

Introduction to Statistical Computing in Scala - an Implementation of the K-Nearest Neighbors classifier

Roxana Tesileanu, *Research Assistant, National Institute of Forest Research and Management (INCDS), Romania*

Abstract—Statistical computing in ecology evolves at a high speed, mainly because researchers have recognized the advantage of being able to design their algorithms according to their needs. The present paper introduces the implementation in Scala of the k-Nearest Neighbors (kNN) classifier based on Euclidean distances, which can be applied also on small datasets, a situation commonly encountered in ecological research.

Index Terms—machine learning in ecology, k-Nearest Neighbors classification, Scala

I. INTRODUCTION

One of the drivers of the machine-learning progress is the great amount of data born within and gathered by networked and mobile computing systems which necessitates further processing in order to gain insights into the specific fields from which it originates (Jordan and Mitchell 2015, p. 256). The "Big Data phenomenon" is real but, unfortunately, it can not be generalized to all research areas, especially some areas of ecological research which investigate systems which are by nature data-poor and will probably remain as such unless cost-intensive data collection projects are being proposed and financed. This by no means implies that such areas cannot take advantage of the progress of machine-learning and take the best out of the existing datasets. On the contrary, machine-learning, being a study field which "sits at the crossroads of computer science, statistics and a variety of other disciplines concerned with automatic improvement over time, and inference and decision-making under uncertainty" (Jordan and Mitchell 2015, p. 256), is welcomed in real-world environmental decision-making facing the need of proposing courses of action under the cloud of uncertainties and regarding issues characterized by multiple attributes (Vatn and Bromley, 1994). Moreover, machine-learning is used not only in analyzing data from observational studies (which might benefit from larger time series datasets), but also from experimental studies (which can't produce large amounts of data but benefit from an experimental design, delivering data which provides stronger inference about effects and causal relationships - Quinn and Keough, 2002) (Harrington 2012). Indeed, a combination of machine-learning classifiers with the benefits of a MANOVA experimental design used for collecting training data, might produce useful research results, in ecological research too. Thus, in order to be able to use the power of machine-learning algorithms adapted to small ecological datasets, a statistical package written in Scala is currently under development at INCDS. The idea of adapting machine-learning algorithms to

real-world data requirements is not new. Hastie and Tibshirani (1996), describe the Discriminant Adaptive Nearest Neighbor (DANN) procedure as a locally adaptive form of nearest neighbors classification using a modified linear discriminant analysis (LDA) procedure to estimate an effective metric for computing neighborhoods.

II. BASIC VECTOR OPERATIONS IN SCALA

III. READING FILES FOR CLASSIFICATION TASKS

IV. K-NEAREST NEIGHBORS (kNN) CLASSIFICATION

V. THE MACHINE-LEARNING PIPELINE FOR REDUCING ERROR

VI. CONCLUSION

APPENDIX A

SOURCE FILE -

SRC/MAIN/SCALA/MAI/SCALAML/BASICVECTOROP.SCALA

.....

APPENDIX B

SOURCE FILE -

SRC/MAIN/SCALA/MAI/SCALAML/READFILE.SCALA

.....

APPENDIX C

SOURCE FILE -

SRC/MAIN/SCALA/MAI/SCALAML/kNN.SCALA

.....

APPENDIX D

SOURCE FILE -

SRC/MAIN/SCALA/MAI/SCALAML/AUTONORM.SCALA

.....

APPENDIX E

SOURCE FILE -

SRC/MAIN/SCALA/MAI/SCALAML/TESTFRAMEWORKCLASSIF.SCALA

.....

ACKNOWLEDGMENT

The author would like to thank Jeff Druce and Mike Reposa for very useful learning tips.

Roxana Tesileanu My bio.