BTU Cottbus, Chair of IT Security
Prof. Dr.-Ing. A. Panchenko
T. Ziemann
E. Strehle

Winter term 2019/2020

Part IV: Classification & Evaluation

Deadline: 15th January 2020

Study Project: Adversarial Machine Learning

Introduction

In the previous milestones we have collected network traffic corresponding to fetches of a given set of websites. From this traffic we abstracted to so-called features which are basic mathematical entities that describe the patterns included in the traffic in an abstract way. After removing statistical outliers, we observed noticeable differences by plotting the obtained feature vectors. The plots suggests that the network traffic, more specifically the corresponding feature vectors, are classifiable using the well studied methods of statistical learning. To continue this project we will apply methods of the field of machine learning now to be able to classify the different websites. Therefore you have to inform yourself about machine learning, more specifically classification methods, and already existing libraries to apply them.

Preliminaries

This time it is not really possible to solve the following sub-tasks independently from each other. Therefore read through all the given tasks and try to work out a scheduling before implementing anything. It is necessary to understand all the basics first. Afterwards the implementation/application of the gained knowledge will be easier.

Research

Classification is one of the main fields of machine learning. In the first step you have to inform yourself about different machine learning techniques applicable for classification problems. What are the mathematical models behind them? What input is needed and how you have to adapt the representation of your data set? Your research have to contain at least the following machine learning techniques:

a) Support Vector Machines (SVM), especially multiclass SVM

b) Random Forest

c) k-Nearest Neighbors (k-nn)

d) Neuronal Networks

Tasks

The following sub-tasks will guide you through the implementation and evaluation of the different classification methods. We will proceed in the following steps, namely the preparation of the data, the training of the classification model and the evaluation of the so obtained model. Each step and its requirements are prescribed in form of the sub-tasks below:

Subtask 01: Preparation of the Data

We start by preparing our data. Therefore we need to parse our already existing data set and convert it in a format compatible with the classification algorithm we have to apply. Take notice, that different libraries may assume a different data representation.

Subtask 02: Train your Models

The next step is to apply all four machine learning techniques mentioned above to your data set. Therefore you do not have to implement these techniques by yourself. You are allowed to use already implemented libraries. But include them within your project. To train the models we need to provide training data in the sense of our collected feature vectors. However, if we would train on all data available we have to left data to test our trained classifier in the next step. Therefore, we need to split our data sets in two parts, namely the *training set* and the *testing set*. Think about a reasonable splitting aspect ratio and discuss the advantages/disadvantages of different ratios/approaches.

Hint: Notice that the testing (evaluation) of the model is often directly combined with the training of the model (cf. Subtask 03). Thus, sometimes it is not possible to separate the testing from the training.

Subtask 03: Evaluation

In the last step of this milestone we need to evaluate the applied methods. The evaluation and visualization of machine learning algorithms is one of the most recent and most discussed research topics in the field of machine learning. Due to the big amount of data it is (nearly) impossible to get any (theoretical) insights on the learned behavior. The only possibility to verify or evaluate your trained model is to observe different test scenarios and examine their propriety. Therefore, the design of the testing-stage is one of the most critical ones within this study project. To observe different aspects of the trained model there are different known techniques, like *k-fold cross-validation* and *confusion matrices*. Inform yourself about these terms and understand there specific aspect of evaluation. Afterwards evaluate the four machine learning techniques by considering the following points:

a) Apply *k-fold cross-validation* and test your classifier.
   a) What are useful values for `k`?
   b) Name the accuracies for each fold as well as the mean accuracy over all folds.
   c) How sure you could be about the obtained mean accuracy with respect to variances within your data set; respective the selection of training and testing set? State the accuracy insurance by means of confidence intervals.

b) Create and analyze the `confusion matrix`.
   a) How many / Which websites are easy to fingerprint across all classifiers?
   b) How many / Which websites are difficult to fingerprint across all classifiers?
   c) How many / Which websites are not correctly classified by all classifiers?

c) Evaluate the rates of *True Positives*, *False Positives*, *True Negative* and *False Negatives*. How are these terms related to the `confusion matrix` and your from the cross-validation obtained accuracy?

d) Document and visualize your validation process. Think about reasonable representation of your evaluation, e.g, plots of confidence intervals, confusion matrices, comparison tables, etc.

e) Could your classifier be improved? How?

Presentation / Consultation

Prepare a consultation. You shall explain the four machine learning techniques, how you applied and evaluate them. Furthermore present and explain your implementations. Give a proof of work through a demonstration of your program/scripts. Each group has to hold a consultation.

Three groups have to hold a presentation. These groups will receive detailed information about which machine learning technique has to be presented. Of course the presentation must contain a proof of work.

Good luck!