

Machine Learning Project

Malicious URL detection

Navneet Agarwal¹

Sohum Dhar²

Tanmay Parekh³

C Vishvesh⁴

¹Department of Computer Science
IIT Bombay

Project Presentation, 2016

Outline

Description

- Problem Statement

- Motivation

Approach and Implementation

- Approach Towards the Problem

- Support Vector Machines

- Perceptron

- Neural Networks!!!

Future Work

- Future Work

Outline

Description

Problem Statement

Motivation

Approach and Implementation

Approach Towards the Problem

Support Vector Machines

Perceptron

Neural Networks!!!

Future Work

Future Work

Problem Statement

- ▶ Malicious web sites are a cornerstone of Internet criminal activities. As a result, there has been a broad interest in developing system to prevent the end user from visiting such sites.

Problem Statement

- ▶ Malicious web sites are a cornerstone of Internet criminal activities. As a result, there has been a broad interest in developing system to prevent the end user from visiting such sites.
- ▶ We aim to explore various learning approaches for detecting malicious web sites using lexical and host-based features of the associated URLs.

Outline

Description

Problem Statement

Motivation

Approach and Implementation

Approach Towards the Problem

Support Vector Machines

Perceptron

Neural Networks!!!

Future Work

Future Work

Motivation for the Problem

- ▶ To keep users in a private network safe from external attacks from hackers, who launch attacks using malicious(e.g. phishing) web sites! The firewall can identify what url-request should be allowed to pass and dropping the malicious requests.

Motivation for the Problem

- ▶ To keep users in a private network safe from external attacks from hackers, who launch attacks using malicious(e.g. phishing) web sites! The firewall can identify what url-request should be allowed to pass and dropping the malicious requests.
- ▶ Users can even be attacked from hackers/attackers by viruses/autorun scripts being downloaded from a malicious url! These scripts can corrupt our machines or send valuable information to the attackers! We certainly don't want anything 'bad' to happen behind our back while we are online!

Motivation for the Problem

- ▶ To keep users in a private network safe from external attacks from hackers, who launch attacks using malicious(e.g. phishing) web sites! The firewall can identify what url-request should be allowed to pass and dropping the malicious requests.
- ▶ Users can even be attacked from hackers/attackers by viruses/autorun scripts being downloaded from a malicious url! These scripts can corrupt our machines or send valuable information to the attackers! We certainly don't want anything 'bad' to happen behind our back while we are online!
- ▶ Many times the network administrator wants to prevent request to certain kinds of web sites, deeming them as malicious! e.g. in China, sites on Tiananmen Square can be blocked in the network... or parental control...

Motivation for the problem

contd. ...

- ▶ We can prevent attacks on cloud servers, internet-backbone-routers, servers . E.g. attacks where the cloud servers/backbone-routers are being requested to get a malicious web page, which carries hidden scripts which can bring down the server...

Motivation for the problem

contd. ...

- ▶ We can prevent attacks on cloud servers, internet-backbone-routers, servers . E.g. attacks where the cloud servers/backbone-routers are being requested to get a malicious web page, which carries hidden scripts which can bring down the server...
- ▶ Many times some malicious web sites track user browsing habits and sell this information to interested parties, e.g. for advertising. These use cookies, multiple advertisement links etc. to track the user! These sites can be blocked ensuring user privacy.

Motivation for the problem

contd. ...

- ▶ We can prevent attacks on cloud servers, internet-backbone-routers, servers . E.g. attacks where the cloud servers/backbone-routers are being requested to get a malicious web page, which carries hidden scripts which can bring down the server...
- ▶ Many times some malicious web sites track user browsing habits and sell this information to interested parties, e.g. for advertising. These use cookies, multiple advertisement links etc. to track the user! These sites can be blocked ensuring user privacy.
- ▶ Many times the (naive) users are unaware of such activities on the internet!

Outline

Description

Problem Statement

Motivation

Approach and Implementation

Approach Towards the Problem

Support Vector Machines

Perceptron

Neural Networks!!!

Future Work

Future Work

Approach

- ▶ We use the sparse dataset available at <https://archive.ics.uci.edu/ml/machine-learningdatabases/url/> and try to understand the kind of feature space based on the description.

Approach

- ▶ We use the sparse dataset available at <https://archive.ics.uci.edu/ml/machine-learningdatabases/url/> and try to understand the kind of feature space based on the description.
- ▶ We try different models on this dataset, based on understanding of the dataset and the feature space.

Approach

- ▶ We use the sparse dataset available at <https://archive.ics.uci.edu/ml/machine-learningdatabases/url/> and try to understand the kind of feature space based on the description.
- ▶ We try different models on this dataset, based on understanding of the dataset and the feature space.
- ▶ We divide the dataset into training and validation to ensure that our model works well on new, unseen urls based on patterns in the url information.

Approach

- ▶ We use the sparse dataset available at <https://archive.ics.uci.edu/ml/machine-learningdatabases/url/> and try to understand the kind of feature space based on the description.
- ▶ We try different models on this dataset, based on understanding of the dataset and the feature space.
- ▶ We divide the dataset into training and validation to ensure that our model works well on new, unseen urls based on patterns in the url information.
- ▶ We apply the following three models based on what we have learnt in class... :)
 - i Support Vector Machine
 - ii Perceptron
 - iii Neural Networks!

Outline

Description

Problem Statement

Motivation

Approach and Implementation

Approach Towards the Problem

Support Vector Machines

Perceptron

Neural Networks!!!

Future Work

Future Work

SVMs!!

- ▶ SVMs works well for classification problem, maximizing the margins to get a separating hyperplane with enough breathing space!

SVMs!!

- ▶ SVMs works well for classification problem, maximizing the margins to get a separating hyperplane with enough breathing space!
- ▶ We use the fact that SVMs are able to work with higher implicit dimensions and even handle sparse datasets as ours.

Outline

Description

Problem Statement

Motivation

Approach and Implementation

Approach Towards the Problem

Support Vector Machines

Perceptron

Neural Networks!!!

Future Work

Future Work

Perceptron Update!

- ▶ Perceptron update works well for classification problem, with a separating hyperplane. We use the feature-set and believe that there is separating hyperplane and run the algorithm for perceptron update algorithm for some number of iterations.

Perceptron Update!

- ▶ Perceptron update works well for classification problem, with a separating hyperplane. We use the feature-set and believe that there is separating hyperplane and run the algorithm for perceptron update algorithm for some number of iterations.
- ▶ However the perceptron hyperplane does not give enough breathing space for the dataset!
- ▶ Moving on to the Neural Network... :)

Outline

Description

Problem Statement

Motivation

Approach and Implementation

Approach Towards the Problem

Support Vector Machines

Perceptron

Neural Networks!!!

Future Work

Future Work

Neural Networks!

- ▶ In machine learning and cognitive science, artificial neural networks (ANNs) are a family of models inspired by biological neural networks (the central nervous systems of animals, in particular the brain) and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which exchange messages between each other. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning.

Neural Networks!

- ▶ In machine learning and cognitive science, artificial neural networks (ANNs) are a family of models inspired by biological neural networks (the central nervous systems of animals, in particular the brain) and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which exchange messages between each other. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning.
- ▶ We implement our own neural network and tune various parameters like learning rate and number of hidden layers and number of units in each hidden layer etc. We implement back-propagation algorithm with squared error function (for simplicity).
- ▶ We use cross validation to analyse the results.

Neural Networks!

- ▶ In machine learning and cognitive science, artificial neural networks (ANNs) are a family of models inspired by biological neural networks (the central nervous systems of animals, in particular the brain) and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which exchange messages between each other. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning.
- ▶ We implement our own neural network and tune various parameters like learning rate and number of hidden layers and number of units in each hidden layer etc. We implement back-propagation algorithm with squared error function (for simplicity).
- ▶ We use cross validation to analyse the results.
- ▶ Neural Network beats SVM!!!

Outline

Description

- Problem Statement
- Motivation

Approach and Implementation

- Approach Towards the Problem
- Support Vector Machines
- Perceptron
- Neural Networks!!!

Future Work

- Future Work

What Next? Future Work...

Improving the Model

- ▶ We can analyse the feature set more with better understanding of the feature space. We can use better dimensionality reduction approaches for huge feature spaces(for NNs).

What Next? Future Work...

Improving the Model

- ▶ We can analyse the feature set more with better understanding of the feature space. We can use better dimensionality reduction approaches for huge feature spaces(for NNs).
- ▶ We can use regularization for NNs to ensure our model works very well for unseen data.

What Next? Future Work...

Improving the Model

- ▶ We can analyse the feature set more with better understanding of the feature space. We can use better dimensionality reduction approaches for huge feature spaces(for NNs).
- ▶ We can use regularization for NNs to ensure our model works very well for unseen data.

See it work!

What would be better than implementing our model on a live network to test it in practice for random and real life test data!! Maybe we could implement our model on our personal wireless access point, just for fun :) !

Conclusion

- ▶ It was a great experience working with different models learning and implementing new algorithms
- ▶ Neural Networks are great models to work with! Hoping to see a lot more of Machine Learning ahead... :D