DISI – University of Trento

Master in Computer Science AA 2018/2019

Simulation and Performance Evaluation

Assignment 1

Stochastic Cryptoanalysis

Renato Lo Cigno, Michele Segata

April 12, 2019

In the country of Povolandia, an organized criminal group has found a smart way to encrypt their messages to escape interception. They use N samples of noise generated from a gamma distribution to encrypt the characters, and each character is mapped onto the gamma distribution parameters, for instance "$a \leftrightarrow (\alpha = 1, \lambda = 1)$". Furthermore, to make the samples look like noise they randomly multiply each sample by either 1 or -1.

The gamma distribution is defined as

$$f_X(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)};\ \alpha, \lambda, x > 0; \quad \Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\, dy$$

Povolandia Security agency, the PSA, through interception, has discovered the general scheme used as described above, and has recorded some files with random samples that represent some known specific word, or sentences, or in general group of letters. Furthermore, they have an intercepted file with an important message to decrypt, but they don't know how to make any sense of this all, so they ask your help.

You are given a `zip` archive including a set of comma-separated values files. The format is the same for all of them and it is shown in Fig. 1. It is basically a single column with the samples. One of those, called `secret.csv` is the file including the message to be decrypted. The other ones are the words known to the the PSA. The filename indicates the known word, while the content is the set of samples used to encode the word. For example, if the filename is `abcd`, the content will include the samples that the criminals have used to encode the letters `a`, `b`, `c`, and `d`.

The PSA also knows some information that might help you in your task. They know that the alphabet is only composed by the 26 lowercase letters of the English alphabet plus the space. There are thus no punctuation or numbers. All the sentence are written in English. The PSA also warns you that their database is not complete. For example, if the secret sentence is `hello world` and your archive only includes the file `dehlw`, however food in processing you are, you can only decode `hell**w**ld`, where `*` means a character not decoded, but at this point you can use your own language skills to infer some of the `*`, and iterate decoding. The PSA database should cover about 80% of the characters of the secret sentence, so it should not be too difficult to recover the sentence, so the goal is to recover the sentence as complete and correct as possible. However, even if you don't manage to recover the complete sentence, you will be evaluated on the approach used to recover the known letters and on the inference methodology.

The goals of the assignment are four:

1. Find the correct value of N;
2. Recover the characters using the database of known letters provided by the PSA;

```
"sample"
-0.911136916097835
0.723910250120982
12.2460337091729
-11.4493983914601
0.206203636353832
-1.15346373671099
1.62845063831073
0.224569695620115
-1.38968550716255
. . .
```

Figure 1: Format of the intercepted files

3. Find the meaning of the intercepted sentence;
4. Find as many mappings “character” $\leftrightarrow (\alpha, \lambda)$ as you can.

With respect to point 4, this means enriching the PSA database with the characterization of the letters they yet don’t know, listing them in a table like Tab. 1. As you might need to generate many character mappings, we provide you with a script that takes the mappings from a `csv` file and outputs the LaTeX code for generating a table like Tab. 1. Your data analysis script should generate a `csv` file according to the format shown in Fig. 2. You can then pass this file as a command line argument to the script we provide to generate the table. We provide you with an `R` and a `python` script. If you are using `R`, run the script with

```
Rscript table.R output.csv
```

If you are using `python`, run the the script with

```
python table.py output.csv
```

Simply copy the textual output of the script into your report. Be sure to have `\usepackage{booktabs}` in the preamble of your report.

Table 1: List of decoded characters.

Letter	α	λ
a	2.00	2.00
b	3.00	3.00
c	4.00	4.00
...	...	...

```
letter,alpha,lambda
a,2,2
b,3,3
c,4,4
. . .
```

Figure 2: Format of the output csv file.

Each student will find his/her own dataset inside the `assignment1-data` folder of the Classroom Google Drive storage (format `<surname>.zip`). Inside the same folder, you also find the scripts that you can use to generate Tab. 1. To write your report use the LaTeX template[1] we suggested and do not write more than 3 pages. Deliver the PDF file of the report and the R, Matlab, or python script you used for processing and plotting as a single .zip or .tar file through Classroom; if the script does not run on a standard Linux box, we simply notify you that it does not work, and we will not attempt correction. Keep your code **CLEAN, ORGANIZED, and COMMENTED**. Do not send us your source code with unused portions commented out, blocks of code with no comments, or with monolithic pieces of code. Split your code in functions. Do not include the data set in the zip file, we already have it! The script MUST work assuming that the dataset is in the same folder of the script. DO NOT use absolute folders like

```
ds <- read.csv('/home/john.doe/Documents/spe/doe.csv')
```

but rather

```
ds <- read.csv('./doe.csv')
```

The deadline for the assignment is May 30 . . . 2019!!

If you have some doubts, just write us an email or ask in class.

Have Fun!

[1] `http://disi.unitn.it/locigno/teaching-duties/spe/assignment-template.zip`