Sommario

[Introduction 1](#_Toc529823509)

[Drebin Dataset 1](#_Toc529823510)

# Introduction

In this report I’ll describe the implementation of an Android malware detector and classifier based on machine learning techniques. In particular I’ll use Supervised Learning methods, where the ground truth is provided by a huge open dataset called “Drebin” containing features extracted by malwares and benign programs collected through both official and unofficial application stores.

The problem of malware detection is a binary classification problem, where the two classes in which programs must be divided into are the positive class, denoted with the symbol “+” and the negative class, denoted with “-”. The former will contain malwares and the latter benign programs.

The second problem is a different classification one, in which malwares must be divided into classes according to the “family” they belong to, where a family is a set of malwares with the same behavior that usually have similar features.

In this report I’ll describe the techniques used to solve both these problems and the relative implementation in Python.

# Drebin Dataset

Drebin dataset is an open dataset containing features from 123,453 benign applications and 5560 malwares. The dataset is structured as follows: the root folder contains a subfolder “features\_vector” and a csv file “sha\_families” containing couples «sha , family» where “sha” is the sha-256 hash of the malware and “family” is the family the malware belongs to. The subfolder contains a txt file for each program, named with the hash of the program that the file is related to. Each file contains a line for each feature of the program. Features can be divided into two categories: static and dynamic.

Static features are string extracted by the Manifest file, that provides general information on the application, static features can be divided in turn in four subcategories: hardware components, requested permissions, app components and filtered intents.

Dynamic features are obtained by disassembling the executable of the applications and they include all APIs calls in the applications and all the strings (including network addresses).

# Malware Detection

In this section I’ll describe the solution to the first problem: given a program the algorithm must classify it as benign or malware. This is a problem of binary classification, and we’ll assume that programs are linearly separable.

The problem can be geometrically interpreted in a N-dimensional vector space, where N is the cardinality of the set of all features collected from all programs in the dataset.

In this space a program is a N dimensional vector:

𝓋 ∈ ℝN  s.t. 𝓋i ­­=

To represent a program in this way we must first index all features with a progressive number from 1 to N.

The solution of the problem is the hyperplane that better partition the ℝN space in two subspaces:

ℝn = 𝔸 ⊕ 𝔹

Such that a program 𝓍 can be classified according to this partition as follows:

𝓍 =

Once we have the equation of the hyperplane:

𝓎(𝓍) = (𝓌­­­1 , 𝓌­­­2 , … , 𝓌­­­N ) + 𝓌­­­0

It’s easy to classify a new instance:

𝓍 =

So the learning process consists on learning the weights 𝓌­­­i of the hyperplane. There’re a lot of algorithms for computing these weights according to a training set, in my implementation I used SVM, that’s a method that finds the optimum value for each weight 𝓌­­­i maximizing the distance from the hyperplane to its closer point. This method is reliable because it’s able to find the optimum solution even in presence of outliers (i.e. a small number of points that are distant from the others).