CENTER FOR CYBER SECURITY TRAINING

Applied Data Science for Cyber Security

Abstract

This interactive course will teach security professionals how to use data science techniques to quickly manipulate and analyze network and security data and ultimately uncover valuable insights from this data. The course will cover the entire data science process from data preparation, feature engineering and selection, exploratory data analysis, data visualization, machine learning, model evaluation and optimization and finally, implementing at scale—all with a focus on security related problems.

Learning Objectives

Participants will learn how to read in data in a variety of common formats then write scripts to analyze and visualize that data. Students will learn to:

- Writing scripts to efficiently read and manipulate CSV, XML, and JSON files.
- Quickly and efficiently parsing executables, log files, pcap and extracting artifacts from them.
- Making API calls to merge datasets.
- Use the Pandas library to quickly manipulate tabular data.
- Effectively visualizing data using Python.
- Preprocessing raw security data for machine learning and feature engineering.
- Building, applying and evaluating machine learning algorithms to identify potential threats.
- Automating the process of tuning and optimizing machine learning models.
- Hunting anomalous indicators of compromise and reducing false positives.

- Use supervised learning algorithms such as Random Forests, Naive Bayes, K-Nearest Neighbors (K-NN) and Support Vector Machines (SVM) to classify malicious URLs and identify SQL Injection.
- Apply unsupervised learning algorithms such as K-Means Clustering to detect anomalous behavior.

Finally, we will introduce the students to cutting edge Big Data tools including Apache Spark (PySpark), Apache Drill, and GPU accelerated parallel computing frameworks and demonstrate how to apply these techniques to extremely large datasets.

Target Audience

If you're in network security and have been thinking of getting into data science... this course is for you.

However, this course is also designed for data scientist who have wanted to get into Cybersecurity. We teach fundamental to advanced Data Science and how to apply it to Cybersecurity.

Course Outline

Introduction: Data Preparation with Pandas

- What is Pandas and why use it?
- The Series, DataFrame, and Panel objects
- The Pandas ecosystem: Scikit-learn, Seaborn, Bokeh

Vectorized Computing in One Dimension: The Series Object

- Creating a series
- Describing data
- Filtering data
- Other operations on data
- **Activity: Worksheet**

Vectorized Computing in Two Dimensions: The DataFrame

- Creating a DataFrame
- Reading logfiles, APIs and other sources
- Manipulating data in data frames

- Applying functions to data frames
- Aggregating data in data frames
- **Activity: DataFrame Worksheet**

Statistical Summaries

- 5-Number summaries
- Normalizing data
- Understanding Distributions
- Correlations
- Confidence Intervals and P-Values
- **Exercise: Complete EDA Worksheet**

Concepts of Data Visualization

- Creating effective visualizations
- Choosing the correct visualization
- Using visualization to explore data

Practical Data Visualization

- Using Matplotlib to create basic charts
- Overview of advanced charts with Seaborn
- **Exercise: Data Visualization Worksheet**

Introduction to machine learning

- An overview of machine learning; solving security problems with machine learning; where does machine learning fit in your security ecosystem?

The machine learning process

- Prework (thinking through the question, figuring out the data);
- Exploring your data
- Gathering and cleaning the data
- Engineering and selecting features selecting, building, and training your model
- Evaluating your model’s performance
- Putting your model in production

Feature engineering and selection

Selecting, preparing, and visualizing features

Classification models

- Logistic regression
- k-NN classifier
- Decision trees
- Random forest
- **Hands-on exercise**: Build a classifier to classify malicious URLs

Evaluating Model Performance

- Understanding Performance Metrics: Accuracy, precision, and recall
- Visualizing confusion matrices and classification reports

Fine-tuning your model

- Grid search for hyper parameter tuning;
- Model selection and evaluation
- Pipelines
- **Hands-on exercise**: Tuning your model

Advanced topics

- Neural nets and their applications to security
- Deep learning case studies
- Hunting with data science

Overview of Unsupervised Learning

Measuring distances

- Euclidian distances
- Other distance functions

Clustering Algorithms

- K-Means
- DBSCAN

Evaluating performance of Unsupervised Models

- Performance metrics for clustering...a little harder;
- Using Yellowbrick to visualize model performance
- **Hands-on exercise**: Detecting anomalies using clustering techniques

REQUIREMENTS

Student Machine/ Laptop Requirements

Center for Cyber Security Training provides a windows based virtual machine for each student. All exercises are performed in that environment.

We recommend at least 8GB of ram and the students will need 30GB of available hard disk space.

Students Knowledge Pre-Requisites:

- Fundamentals Object-Oriented Programming
- All Skill-Levels (Challenges for Beginner to Pro Programmers)
- All Backgrounds (From IT, Cyber SME to Manager)