Multivariate Prediction for Learning in Relational Graphs

Yi Huang, Volker Tresp and Hans-Peter Kriegel

Siemens       Siemens       Munich University

- The term *Linked Data* is used in context of the Semantic Web as the effort to link and jointly explore several structured data sources.

- **Data Situation:**
  - Data are heterogeneous with many different entity types and relationships
  - Relations are often extremely sparse, e.g., only a tiny subset of all possible persons are someone's friends
  - Information is missing e.g., for privacy reasons

- This presentation concerns learning with data in relational formats
  - In focus are simplicity, scalability and ease of use
  - Learned tuples and their certainty values can be stored in the data base and can be integrated in querying (SQL-queries)

- **Graphical Representation:**
  - Circles: constants
  - Rectangles: relationships
  - Diamonds: tuples, i.e., *Random Variables*

A graphical representation for the data in RDB

- **Person**
  - *Jack*
  - *Joe*
  - *Jane*

- **Knows**
  - *Jack* *friend* *Joe*
  - *Joe* *friend* *Jane*
  - *Jane* *friend* *Jack*

- **Income**
  - *High*
  - *Low*

- **IncomeScale**
  - *Low*
Multivariate Prediction for Learning in Relational Graphs (cont.)

- **Statistical Framework:**
  - **Statistical units:** objects of a certain type
  - **Population:** set of statistical units under consideration
  - **Sample:** random subset of the population
  - **Multivariate outputs** (response variables): tuple variables concerning a statistical unit
  - **Inputs** (covariates): (additional) features derived from tuples

- **Predictive Setting:**
  - **Transduction:** prediction of new tuples for objects in the sample
  - **IntraRDB-Induction:** prediction of new tuples for objects in the RDB but outside of the sample
  - **ExtraRDB-Induction:** prediction of new tuples for objects outside of the RDB