Inductive Programming: Tutorial 7
Data Wrangling

Stephen Muggleton

The aim of this tutorial is to help you understand concepts in Lecture 7, involving Data Wrangling.

Question 1

1. What is Data Wrangling?
2. What is the motivation for developing Inductive Programming techniques for Data Wrangling?

Solution

1. Data wrangling is the process of transforming and mapping data from one raw data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.
2. Data Wrangling is widely used across Business, Science and Medicine. It requires large amounts of programming effort and the resulting programs are often error-prone.

Question 2

Give an example of a Data Wrangling date transformation problem.

Solution

Id	**Input**	**Outputs**
1	25-03-74	25/03/74
2	29-03-86	29/03/86
8	...	...

Question 3

Give an example of an unsolved Inductive Programming problem revealed in the MagicHaskeller experiment.

Solution The accuracy of the Inductive Programming was highly dependent on selection of the set of primitives used. However providing all the background predicates also decreases accuracy. This is called the *Relevance* problem.

Question 4

Give an example of Inductive Programming being applied to a scientific data analysis problem.

Solution An example of such a problem is that of extracting predation facts, involving pairs of species, from ecological papers. This is exemplified below.

Harpalus rufipes *eats* large prey such as **Lepidoptera**.

Bembidion lampros: In cereals the main *food* was **Collembola**.

Question 5

Give an example of Inductive Programming being applied to a medical data analysis problem.

Solution An example of such a problem is that of extracting facts from patient records. For instance, the three marked entries below may need to be identified.

P003

56

Diagnosis: **carcinoma** , lung disease: unknown

20.78