Introduction to Predictive Modelling: A Machine Learning Perspective

Elio Amicarelli
Warwick Q-Step Methods
Spring Camp 2017

March 10, 2017

Description

We live in exciting times. The domain of data analysis is in continuous expansion thanks to the unparalleled amount of data we are producing and the increasing computational power at our disposal. At the forefront of this expansion there is the science of prediction. Statistical prediction, a research field once populated by few and looked at with suspicion by many, is experiencing today an unmatched level of methodological dynamism and popularity. In this context, social scientists are increasingly interested in developing predictive models and mastering Machine Learning techniques already adopted in other fields of applied and academic research. This workshop will introduce the participants to the core concepts underlying the theory and practice of predictive modelling from a Machine Learning perspective. The aim is to provide participants with a set of concepts and tools indispensable to *i)* effectively understanding and engaging with predictive-oriented research, *ii)* knowingly applying predictive models in their own research and, *iii)* confidently continuing their studies in predictive modelling.

Prerequisites

There are no specific prerequisites to attend this workshop. Previous exposure to basic statistics and data analysis concepts (e.g. generalized linear models) and familiarity with the R programming language is desirable but not mandatory.

Structure

The workshop is articulated in 5 sections.

Section 1: Introduction and plan of action

In the first section we will introduce the attendees to the goals and vocabulary of predictive modelling. We will also give a schematic overview of the predictive modelling pipeline and illustrate how the different steps in this pipeline are mirrored in the workshop's roadmap.

Section 2: Core concepts in predictive modelling

In this section the participants will learn the core theoretical concepts underlying the development of good predictive models. The discussion will be mainly based on how model complexity affects the generalization performances of a predictive model and how the analyst should select the appropriate degree of model complexity. At the end of this section the participant should have a working knowledge of the following concepts and their relationships: bias-variance tradeoff, overfitting, underfitting, training set, validation set, test set, cross-validation.

Section 3: Tree-based methods

Section 3 will give an introduction to the functioning and use of tree methods for classification and regression tasks. The attendees will learn how classification and regression trees are built and what the strengths and weaknesses that characterize them are. The attendees will be shown the practical implications of model complexity on predictive performances as discussed in Section 2 through the lenses of tree methods. In doing so, we will also give a concise account of the most important metrics used in the Machine Learning literature to assess classifiers' predictive performances.

Section 4: Ensemble methods

The aim of this section is to introduce the attendees to the idea of ensemble learning; combining several models can be a fruitful strategy to improve upon their individual predictive performances. This concept will be examined in relation to the theoretical aspects discussed in Section 2 and to the characteristics of tree methods as discussed in Section 3. The participants will learn how Random Forests can be built by combining several trees and why this ensemble method improves upon the predictive performances of individual trees. An introduction to a second ensemble method, namely Adaptive Boosting, will be provided, time permitting.

Section 5: Hands-on and where to go from here

In the final part of the workshop, participants will apply the concepts and models discussed to face a predictive task and to evaluate the predictions of their models. To conclude, useful learning resources will be shared with the participants along with some tips regarding how to best build upon the knowledge acquired during this workshop.

Background readings

1) Breiman, L. (2001). Statistical modeling: The two cultures (**with comments and a rejoinder by the author**). *Statistical science*, 16(3), 199-231. Ungated version available here.

2) Shmueli, G. (2010). To explain or to predict?. *Statistical science*, 25(3), 289-310. Ungated version available here.

3) Kennedy, R., Wojcik, S., & Lazer, D. (2017). Improving election prediction internationally. *Science*, 355(6324), 515-520.
Note: This is a relatively technical paper. I would like you to try to read it before and after the workshop.

4) Ward, M. D., Greenhill, B. D., & Bakke, K. M. (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4), 363-375.

For participants particularly interested in peace and conflict research I also recommend to read the following:

5) Hegre, H., Metternich, N. W., Nygrd, H. M., & Wucherpfennig, J. (2017). Introduction: Forecasting in peace research. *Journal of Peace Research*, 0022343317691330. Ungated version available here.