Declarative data generation with ProbLog

Anton Dries ANTON.DRIES@CS.KULEUVEN.BE
Department of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium

Keywords: data generation, probabilistic logic programming

Abstract

In this paper we describe a novel declarative approach to data generation based on probabilistic logic programming. We show that many data generation tasks can be described as a probabilistic logic program. To this end, we extend the ProbLog language with continuous distributions and we develop a simple sampling algorithm for this language. We demonstrate that many data generation tasks can be described as a model in this language and we provide examples of generators for attribute-value data, sequences, graphs and logical interpretations and we show how to model common extensions such as noise, missing values and concept drift.

This paper was originally published at SOICT'15 (Dries, 2015).

Introduction

Machine learning is an experimental science in a sense that many of its results are validated in an empirical way. As in any other science, the goal of experimentation is twofold: (1) to validate the developed methodology or hypothesis, and (2) to study the behavior of a system in varying conditions in order to gain a deeper understanding.

Where real data is required to achieve the first goal, artificial data is of vital importance to achieve a better understanding of the algorithms and methodologies of machine learning. In contrast to real-world data it has several benefits that make it suitable for large-scale experimentation:

- It allows for controlled experiments.
- The ground truth is known.
- It is available in abundance without legal, ethical and commercial restrictions.

As a result, artificial data is often used to determine relationships between data characteristics and algorithm performance and to analyze the behavior of algorithms in changing environments.

In this paper, we develop a new methodology for data generation. Our framework follows a declarative approach based on probabilistic logic programming. It has several advantages over existing work:

- It is declarative, that is, the user only needs to specify distributions on variables and relations between them. The data generation process is completely handled by the underlying engine. This also increases transparency and reproducibility because all the properties of the data are fully described in a relatively simple model.
- It is flexible and general. The framework is based on Prolog, which is a general purpose programming language. It is therefore capable of describing arbitrarily complex distributions and structures.
- It is modular. It allows different components to be defined separately and be reused.
- It is based on existing technology, namely Prolog and ProbLog, which means that there are many resources available that describe how to program these models.

Appearing in *Proceedings of Benelearn 2016*. Copyright 2016 by the author(s)/owner(s).

References

Dries, A. (2015). Declarative data generation with problog. *The Sixth International Symposium on Information and Communication Technology*. Hue, Vietnam: ACM.