AGENDA - PART 1

10min  INTRO + SETUP
30min  ML CRASH COURSE
30min  ML IDS/WAF PLAY TIME
40min  ML IDS/WAF POISONING
10min  BREAK
AGENDA - PART 2

20min  DEEP LEARNING PRIMER
30min  PWNING DEEP LEARNING SYSTEMS
30min  SPEAR PHISHING WITH DEEP LEARNING
40min  SECRET PROJECT ;}
Let $f$ be a function representing this training session. Let $x$ be the amount of math involved.

$$\lim_{x \to 0} f(x)$$
OUR TOOLZ

- scikit-learn - implements a range of machine learning algorithms
- TensorFlow - library for numerical computation using flow graphs / deep learning
SCIKIT-LEARN

- easy-to-use, general-purpose toolbox for machine learning in Python.
- supervised and unsupervised machine learning techniques.
- Utilities for common tasks such as model selection, feature extraction, and feature selection
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license
TENSORFLOW

- Open source, originally Google Brain
- Widely used in both research and production
- Main popular use is for deep learning/neural nets
  - But not restricted to just deep models
- Multiple GPU Support
YOU NEED THESE

**NumPy**
Base N-dimensional array package

**SciPy library**
Fundamental library for scientific computing

**Matplotlib**
Comprehensive 2D Plotting

**IP[y]: IPython**
Enhanced Interactive Console

**Sympy**
Symbolic mathematics

**pandas**
Data structures & analysis
CROSS VALIDATION
MACHINE LEARNING 101

Types of machine learning use cases:

- Regression
- Classification
- Anomaly detection
- Recommendation

won’t cover here, but check out this talk

This covers EVERYTHING. (almost)
LINEAR REGRESSION

2d linear regression
POLYNOMIAL REGRESSION

regression line
SOLVING
SOLVING

\[ z = \sin\left(\frac{1}{2}x^2 - \frac{1}{4}y^2 + 3\right)\cos(2x + 1 - \exp(y)) \]
LINEAR/LOGISTIC REGRESSION

- /home/ml/Desktop/intro/00-linear-regression.ipynb
- /home/ml/Desktop/intro/01-logistic-regression.ipynb
ANOMALY DETECTION
ANOMALY DETECTION

- **Outliers vs. novelties**
  - **novelties**: unobserved pattern in new observations not included in training data

- **Simple statistics/forecasting methods**
  - Exponential smoothing, Holt-Winters algorithm

- **Machine learning methods**
  - Elliptical envelope, density-based, clustering, SVM
ANOMALY DETECTION

- /home/ml/Desktop/intro/02-anomaly-detection.ipynb
- /home/ml/Desktop/anomaly/anomaly-detection-eg.ipynb
CLASSIFICATION

labeled data - do you have it?
CLASSIFICATION

yes! lots!

supervised learning

no :(

unsupervised learning

only a little bit

(semi-supervised learning)
SUPERVISED CLASSIFICATION

● Many different algorithms!
● We will go through five:
  ○ Naive Bayes
  ○ K-nearest neighbors
  ○ Support Vector Machines
  ○ Decision Trees
BAYES THEOREM

The probability of an event happening is based on prior knowledge of conditions that might be related to the event.

- $P(U)$: 0.5%
- $P(U \cap +)$: 99% (0.495%)
- $P(U \cap -)$: 1% (0.005%)
- $P(\bar{U})$: 99.5%
- $P(\bar{U} \cap +)$: 0.995%
- $P(\bar{U} \cap -)$: 98.505%
- $P(\bar{U} | U)$: 1%
- $P(\bar{U} | \bar{U})$: 99%
NAIVE BAYES CLASSIFIER

- SUPERVISED LEARNING
CLASSIFYING SPAM
THE DATASET: 2007 TREC Public Spam Corpus

http://plg.uwaterloo.ca/~gvcormac/treccorpus07/
MULTICLASS CLASSIFICATION

2 ways to do it:

- 1-vs-rest
  - 1 binary classifier per class
- 1-vs-1
  - 1 binary classifier per pair of classes
  - $K^*K^*/2$ classifiers for a K-class problem
K-NEAREST NEIGHBORS CLASSIFIER (kNN)
SUPPORT VECTOR MACHINES (SVM)

/home/ml/Desktop/intro/03-svm.ipynb
DECISION TREE CLASSIFIER

visualization
UNSUPERVISED CLASSIFICATION

- Mainly refers to clustering

- Four types:
  - Centroid: K-Means
  - Distribution: Gaussian mixture models
  - Density: DBSCAN
  - Connectivity: Hierarchical clustering
K-MEANS CLUSTERING

- UNSUPERVISED LEARNING

/home/ml/Desktop/intro/04-kmeans-pca.ipynb
SO MANY ALGORITHMS.
HOW TO PICK. ???????????
XSS DETECTION
CLASSIFYING PACKETS

data

/home/ml/Desktop/network/kdd-packet-classification.ipynb
CLASSIFYING MALWARE (with static PE features)
pefile dump

----------FILE_HEADER----------
[IMAGE_FILE_HEADER]
Machine:        0x14C
NumberOfSections:   0x4
TimeDateStamp:    0x851C3163
[INVALID TIME]
PointerToSymbolTable: 0x74726144
NumberOfSymbols:  0x455068
SizeOfOptionalHeader: 0xE0
Characteristics:  0x818F

----------OPTIONAL_HEADER----------
[IMAGE_OPTIONAL_HEADER]
Magic:           0x10B
MajorLinkerVersion: 0x2
MinorLinkerVersion: 0x19
SizeOfCode:       0x200
SizeOfInitializedData: 0x45400
SizeOfUninitializedData: 0x0
AddressOfEntryPoint: 0x2000
BaseOfCode:       0x1000
BaseOfData:       0x2000
ImageBase:        0xDE0000
SectionAlignment: 0x1000
FileAlignment:    0x1000
MajorOperatingSystemVersion: 0x1
MinorOperatingSystemVersion: 0x0

----------FILE_HEADER----------
[IMAGE_FILE_HEADER]
Machine:        0x14C
NumberOfSections:   0x4
TimeDateStamp:    0x851C3163
[INVALID TIME]
PointerToSymbolTable: 0x74726144
NumberOfSymbols:  0x455068
SizeOfOptionalHeader: 0xE0
Characteristics:  0x818F

----------OPTIONAL_HEADER----------
[IMAGE_OPTIONAL_HEADER]
Magic:           0x10B
MajorLinkerVersion: 0x2
MinorLinkerVersion: 0x19
SizeOfCode:       0x200
SizeOfInitializedData: 0x45400
SizeOfUninitializedData: 0x0
AddressOfEntryPoint: 0x2000
BaseOfCode:       0x1000
BaseOfData:       0x2000
ImageBase:        0xDE0000
SectionAlignment: 0x1000
FileAlignment:    0x1000
MajorOperatingSystemVersion: 0x1
MinorOperatingSystemVersion: 0x0

----------DOS_HEADER----------
[IMAGE_DOS_HEADER]
e_magic:         0x5A4D
e_cblp:          0x50
e_cp:            0x2

----------NT_HEADERS----------
[IMAGE_NT_HEADERS]
Signature:       0x4550

----------PE Sections----------
[IMAGE_SECTION_HEADER]
Name:            CODE
Misc:            0x1000
Misc_PhysicalAddress: 0x1000
Misc_VirtualSize:   0x1000
VirtualAddress:    0x1000
SizeOfRawData:     0x1000
PointerToRawData:  0x1000
PointerToRelocations: 0x0
PointerToLinenumbers: 0x0
NumberOfRelocations: 0x0
NumberOfLinenumbers: 0x0
Characteristics:   0xE0000020
Flags: MEM_WRITE, CNT_CODE, MEM_EXECUTE, MEM_READ
Entropy: 0.061089 (Min=0.0, Max=8.0)

[IMAGE_SECTION_HEADER]
Name:            DATA
Misc:            0x45000
Misc_PhysicalAddress: 0x45000
Misc_VirtualSize:   0x45000
VirtualAddress:    0x2000
SizeOfRawData:     0x45000
PointerToRawData:  0x45000
PointerToRelocations: 0x0
PointerToLinenumbers: 0x0
NumberOfRelocations: 0x0
NumberOfLinenumbers: 0x0
Characteristics:   0xC0000040
Flags: MEM_WRITE, CNT_INITIALIZED_DATA, MEM_READ
Entropy: 7.980693 (Min=0.0, Max=8.0)

----------Parsing Warnings----------
Suspicious NumberOfRvaAndSizes in the Optional Header. Normal values are never larger than 0x10, the value is: 0xdfffddde

Error parsing section 2. SizeOfRawData is larger than file.
FEATURE VECTOR

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>memtest.exe</td>
<td>631ea355665f28d4707448e442fb5b8</td>
<td>332</td>
<td>224</td>
<td>258</td>
<td>9</td>
<td>0</td>
<td>361984</td>
<td>115712</td>
<td>0</td>
<td>6135</td>
<td>4096</td>
<td>372736</td>
<td>4194304</td>
<td>4096</td>
<td>512</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1036</td>
<td>288</td>
<td>1024</td>
<td>485887</td>
<td>16</td>
<td>1024</td>
<td>1048576</td>
<td>4096</td>
<td>1048576</td>
<td>16</td>
<td>8</td>
<td>5.7668065537</td>
<td>3.60742957555</td>
</tr>
</tbody>
</table>

| VirusShare_76c2574c22b44f69e3ed519d36bd8dff | 76c2574c22b44f69e3ed519d36bd8dff | 332 | 224 | 258 | 10 | 0 | 28672 | 445952 | 16896 | 14819 | 4096 | 32768 | 4194304 | 4096 | 512 | 5 | 0 | 6 | 5 | 0 | 39777216 | 1024 | 680384 | 2 | 34112 | 1048576 | 4096 | 1048576 | 0 | 0 | 0 | 0 | 6.49788465186 | 30634.666667 | 0 | 139264 | 661773.333333 | 3978 | 3362816 | 8 | 172 | 1 | 0 | 21 | 3.42072662405 | 1.86523352037 | 7.9688495098 | 6558.42857143 | 180 | 676 | 24 | 0 | 0 |

legitimate:

malware:
THANK YOU

@cchio
cchio@cs.stanford.edu

@antojosep007
antojoseph007@gmail.com
THANKS

sign up for updates!

mlsec@cs.stanford.edu