

1 The caret package

Prediction is a central problem in machine learning that involves inducing a model from a set of training instances that is then applied to future instances to predict a target variable of interest. Several commonly used predictive algorithms, such as logistic regression, neural networks, decision trees, and Bayesian networks, typically induce a single model from a training set of instances, with the intent of applying it to all future instances.

We call such a model a population-wide model because it is intended to be applied to an entire population of future instances. A population-wide model is optimized to predict well on average when applied to expected future instances. Recent research in machine learning has shown that inducing models that are specific to the particular features of a given instance can improve predictive performances (Gottrup et al., 2005). We call such a model an instance-specific model since it is constructed specifically for a particular instance (case).

The structure and parameters of an instance-specific model are specialized to the particular features of an instance, so that it is optimized to predict especially well for that instance. The goal of inducing an instance-specific model is to obtain optimal prediction for the instance at hand. This is in contrast to the induction of a population-wide model where the goal is to obtain optimal predictive performance on average on all future instances.

The caret package (short for Classification And REgression Training) is a set of functions that attempt to streamline the process for creating predictive models. The package contains tools for:

- data splitting
- pre-processing
- feature selection
- model tuning using resampling

- variable importance estimation