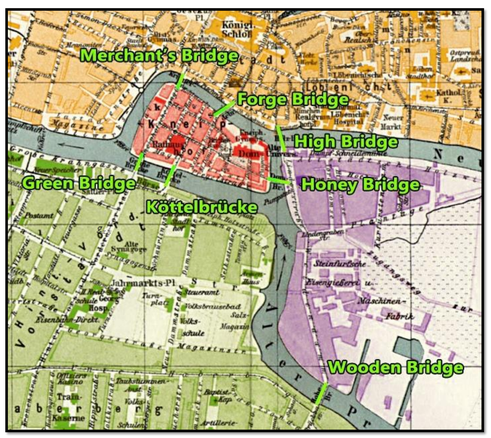
The first recorded proof of graph theory was the Königsberg bridge problem in 1735. Konigsberg was a city separated by a river with two islands in-between. Each landmass was connected to each other by seven bridges. The city wanted a figure out which path through the city where the traveler would cross each bridge once and only once.

Leonhard Euler, pioneered the theory of networks in mathematics when he proved that it was impossible to do so. This was done by constructing a graph where each segment of the city was a node and connected them using the 7 bridges as edges. The result was a simplified path through the city via each bridge. It was found that there was no possible way to go through the entire city by crossing every bridge only one time.

Try it yourself:



Graph of Königsberg



Map of Königsberg

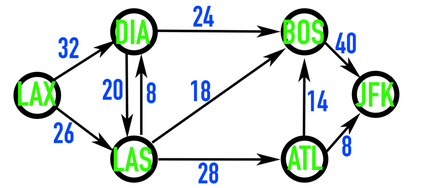
Maximum Flow

Network capacity the the maximum amount of traffic a network can resolve at any given moment. In graph theory, flow networks are directed graphs with a source, a sink, and several nodes connected with edges. Every edge has its own capacity which limits the flow through the edge. The conditions for a flow network are:

- If the node is not a sink, the flow input is equal the flow output.
- The total flow out of the source is equal to the total flow into the sink.

A directed graph is a graph made up of nodes and edges which have a direction representing its possible movement. Maximum flow is the greatest amount of flow that the network can allow to move from the source to the sink.

Imagine a scenario where NYCCT ordered two sets of books for the library from Los Angeles via airmail. However, on the route, the plane must stop at multiple destinations, but each flight can only hold a certain number of books. What is the largest amount of books can we order? This can be solved with multiple maximum flow solution algorithms.

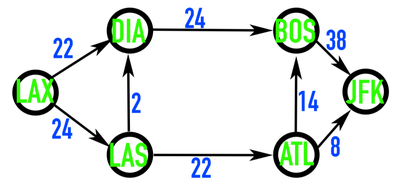


Flow Diagram with Maximum Capacities of books

This solution was solved using the Ford-Fulkerson algorithm.

Each node represents the airports the books arrive at on transit. The source is LAX and the sink is JFK.

Each directed edge is associated with a number called the weight. The weight is the maximum amount of flow an edge can take.



Graph of Maximum Possible Flow of Books

The solution states the maximum possible flow of books from LAX to JFK is 46.

Maximum capacity can also be used to schedule airline flights, economic demand and even street traffic problems.

Conclusions

Graph theory is a method of problem solving being implemented in new ways every day, opening a wide horizon of possibilities in the scientific field. Networks are found all around us, but it is up to us to see what entities of these networks are essential to the life of a system, and what is the most efficient and effective way to maximize these resources; all of which can be found using graph theory.

References

https://www.hackerearth.com/practice/algorithms/graphs/maximum-flow/tutorial/
http://www.orgnet.com/sna.html
https://www.amusingplanet.com/2018/08/the-seven-bridges-of-konigsberg.html
Newman, Mark EJ. "The structure and function of complex networks." *SIAM review* 45.2 (2003): 167-256.
Du, Donglei. "Social network analysis: Centrality measures." *Faculty of Business Administration, Lecture notes E3B 9Y2* (2019).