Villeneuve d'Ascq, June 23rd, 2023

To whom it concerns,

Object: Report on the Doctoral Thesis of M^r Vaclav MACHA entitled "General Framework for Classification at the Top"

The doctoral thesis of M^r Vaclav MACHA deals with the problem of deriving learning strategies to increase the performances of classifiers "at the top", i.e. only for the top-ranked items. These strategies may also be used to increase the accuracy for very small false positive rates, which is a desired property for applications related to sensitive decision-making such as medical imaging, intrusion detection, or steganalysis.

Thesis organization and appreciation:

The dissertation is more than a hundred pages long, it is written in English and is divided into 6 chapters, 4 of them presenting the authors' contributions. The thesis is extremely pleasant to read, and it is very well documented with appropriate references, illustrations, notes, and highlighted properties. The subject being very theoretical, this is a challenging task to perform, and I appreciate the fact that M^r MACHA, by recalling the main properties of the learning strategies, but also presenting theoretical concepts such as dual formulations using kernels or surrogate formulations, make this theoretical manuscript very pedagogical. Note that all the proofs and auxiliary results are located in the appendices, I think that this choice also eases the reading of the manuscript since the reader can be focussed on the theoretical implications of each result.

I describe below the contents of the different chapters.

- After a brief motivation, the introduction presents the context of "Classification at the Top" in a formal way. The classical metrics used in classification are presented together with metrics suited for classification at the top. Different formulations of this optimization problem are proposed: ranking, accuracy at the top or hypothesis testing. The differences rely on the error metrics (false positive and/or false negative) which are used for the objective functions or for the constraints.

- The second chapter introduces the use of surrogate functions, whose goals are to be differentiable while approximating different error rates. This leads to different optimization formulations called TopPushK, Grill, TopMeanK, Pat&Mat, together with Neyman-Pearson formulations considering only false positives as constraints. Mr MACHA takes care to contrast the different approaches by for example presenting a synthetic table listing the specific hyper-parameter, the implication of false positive and false negative, and the way the threshold is computed.

- The three next chapters present different learning strategies. Chapter three proposes a linear classifier that leads to a differentiable formulation for the Pat&Mat formulation and where a Stochastic Gradient Descent (SGD) can be applied. The stability and convexity of other formulations are also studied and M^r MACHA concludes on the fact that the stability of the learning procedure can be mitigated by adopting smaller thresholds, at the cost of a loose approximation of the initial optimization formulation. Again all these different properties are summed up at the end of the chapter to draw important conclusions.

- The fourth chapter presents the dual formulations of the previous linear models, which enables, with the use of kernels, to address non-linear problems. The dual formulations are written for the different strategies and the use of kernel function is combined with a coordinate descent algorithm in order to reach convergence without differentiation. As noted by the author, this formulation is unsuitable for large training databases.

- Chapter five presents a learning strategy that can be applied to mini-batches of samples and which is compatible with the SGD, hence can be for example suitable for deep learning. In this setup, the main difficulty is in the estimation of the global threshold from a set of samples constituting the mini-batches. The inherent bias of the sample gradient is first studied and M^r MACHA proposes a scheme called DeepTopPush, working in the primal space, that reuse samples having top scores in the future mini-batches. A theoretical analysis shows that this strategy decreases the initial bias after several iterations. In my opinion, this formulation is the more original contribution of the thesis, it also takes advantage of all the advances presented in the former chapters.

- Finally, the sixth chapter presents numerical experiments which illustrate the benefits of the theoretical derivations presented in the previous three chapters. The different metrics, and their optimality w.r.t. the different formulations are recalled. Both primal and dual formulations are evaluated for an image recognition task using the metrics but also an appropriate visualization diagram. In this setup, DeepTopPush does not achieve competitive results, but on the contrary, for two other sensitive applications (steganalysis or intrusions detection) the provided ROC curves show the power of DeepTopPush. For example for steganalysis an a FP rate of 10^{-5}, the TP rate is larger than 0.8 when it is close to 0.0 using cross-entropy. Similar results are obtained for malware detection. Such results highlight how this contribution can be efficient for operational classification. As a minor remark, for steganalysis, the names of the image databases should be recalled.

- The conclusion recalls the contributions, whose implementations are available on github and a small paragraph presents a possible extension of this work.

General Conclusions:

In my opinion, this dissertation shows that M^r Vaclav MACHA has acquired important knowledge in machine learning. The manuscript is also easy to read and constitutes in my opinion an excellent starting point for any Master student wanting to start a PhD on a similar subject. I particularly appreciated the fact that the proposed solutions are both original and evaluated using practical data. The author of the thesis proved to have the ability to perform research and to achieve scientific results. I do recommend preparation to receive the Degree of Ph.D.

Patrick Bas,

Research Director at "*Centre National de la Recherche Scientifique*" (CNRS).

tel: +33 (0)3 20 33 54 46

email: Patrick.Bas@cnrs.fr