Advanced School in Artificial Intelligence

Introduction to Machine Learning

Elena Bellodi

elena.bellodi@unife.it

RegioneEmilia-Romagna

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Università degli Studi di Ferrara

Outline

- Machine learning (ML) definitions
- Learning paradigms
    - **supervised**
    - **unsupervised**
    - semi-supervised
    - reinforcement
- Use of Data in ML
    - training, validation and test set
    - generalization, underfitting and overfitting
    - capacity
    - bias and variance
- Learning protocols

Learning Paradigms

- **Learning with Different Outputs**
  - supervised
  - unsupervised
  - semi-supervised
  - reinforcement

- **Learning with Different Protocol**
  - batch learning
  - online learning
  - active learning

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Batch Learning

- Data is presented to the learning algorithm **in its entirety** *at the outset* of the learning process
    - batch of (email, spam?) ⇒ spam filter
    - batch of (patient, cancer) ⇒ cancer classifier

- very common in supervised learning

Online or Incremental Learning

- The data set is given to the algorithm one example at a time
  - *data streams* to be processed on the run (sensor data)
  - useful in case of limitations on computing and storage
- the model needs to be updated each time a new data point arrives

- Supervised learning if data are labeled
- **Reinforcement learning** if the hypothesis ‘**improves**’ through receiving instances *sequentially*
  - online spam filter:

1. observe an email $\mathbf{x}_t$
2. predict spam status with current $g_t(\mathbf{x}_t)$
3. receive ‘desired label’ y_t from user
4. update g_t with $(\mathbf{x}_t , y_t)$

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Active Learning

- ‘Question asking’ (sequentially): during the training stage query a user interactively about the y_n of the **chosen** $\mathbf{x}_{n,}$ as an iterative supervised learning
- Active VS ‘passive’ online learning: improve hypothesis with *fewer labels* (hopefully) by **asking questions strategically**
  - the algorithm could potentially reach a higher level of accuracy while using a smaller number of training labels if it were allowed to choose the data it wants to learn from
  - Useful when unlabeled data is abundant but manually labeling is expensive
- It is part of the human-in-the-loop paradigm
- It is a type of semi-supervised learning, meaning models are trained using both labeled and unlabeled data
- One of the most popular areas in active learning is *natural language processing (NLP)*

Bibliografia

- Peter Flach, «Machine Learning, The Art and Science of Algorithms that Make Sense of Data», 2012, Cambridge University Press
- [Michalski 1986] Michalski, R. S. “Understanding the nature of learning: Issues and research directions” in Michalski, R. S., Carbonell, J. G., and Mitchell, T. M., editors, Machine Learning - An Artificial Intelligence Approach, Volume II, Morgan Kaufmann Publishers, Los Altos, California, pages 3—26, 1986.
- [Simon 1984] Simon, H. A. “Why should machines learn” In Michalski, R. S., Carbonell, J. G., and Mitchell, T. M., editors, Machine Learning - An Artificial Intelligence Approach, Springer-Verlag, Berlin, pages 25—37, 1984.