Home Assignment 4

Started: May 20 at 5:39pm

Quiz Instructions

General Directions to Assignment

This assignment aims to train the students to infer good features to be used in the ML models. To realize this assignment, I will provide:
- A dataset with a modified version of the apache log.
- A utility Python file with a set of methods to load a CSV file into a Pandas Dataframe.

The students need to read the content and follow the directions to realize the assignment and pay attention to what he needs to submit at the end of each task.

Support Files:

- **access_log_format-1.csv (https://experiencia21.tec.mx/courses/112876/files/48522642/download?download_frd=1)**

- **data_utils.py (https://experiencia21.tec.mx/courses/112876/files/48522658/download?download_frd=1)**

Answer Directions: The students need to deliver a Juniper notebook with the required method and the evidence of this use.

Basic Information about Apache Log

Log files are a handy tool for debugging issues within a web application. It is an essential source of information to system administrators and security analysts, enabling them to identify when a website is malfunctioning or presenting any security issues.

One specific log file used in debugging applications (or simply gaining insight into visitor activity) is the access log produced by an Apache HTTP server. Apache access log is one of several log files created by an Apache HTTP server. This log file is responsible for recording data for all requests processed by the Apache server. So, if an individual visits a webpage on your site, the access log file will contain details regarding this event.

This information is valuable in a variety of situations: for example, if a common request is failing for each individual trying to get to a particular web page, the link may be pointing to a page that no longer exists; if a specific page on the site is taking longer than it should to load, log entries could indicate SQL queries that could be refactored to improve performance; if one particular page on the site is trendy, aggregating data from access logs could shine a light on commonly requested resources, thus enabling businesses to increase their popularity by providing more related content.

For example, on the Ubuntu Linux distribution, access log records will be written to the following location by default: */var/log/apache2/access.log*.

Interpreting the Apache Access Logs

The Common Log Format is a standardized text file format used by various web servers in generating server log files, including the Apache HTTP server. A basic example of a register in this file is:

14.139.187.130,hahiss,hahiss@optonline.net,2017-01-01 02:16:51,GET / HTTP/1.1,200,10267,https://www.google.co.in/,"Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/55.0.2883.87 Safari/537.36",FR

- **139.187.130**: IP address of the client that made the request.
- **Hahiss:** user-id.
- **hahiss@optonline.net**: e-mail.
- **2017-01-01 02:16:51**: date and time of the request.
- **"GET /server-status HTTP/1.1"**: request type and resource being requested.
- **200**: HTTP response status code.
- **10267**: the size of the object returned to the client.
- **https://www.google.co.in/ (https://www.google.co.in/)**: destination address
- **"Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/55.0.2883.87 Safari/537.36"**: user-agent
- **FR**: the country where the request is originated.

Question 1 — 20 pts

In the real problems, a lot of dataset in the security is not labeled, or when it is, it is not reliable. Because of this situation, an important task of a Data Analyst is to identify features in the dataset and, using probability, try to infer labels to the dataset.

Using the provided unlabeled dataset, create a python method that calculates how likely is a connection from IP '14.139.187.130', and the US can make a transaction at 017-01-01 06:56:17?

Upload Choose a File

Question 2 — 40 pts

During the sessions, you saw that there are many different metrics to identify when you have an attack against a system. Develop a method that calculates the success transaction rate (response status code == 200) for 30 min. Following, develop another method that calculates the mean and standard deviation of this successful transaction rate. Evaluate the transactions using the following heuristic and set a new label in the successful transactions with two (suspicious) and three standard deviations (high-suspicious) above the mean.

Upload Choose a File

Question 3 — 40 pts

Using the previous result, use probability to calculate how likely each suspicious and high-suspicious transaction is to be legitimate.

Additional Explanation: To identify how is the probability that the previous classification is wrong, calculate for each entry the probability that the connection is legitimate, given the specific user-id, country, and IP that transaction is.

For example, considering that this transaction is classified as 'suspicious':

"68.180.228.229,noahb,noahb@hotmail.com,2017-01-01 03:07:34,GET /maccdc2012/signatures.log.gz HTTP/1.1,200,549,-,Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp (http://help.yahoo.com/help/us/ysearch/slurp)), US"

Using the full dataset to calculate the frequency, calculate this probability:

$$p(\text{normal}) = 1 - p(\text{suspicious} | \text{user}=u, \text{IP}=ip, \text{country}=c)$$

Following the general equation to describe conditional probability:

$$p(A | B,C) = (p(A,B,C) + \alpha) / (p(B,C) + \beta)$$

Considers: α=0.99850757 and β= 1.

Upload Choose a File

Not saved Submit Quiz