Cake and Grief Counseling Will be Available:

Using Artificial Intelligence for Forensics Without Jeopardizing Humanity

Jesse Kornblum

Outline

- Introduction
- Artificial Intelligence
- Spam Detection
- Clustering
- Classification
- Collaborative Filtering
- Questions

KYRUS

Introduction

- Analyzing an infinite number of programs, documents
  - Only five minutes per sample
- Which of them are similar to each other?
- Which of them fit into existing categories?
  - Zeus variant
  - Related to the last problem set
  - Related to the Henderson account?

KYRUS

Assumptions

- There are too many programs to manually review
  - You're not going to look at them all
- Computers are good at computing
  - Humans are not
- Humans are good at categorizing
  - Computers are not

KYRUS

Artificial Intelligence

Artificial Intelligence

= Mathy Stuff

Mathy Stuff, an Example

- Spam Detection
  - Big problem
  - Humans are really good at it
- Reduce to math problem for computers
- Classification problem
  - For each item, is it { spam, ham }

KYRUS

Email Features

- Anything can be a feature
  - Sender
  - Recipient
  - Headers
  - Phrases of text
- Derived features
  - Valid rDNS
  - Sender in address book
  - % of words spelled wrong

Image courtesy of Flickr user doctor_keats and used under Create Commons license.

KYRUS

Naïve Bayes Classification

- Determine which is greater
  - P(spam) or P(ham)
  - P(spam given features) or P(ham given features)

Training Set

- Get a set of emails
- Human labels which are spam, which are ham

SPAM

From: rxbsgw56@qquix.biz
To: jessek@kyr.us
Subject: V1agra!!

T0p quailikty V1aagra delieverd direct to you!

http://sales.v1agara.biz/

HAM

From: mom@aol.com
To: jessek@kyr.us
Subject: Wear a jacket

It's going to be cold while you're in Atlanta. Please wear a jacket so that I don't worry about you.

Love,
Mom

KYRUS

Naïve Bayesian Classifier

- Based on Bayes Theorem
- Probabilities are based on what‘s in the training set

$$P(B|A) = \frac{P(B) * P(A|B)}{P(A)}$$

$$P(spam|f) = \frac{P(s) * P(f|s)}{P(f)}$$

- In other words, count things in the training set and do math on them

KYRUS

Naïve Bayesian Classifier

$$P(spam|f) = \frac{P(s) * P(f|s)}{P(f)}$$

$$P(ham|f) = \frac{P(h) * P(f|h)}{P(f)}$$

Which probability is greater?

Clustering

- Clustering
  - Group together similar things

- Choose a distance metric
- Compute distances between all items
- Group items where distance is less than threshold

Similarity

- What does it mean for two things to be similar?

Similarity

- Depends on:
  - The kind of things be compared
  - How they're being compared

- What makes two programs similar?

Similar Programs

- Do the same thing
- Have the same look and feel
- Connect to the same servers
- Written by the same person
- Used in the same intrusion
- Found in the same problem set

KYRUS

Current Tools

- Cryptographic Hashing
  - Exact match
  - e.g. MD5
- Fuzzy Hashing
  - Similar ones and zeros
- Ad hoc analysis
- Reverse Engineering

KYRUS

sdhash

- Algorithm for matching similar files
- Developed by Dr. Vassil Roussev, University of New Orleans

- Like ssdeep, ignores file type data
- Variable sized hashes
  - About 3% of input size
- Handles data reordering

- Matches more files than ssdeep
  - There is not, technically, “better”
- Code, paper, roadmap:
  - http://roussev.net/sdhash/

KYRUS

sdhash

Distance Metrics

Distance Metrics

Distance Metrics

Distance Metrics

- What's a good distance metric for programs?
- Or: What makes programs similar?

Features

- Signed code?
  - Signed by whom
- Which APIs are called
- How often APIs are called
- Order in which APIs are called
- Entropy
- DLLs used
- Percentage of code coverage
- Magic strings
- N-grams of instructions
- Control-flow graph
- IP addresses accessed
- …

KYRUS

Computing Clusters

- Extract features from all inputs
- Compute distance metric for all pairs of inputs

```
For all inputs a and b:
    if distance(a,b) < threshold
        add_cluster(a,b)
```

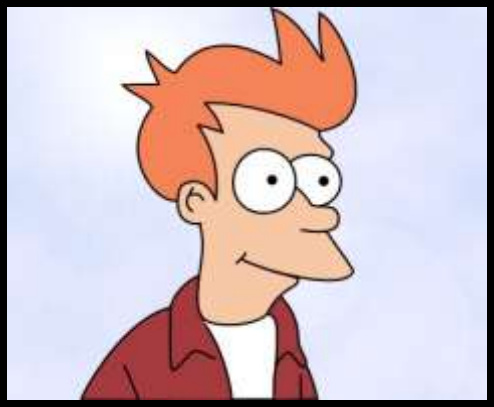
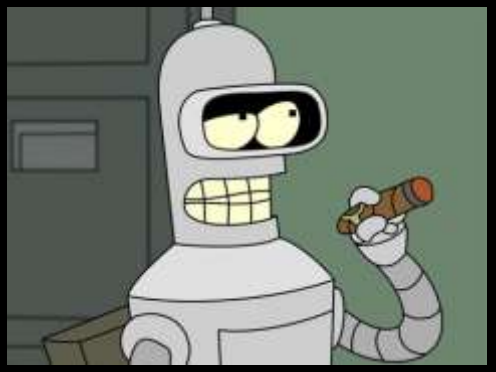
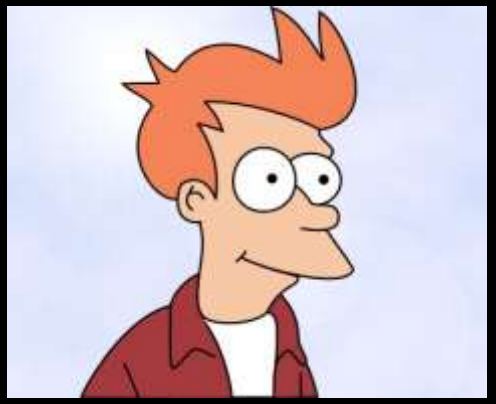
- Exclusive vs. Non-Exclusive clustering
  - Assume A~B and B~C
  - Exclusive: {A,B,C}
  - Non-Exclusive: {A,B} {B,C}

KYRUS

Not Just for Programs

- eDiscovery is all over this
- Commercially available now
- Uses phrases of text as features
  - File format independent
- Distance metrics are statistical or linguistic or …

Clustering vs. Classification

	Clustering	Classification
What Makes Things Similar?		
Which Things are Similar?		

KYRUS

Classification

- Also known as:
  - Predictive Coding
  - Assisted Machine Learning

- Put inputs into a category
  - Zeus variant or Not Zeus variant
  - Related to the Henderson account or not

KYRUS

Classification

- User must create a set of training data
  - HINT: Results of clustering work here!
- Must identify some inputs for each possible outcome
- The more the better

Classification

- Artificial intelligence is just math

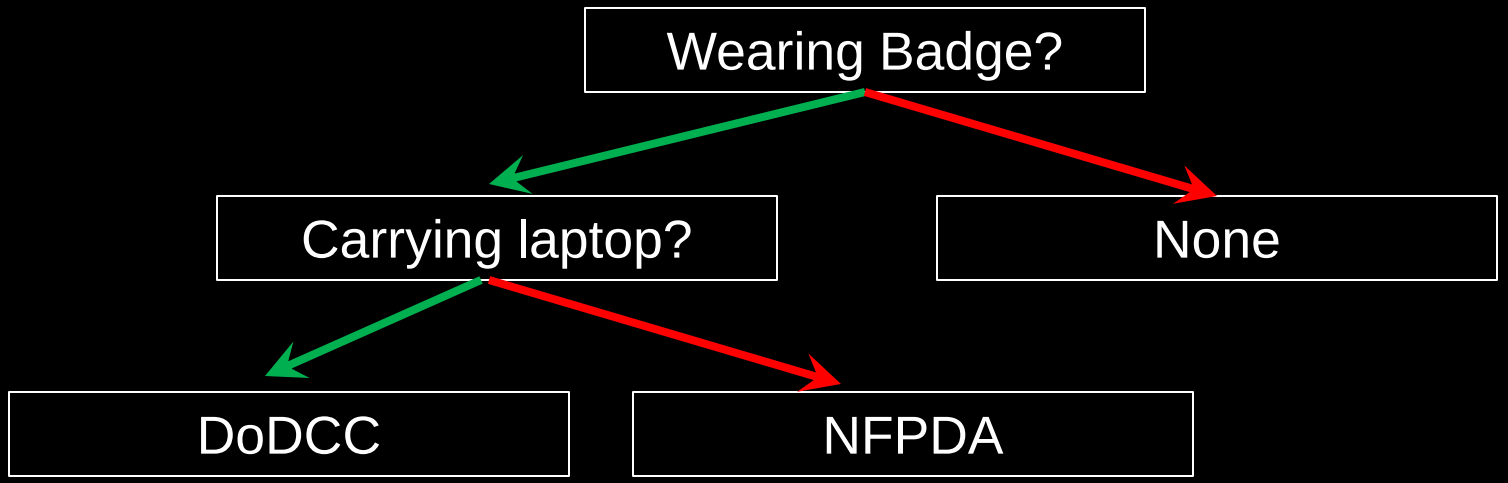
- There are many algorithms:
  - Naïve Bayesian classifier
  - K-Nearest Neighbor
  - Locality Sensitive Hashing
  - Decision Trees
  - Neural Networks
  - Hidden Markov Models

- See Wikipedia article on Classification (machine learning)

KYRUS

Decision Tree

- Build a flowchart of questions on the features
- Each question should divide the data equally
- Which conference is this person attending?
  - DoDCC, NFPDA, None



Decision Tree

- Quick to classify, but slow to construct
- What questions are best at which point in the tree?

- You could make a career out of efficient decision tree generation
  - And people do

Feature Selection

- The Curse of Dimensionality
  - So many dimensions (features) that comparisons become too time consuming or too complex

- No problem
- Select the “important” features
  - (Insert mathy stuff here)

- Example:
  - Presence of crypto constants
  - Depends on context

KYRUS

Classifier Evaluation

- Obviously our decision tree is not perfect
- There are several metrics of classifier performance

- Precision and Recall

- Precision measures false positives
  - P = TP / (TP + FP)

- Recall measures false negatives
  - R = TP / (TP + FN)

- Both are on a scale from zero to one
  - One being perfect

KYRUS

Classifier Evaluation

- Known error rate

- Remember our assumption
  - “There is too much data to manually review”

Classification Packages

- All of these are Free and Open Source:
  - Weka
  - Apache Mahout
  - Malheur
  - LibSVM

- Each package has several modules
- Which is the best?

Classification Systems

- Academia
  - “Solved problem”
- eDiscovery
  - They love this stuff
  - Predictive coding
  - Features are n-grams of text
- For you?
  - Some assembly required
  - Your Agency puts it together

KYRUS

Collaborative Filtering

Customers like you who bought this item also bought

Collaborative Filtering

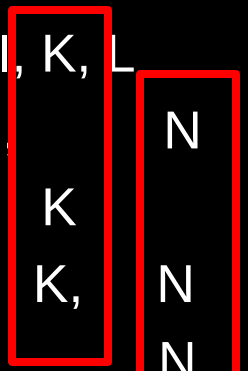
- Looking at Item A

- You've bought:
  - B, C, D, E, F
- Others:
  - D, H, I, K, L
  - C, I, N
  - B, D, F, K
  - B, D, E, F, K, N
  - C, D, N

Collaborative Filtering

- Looking at Item A

- You've bought:
  - B, C, D, E, F
- Others:
  - D, H, I, K, L
  - C, I, N
  - B, D, F, K
  - B, D, E, F, K, N
  - C, D, N



Collaborative Filtering

- Used to predict
  - Human preferences
  - Human patterns
- Forensics is about studying artifacts to infer a human action

- People who have attacked these hosts also attacked...
- People who made events like those in your timeline also did…

Conclusion

- Analyzing an infinite number of programs, documents
  - Only five minutes per sample
  - Use AI. It's math, and computers are good at math
- Which of them are similar to each other?
  - Build clusters
- Which of them fit into existing categories?
  - Zeus variant
  - Related to the last problem set
  - Related to the Henderson account?
  - Build classifiers for these categories

KYRUS

Questions?

Jesse Kornblum
jessek@kyr.us

KYRUS