

Aprendizagem Aplicada à Segurança

Mário Antunes

September 22, 2023

University of Aveiro

Table of Contents

Class Introduction

Grading

Class Schedule

Environment

Bibliography

Name: Mário Antunes

E-Mail: *mario.antunes@ua.pt*

Office: 19.2.15 (IT1)

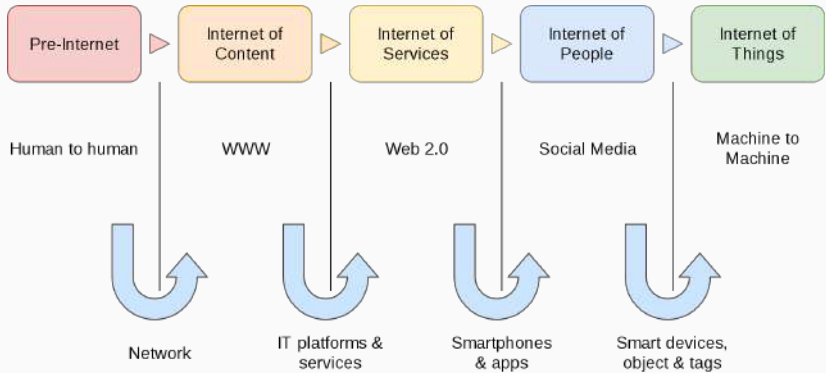


Class Introduction

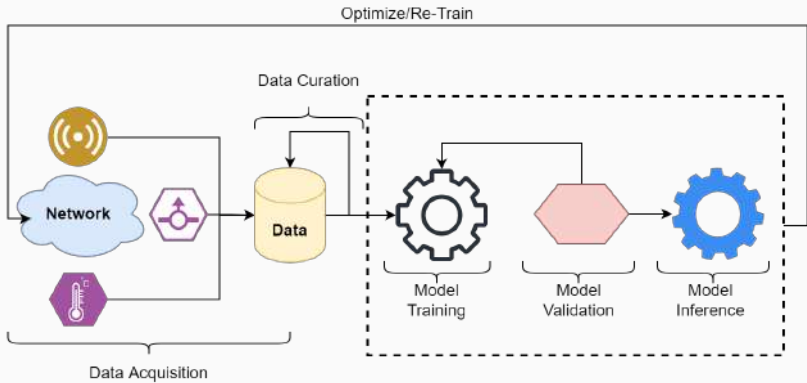
- Given the evolution of the threats
- And the complexity of the systems
- AI/ML are gaining traction as a useful tool



Class Introduction



Class Introduction



- 50% Theory + 50% Practice
- Discrete: 25% Mid-term Exam + 25% Final Exam + 20% Project Idea + 30% Project
- Final: 50% Final Exame + 50% Project

Class Schedule i

Date	Class	Topic
15/09/2023	1	Introduction
22/09/2023	2	
29/09/2023	3	SPAM Detector
06/10/2023	4	
13/10/2023	5	
20/10/2023	6	Anomaly Detection
27/10/2023	7	
03/11/2023	8	Mid-term Exam
10/11/2023	9	
17/11/2023	10	Malware Analysis
24/11/2023	11	
01/12/2023	12	
08/12/2023	13	Project
15/12/2023	14	
22/12/2023	15	



- All of the books are available here:
[*https://learning.oreilly.com/*](https://learning.oreilly.com/)

- [1] S. Halder and S. Ozdemir, *Hands-On Machine Learning for Cybersecurity: Safeguard your system by making your machines intelligent using the Python ecosystem*. Packt Publishing Ltd, 2018.
- [2] C. Chio and D. Freeman, *Machine Learning and Security*. O'Reilly, 2018.

- [3] A. Parisi, *Hands-On Artificial Intelligence for Cybersecurity: Implement smart AI systems for preventing cyber attacks and detecting threats and network anomalies*. Packt Publishing Ltd, 2019.
- [4] E. Tsukerman, *Machine Learning for Cybersecurity Cookbook*. Packt Publishing Ltd, 2019.
- [5] J. P. Mueller and R. Stephens, *Machine Learning Security Principles*. Packt Publishing Ltd, 2019.

Aprendizagem Aplicada à Segurança

Mário Antunes

September 22, 2023

University of Aveiro

Table of Contents

Class Introduction

Grading

Class Schedule

Environment

Bibliography

Name: Mário Antunes

E-Mail: *mario.antunes@ua.pt*

Office: 19.2.15 (IT1)

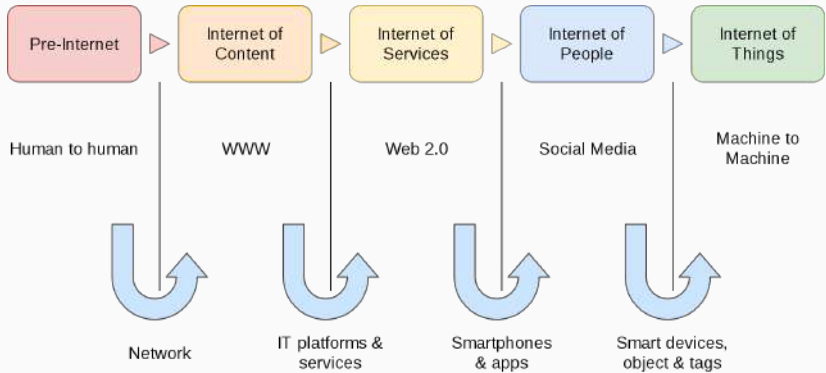


Class Introduction

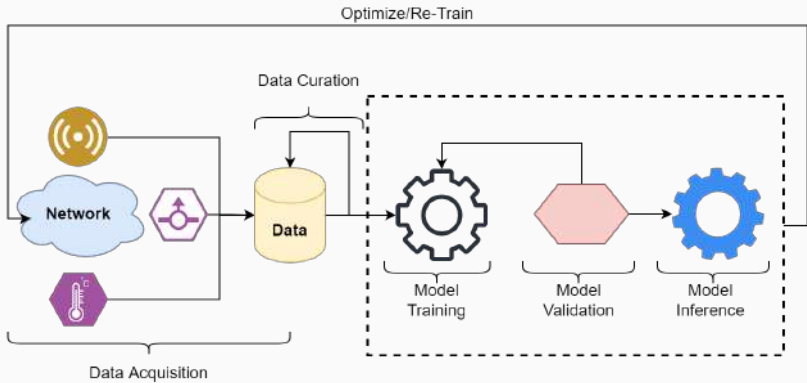
- Given the evolution of the threats
- And the complexity of the systems
- AI/ML are gaining traction as a useful tool



Class Introduction



Class Introduction



- 50% Theory + 50% Practice
- Discrete: 25% Mid-term Exam + 25% Final Exam + 20% Project Idea + 30% Project
- Final: 50% Final Exame + 50% Project

Class Schedule i

Date	Class	Topic
15/09/2023	1	Introduction
22/09/2023	2	
29/09/2023	3	SPAM Detector
06/10/2023	4	
13/10/2023	5	
20/10/2023	6	Anomaly Detection
27/10/2023	7	
03/11/2023	8	Mid-term Exam
10/11/2023	9	
17/11/2023	10	Malware Analysis
24/11/2023	11	
01/12/2023	12	
08/12/2023	13	Project
15/12/2023	14	
22/12/2023	15	



- All of the books are available here:
[*https://learning.oreilly.com/*](https://learning.oreilly.com/)

- [1] S. Halder and S. Ozdemir, *Hands-On Machine Learning for Cybersecurity: Safeguard your system by making your machines intelligent using the Python ecosystem*. Packt Publishing Ltd, 2018.
- [2] C. Chio and D. Freeman, *Machine Learning and Security*. O'Reilly, 2018.

- [3] A. Parisi, *Hands-On Artificial Intelligence for Cybersecurity: Implement smart AI systems for preventing cyber attacks and detecting threats and network anomalies*. Packt Publishing Ltd, 2019.
- [4] E. Tsukerman, *Machine Learning for Cybersecurity Cookbook*. Packt Publishing Ltd, 2019.
- [5] J. P. Mueller and R. Stephens, *Machine Learning Security Principles*. Packt Publishing Ltd, 2019.

Aprendizagem Aplicada à Segurança

Mário Antunes

September 22, 2023

University of Aveiro

Table of Contents

SPAM

SPAM Detection

Binary Classification

Text Mining

Natural Language Processing (NLP)

Classification Model

Model Evaluation

- The term “spam” is internet slang that refers to unsolicited commercial email (UCE).
- The first reported case of spam occurred in 1898, when the New York Times reported unsolicited messages circulating in association with an old swindle.
- The term “spam” was coined in 1994, based on a now-legendary Monty Python’s Flying Circus sketch, where a crowd of Vikings sings progressively louder choruses of “SPAM! SPAM! SPAM!”

SPAM



Dear Sir,

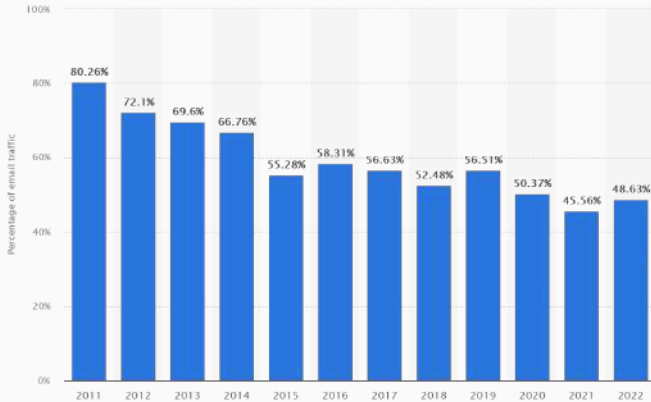
I am prince [REDACTED] from Nigeria. Your help would be very appreciated. I want to transfer all of my fortune outside if Nigeria due to a frozen account, If you could be so kind and transfer small sum of 3 500 USD to my account, I would be able to unfreeze my account and transfer my money outside of Nigeria. To repay your kindness, I will send 1 000 000 USD to your account.

Please contact me to proceed

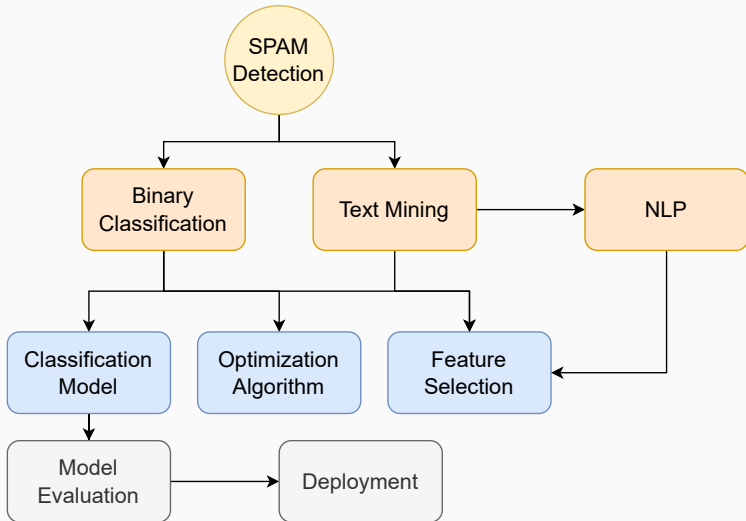
Prince [REDACTED]

- Huge list of *https://en.wikipedia.org/wiki/Anti-spam_techniques*
- From common sense to *Bayesian spam filtering*
- Unfortunately it is a costly battle

Fight against SPAM

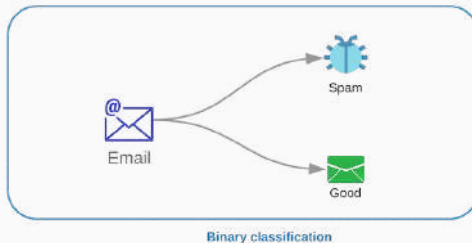


SPAM Detection



Binary Classification

- Binary classification is the task of classifying the elements of a set into two groups (each called class) on the basis of a classification rule.
- For this application one message can either be spam or ham.



- Text mining is the process of deriving high-quality information from text.
- Combines concepts from Machine Learning, Linguistic and statistical analysis.
- In this area we will explore the methods used to rank words/tokens and the BoW model.

Bag of Words (Bow) model

	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

Natural Language Processing (NLP)

- NLP gives the computers the ability to understand text.
- Combines *Syntax* and *Semantic* into the analysis.
- One famous examples are the Large Language Models (LLMs) that power OpenAI Chat GPT.

Classification Model

- SPAM detection is “considered” a toy example.
- As such, we will explore two of the simplest learning models: Naive Bayes and Logistic Regression.

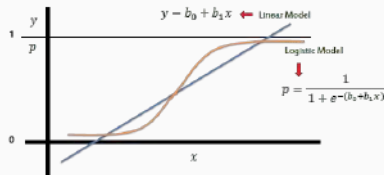
Likelihood of the Evidence given that the Hypothesis is True

Prior Probability of the Hypothesis

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

Posterior Probability of the Hypothesis given that the Evidence is True

Prior Probability that the evidence is True



Model Evaluation

- Classification model can be evaluated using a confusing matrix
- The simplest methods to evaluate a model is through accuracy: $acc = \frac{TP+TN}{TP+TN+FP+FN}$

	Predicted Positive	Predicted Negative	
Actual Positive	TP <i>True Positive</i>	FN <i>False Negative</i>	Sensitivity $\frac{TP}{(TP + FN)}$
Actual Negative	FP <i>False Positive</i>	TN <i>True Negative</i>	Specificity $\frac{TN}{(TN + FP)}$
	Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Model Evaluation

