Informed Machine Learning

An Overview

Gkatsis Vassilis

PhD candidate @ DIT UOA

December 15, 2023

What is it?

Definition

The set of methods used to integrate prior knowledge into Machine Learning systems.

Define prior knowledge.

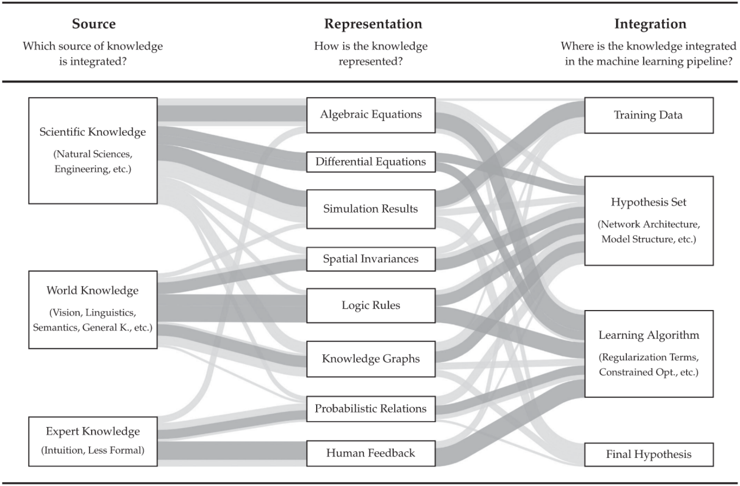
- Laws of Nature
- Experimental/Simulation Results
- Entity Relations
- Experts Intuition
- Annotator Feedback

What's the use of it?

Scenarios with:

- Data Scarcity (e.g. expensive data acquisition/labeling.)
- Expensive simulations.
- High need for meeting constraints.
- Need for transparency.

Structure of Informed ML



Source
Which source of knowledge is integrated?
Representation
How is the knowledge represented?
Integration
Where is the knowledge integrated in the machine learning pipeline?
Scientific Knowledge
(Natural Sciences, Engineering, etc.)
World Knowledge
(Vision, Linguistics, Semantics, General K., etc.)
Expert Knowledge
(Intuition, Less Formal)
Algebraic Equations
Differential Equations
Simulation Results
Spatial Invariances
Logic Rules
Knowledge Graphs
Probabilistic Relations
Human Feedback
Training Data
Hypothesis Set
(Network Architecture, Model Structure, etc.)
Learning Algorithm
(Regularization Terms, Constrained Opt., etc.)
Final Hypothesis

Figure: Laura von Rueden et al. “Informed Machine Learning – A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems

Human Feedback

in Reinforcement Learning (Definition)

Problem Definition

Train agents on simple robot tasks (e.g. walking or hopping) and simple games (e.g. space invaders and pong) while minimizing the cost of human-agent interaction (e.g. in terms of human time spent).

Christiano et al., “Deep Reinforcement Learning from Human Preferences”.

Human Feedback

in Reinforcement Learning (Need for IML)

- Finding an optimal policy is complicated.
- Traditionally agent-human interaction is costly.
- We want to eploit human intuition.

Christiano et al., “Deep Reinforcement Learning from Human Preferences”.

Human Feedback

in Reinforcement Learning (Solution)

- Provide a human with 2 choices for the next instance.
- Minimize the difference between agent-human selections.

Christiano et al., “Deep Reinforcement Learning from Human Preferences”.

Knowledge Graphs

in Multi-omic Data Analysis Example (Definition)

Problem Definition

Predict progression-free interval using different -omic data (e.g. genomic, proteomic) available in small amounts.

Ma and Zhang, "Multi-view Factorization AutoEncoder with Network Constraints for Multi-omic Integrative Analysis".

Knowledge Graphs

in Multi-omic Data Analysis Example (Need for IML)

- Large number of features small number of available data.

Ma and Zhang, “Multi-view Factorization AutoEncoder with Network Constraints for Multi-omic Integrative Analysis”.

Knowledge Graphs

in Multi-omic Data Analysis Example (Solution)

- Obtain a feature representation for each sample "X_i".
- Obtain a genomic features interaction ("similarity") graph "G_{ij}".
- Constrain representations inconsistent with the interaction graph.

Consistency Check

If G_{ij} is high → $||\mathrm{X}_i - X_j||^2$ should be low.

Ma and Zhang, “Multi-view Factorization AutoEncoder with Network Constraints for Multi-omic Integrative Analysis”.

Simulation Results

in Grass Pasture Nitrogen Response Rate Prediction (Definition)

Problem Definition

Predict Grass Pasture Nitrogen Response Rate when data concerning initial conditions are scarce or infrequent.

Pylianidis et al., “Simulation-assisted machine learning for operational digital twins”.

Simulation Results

in Grass Pasture Nitrogen Response Rate Prediction (Need for IML)

- Traditionally solved with simulation models.
- Occasionally data are not available.
- Or may not be available in desired frequency.

Pylianidis et al., “Simulation-assisted machine learning for operational digital twins”.

Simulation Results

in Grass Pasture Nitrogen Response Rate Prediction (Solution)

Provided a Simulation Model

- Inefficient in scarce data conditions.
- Can create training data for a ML model.

Train a ML model.

- Use simulated data as train/test.
- Test it in various data scarcity settings.

Pylianidis et al., “Simulation-assisted machine learning for operational digital twins”.

Logic Rules

in Image Classification with Semantic Based Regularization (Definition)

Problem Definition

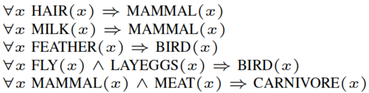
Reduce the required amount of data needed to train an image classification Neural Network by providing external knowledge in the form of Logic Rules.

Diligenti, Roychowdhury, and Gori, "Integrating Prior Knowledge into Deep Learning".

Logic Rules

in Image Classification with Semantic Based Regularization (Solution)

- Knowledge base in FOL form describing relationships between classes.
- Transform it into continuous constraints with Statistical Relational Learning.



$$\begin{aligned}
&\forall x\ \mathrm{HAIR}(x) \Rightarrow \mathrm{MAMMAL}(x)\\
&\forall x\ \mathrm{MILK}(x) \Rightarrow \mathrm{MAMMAL}(x)\\
&\forall x\ \mathrm{FEATHER}(x) \Rightarrow \mathrm{BIRD}(x)\\
&\forall x\ \mathrm{FLY}(x) \wedge \mathrm{LAYEGGS}(x) \Rightarrow \mathrm{BIRD}(x)\\
&\forall x\ \mathrm{MAMMAL}(x) \wedge \mathrm{MEAT}(x) \Rightarrow \mathrm{CARNIVORE}(x)
\end{aligned}$$

Diligenti, Roychowdhury, and Gori, "Integrating Prior Knowledge into Deep Learning".

Logic Rules

in Image Classification with Semantic Based Regularization (Solution)

$$C_e[\boldsymbol{f}(\mathcal{X})] = \sum_{k=1}^{T} \Bigg(\overbrace{||f_k||^2}^{Reg} + \overbrace{\lambda_l \sum_{x \in \mathcal{E}_k} L(f_k(x), y_k(x))}^{Labeled} \Bigg) + \\ + \overbrace{\sum_{h=1}^{H} \lambda_h L_c \left(\Phi_h\left(\boldsymbol{f}(\mathcal{X})\right)\right)}^{Logic} \qquad (1)$$

Diligenti, Roychowdhury, and Gori, “Integrating Prior Knowledge into Deep Learning”.

Algebraic Equations I

in Fluid Simulation (Definition)

Problem Definition

Predict the position and velocity of a particle in the next frame based on the position and velocity of all particles in the current frame.

Ladický et al., "Data-driven fluid simulations using regression forests".

Algebraic Equations I

in Fluid Simulation (Need for IML)

Required feature vector properties:

- Model the behavior of particles without explicitly calculating nearest neighbors.
- Evaluation of a particular dimension of a feature vector should be possible in constant time.
- Small input changes should cause small output changes.

Ladický et al., “Data-driven fluid simulations using regression forests”.

Algebraic Equations I

in Fluid Simulation (Solution)

Solution:

The final feature vector is a concatenation of the feature vectors of each component of the Navier-Stokes equations, calculated on a large fixed set of randomly sampled boxes R.

Ladický et al., "Data-driven fluid simulations using regression forests".

Algebraic Equations II

in Electrochemical Micro-Machining (μ-ECM) (Definition)

Problem Definition

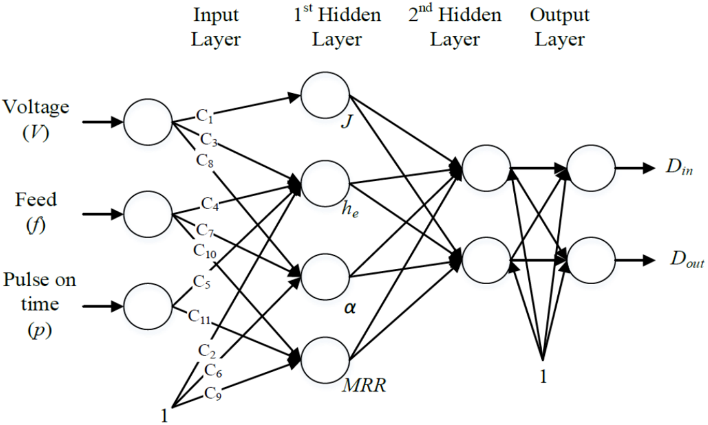
Enhance Input Parameters - Key Performance Indicator prediction by integrating prior knowledge in electrochemical micro machining processes.

Lu et al., "Physics-Embedded Machine Learning".

Algebraic Equations II

in Electrochemical Micro-Machining (μ-ECM) (Solution)

Four influential intermediate outputs are linearly connected with the input parameters.



Future Work

- Integration of Informed ML with Active Learning.
- Creation of a unified interdisciplinary method.