DataPhilly Workshop: storedDATAstories

P J Kowalczyk

2020-09-10

Introduction

This workshop will take one through the steps associated with an end-to-end machine learning campaign: data retrieval; data curation; model construction, evaluation, selection and interpretation; and reporting. Particular attention will be paid to reporting, i.e., building a narrative. Examples will be presented demonstrating how one might generate multiple output formats (e.g., HTML pages, presentation slides, PDF documents) starting with the same code base.

As a specific example, a data narrative will be built showing how one might build predictive models for the toxicity of organic molecules. Reports will be presented as (1) an HTML file, (2) a PDF or Word document (in a format acceptable for journal submission), and (3) a slide presentation.

While the workshop's example comes from the field of cheminformatics, the computational tools used and the exercises presented are applicable to any field where an investigator is interested in building predictive models, and describing these models to colleagues and associates.

At the workshop's conclusion attendees will have worked through exercises that may serve as templates to be used with their data as they build their data narratives.

Machine Learning Workflows

CRISP-DM

CRoss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining is an open standard process model that describes common approaches used by data mining experts. The sequence of the phases is not strict and moving back and forth between different phases is always required. The arrows in the process diagram indicate the most important and frequent dependencies between phases. The outer circle in the diagram symbolizes the cyclic nature of data mining itself. A data mining process continues after a solution has been deployed. The lessons learned during the process can trigger new, often more focused business questions, and subsequent data mining processes will benefit from the experiences of previous ones.

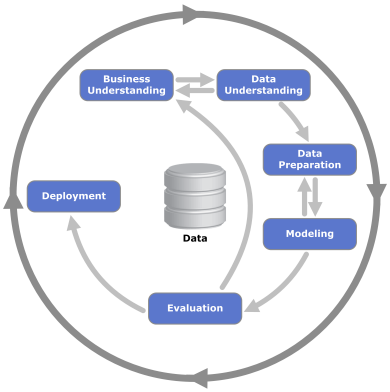

Figure 1: CRISP-DM.

R in Action

While the material presented during this Workshop will address all steps in an end-to-end machine learning workflow, particular attention will be paid to *reporting* the steps and outcomes.

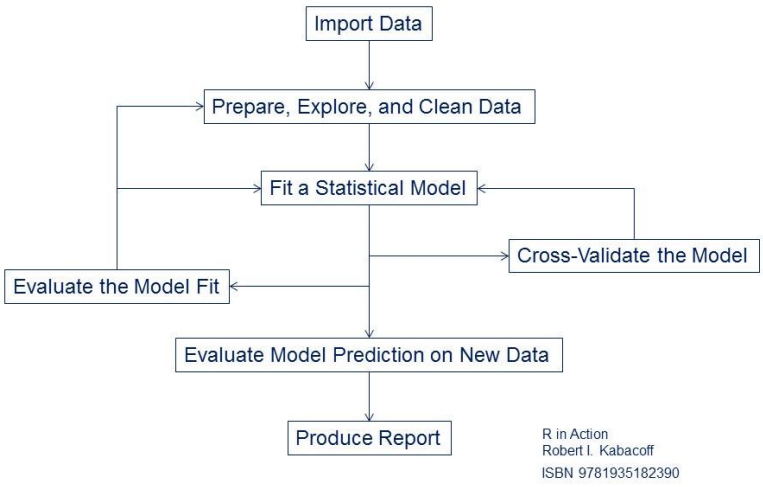

Figure 2: Machine Learning Workflow, taken from 'R in Action'.

Tables

Table 1: A subset of the H3 dataset.

	TargetID	AnalogSeriesID	pKi
253	CHEMBL264	22123	6.76
513	CHEMBL264	225	7.43
63	CHEMBL264	21586	9.00
541	CHEMBL264	5075	8.52
331	CHEMBL264	8347	9.42
920	CHEMBL264	1840	8.50
151	CHEMBL264	359	8.99
450	CHEMBL264	21586	8.83
527	CHEMBL264	15079	7.33
727	CHEMBL264	11487	9.70

Data Summary

Figure 1 is a **png** file; Figure 2 is a **jpg** file. It is also possible to dynamically generate figures.

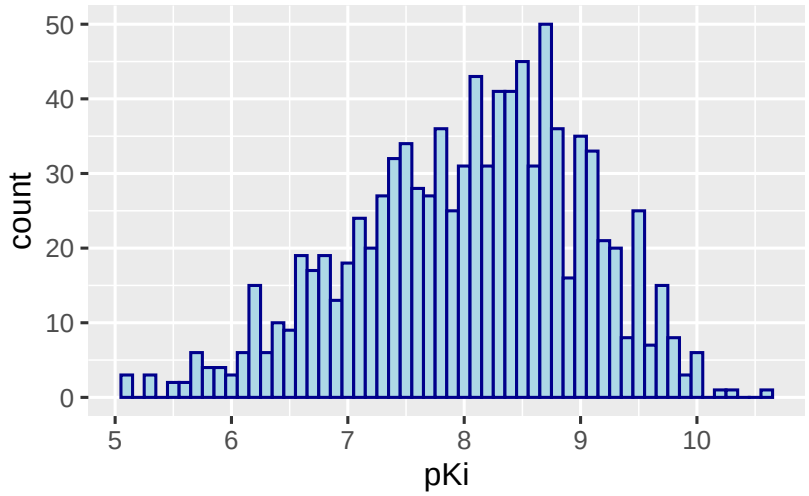

Figure 3: Distributiuon of pKi Data. This figure is generated when the document is Knit.