Final Projects

Choose your destiny

- Measuring distance of HMM Models
  How distant are the models of each sample? [Unsupervised]

- Boosting HMM
  Testing different boosting techniques to enhance accuracy

- Clustering for ciphertext classification
  Let’s tell apart encryption techniques [Unsupervised]

- Featuring Deep Learning in the Generality vs Accuracy dilemma
  How robust Deep Learning really is?

- HMM2Vec to conquer the world
  “Individual” B matrices as feature vectors to enhance “old-style” classification

Measuring distance of HMM Models

- Similar dataset to the previous research work
- This can be applied to anything (malware detection, etc…)
- The idea is to generate a model per each sample
- Then, measuring the distance among them to “cluster” families together

Boosting HMM

- A continuation of the work done in Midterm#2
- The idea is to rely on “random restarts” to enhance the accuracy obtained in your Midterm#2 experiments
- Will boosting be more accurate?

Boosting HMM

- What you should have learned from the Midterm#2
  Best value for M and N

- How many restarts?
  Many

- Which experiments to perform?
  The same as in Midterm#2

- Can I combine boosting and bagging [stacking]?
  Oh, well, that would be great

Clustering for ciphertext classification

- For this project you require a clustering library

- Also, you need a dataset of encrypted texts

You can start with two ciphers:

Simple Substitution

http://practicalcryptography.com/ciphers/simple-substitution-cipher/

Columnar Transposition

http://www.practicalcryptography.com/ciphers/columnar-transposition-cipher/

Clustering for ciphertext classification

How will you build the clusters?

- You need information taken from the ciphertext

Possible candidates:

- Entropy
- Index of Coincidence (IC)
- HMM score from model trained on the cipher
- HMM score from model trained on the English language

➢This would generate a 4-dimensional space

More?

Clustering for ciphertext classification

1st experiment

- The plaintext should remain constant
- The key should change per each sample

2nd experiment

- The plaintext should change per each sample
- The key should remain constant

Clustering for ciphertext classification

- Extrinsic or Intrinsic?

  Purity and Silhouette coefficient (cohesion and separation)

- How many dimensions?

  At least 2 or 3

- How many ciphers?

  At least the two proposed

Featuring Deep Learning in the Generality vs Accuracy dilemma

- Applicable with ease on the malware detection problem
- The idea is to generate a multi-family model for multi-class classification
- Vanilla version has been done already
  We got promising results!
- Here, we want to “stack” together scores from HMM, SVM, and maybe more

HMM2Vec to conquer the world

- Brand new algorithm

  We will show it to the world in January 2021

- The idea is to focus on the B matrices to furnish input to other classifiers

- Results have been compared to Word2vec (Natural Language Processing algorithm)

  But the new heavy weight world champ is Google BERT...

- Ideas to create a control-loop: HMM2Vec2HMM crazyness!