Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Industriales

Máster en Automática y Robótica

Applied Artificial Intelligence

Assignment 1.1: Machine Learning Internet search

Josep María Barberá Civera (17048)

February 3, 2024

ASSIGNMENT 1.1

Machine Learning Internet search

First of all, let's define the term *Machine Learning* (ML). Wikipedia [1] can help us to get a first definition:

> "Machine learning (ML) is a **field of study** in **artificial intelligence** concerned with the development and study of **statistical algorithms** that can learn from data and generalize to unseen data, and thus perform tasks without explicit instructions."

Well, now we have a first idea of what the ML is. In the following sections we are going to go deeper in this understanding.

1.1 Key Concepts on ML

As a first source of information we can ask *ChatGPT* for the key concepts on ML. The output is very complete:

> "Data, Training, Supervised Learning, Unsupervised Learning, Validation and Testing, Feature Engineering, Overfitting and Underfitting, Algorithm Types, Bias and Fairness and Hyperparameters."

But we should consider other information sources like the second chapter from Sandra Vieira et al. [2]. There, they conclude with the key points about Machine Learning. Despite they focus on supervised ML only, and we think it could be noteworthy:

> "The supervised machine learning pipeline is comprised of six main stages: (i) **problem formulation**, (ii) **data preparation**, (iii) **feature engineering**, (iv) **model training**, (v) **model evaluation**, and (vi) **post hoc analysis**. A well-defined supervised machine learning problem should include a clear definition of a feature set, target variable, and task. Careful data preparation, including checking for outliers and possible confounding variables, should be carried out to avoid spurious interpretations. Feature engineering is a critical step in the standard machine learning pipeline; most of the time devoted to machine learning is spent on crafting useful features. Data transformations such as dimensionality reduction, automated feature selection, and scaling/normalization should be implemented as part of the CV scheme. There are several performance metrics to choose from; it is good practice to present the confusion matrix for classification problems, from which most classification metrics can be derived. Testing performance for statistical significance is especially important in [...] studies where samples tend to be small, resulting in a higher risk of overoptimistic results.

1.2 Most used ML techniques

From *ScienceDirect* we have found the next summary of remarkable ML techniques:

> "There are three major categories of algorithms used in machine learning: **supervised learning, unsupervised learning, and reinforcement learning**. (Aditya P. Mathur et al., 2022) Supervised learning algorithms require labeled data for training and include **Support Vector Machine (SVM), Naïve Bayes Classification, Mathematical Regression, Decision Trees, and Artificial Neural Networks (ANN)**. (Hoda K. Mohamed et al., 2021) Unsupervised learning algorithms do not require labeled data and include **clustering and Hidden Markov Model** (HMM). (Hoda K. Mohamed et al., 2021) Reinforcement learning algorithms learn through trial and error and are used in situations where the algorithm interacts with an environment to achieve a goal. (Aditya P. Mathur et al., 2022)."

But trough Scopus search it has been easily reported that currently there are some other moderns machine learning paradigms. For example the next ones are sourced from [3]:

- Multi-label learning;
- Semi-supervised learning;
- One-class classification;
- Positive-unlabeled learning;
- Transfer learning;
- Multi-task learning;
- Few/one-shot learning.

1.3 Most well-known ML courses (including MOOCs)

Based on [4] the best ML Courses for 2024 are the following:

- Machine Learning — Coursera
- Deep Learning Specialization — Coursera
- Machine Learning Crash Course — Google AI
- Machine Learning with Python — Coursera
- Advanced Machine Learning Specialization — Coursera*
- Machine Learning — EdX
- Introduction to Machine Learning for Coders — Fast.ai

1.4 Most famous ML data bases and competitions

Regarding the data bases the top ten can be the following: QuestDB (for Time Series), SingleStore (a Data-Intensive database), ClickHouse (for real-time ML apps), MindsDB, Datastax & Cassandra (streaming), MariaDB, mongoDB but also Redis and of course Kaggle.

Regarding the most popular ML competitions, based on [5] the list should include: Kaggle, DrivenData, Alcrowd, Zindi, CodaLab, Tianchi, Signate, EvalAI, Waymo, Xeek, DS Works, MOFC, Micropredicion, Battlecode, Topcoder, Hugging Face Competitions, ML Contest.

References

[1] Wikipedia contributors. *Machine learning — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=1202291005. [Online; accessed 3-February-2024]. 2024.

[2] S. Vieira, W. H. Lopez Pinaya, and A. Mechelli. "Chapter 2 - Main concepts in machine learning". In: *Machine Learning*. Ed. by A. Mechelli and S. Vieira. Academic Press, 2020, pp. 21–44. ISBN: 978-0-12-815739-8. DOI: https://doi.org/10.1016/B978-0-12-815739-8.00002-X. URL: https://www.sciencedirect.com/science/article/pii/B978012815739800002X.

[3] F. Emmert-Streib and M. Dehmer. "Taxonomy of machine learning paradigms: A data-centric perspective". In: *WIREs Data Mining and Knowledge Discovery* 12.5 (2022), e1470. DOI: https://doi.org/10.1002/widm.1470. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1470. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1470.

[4] B. Martin. *7 best machine learning courses for 2024*. Jan. 2024. URL: https://www.learndatasci.com/best-machine-learning-courses/.

[5] *ML Contests — Top Machine Learning Competition Platforms*. Mar. 2023. URL: https://mlcontests.com/competition-platforms/.