Raw Parsed


```
{
  "query": "machine learning",
  "num_results": 3,
  "results": [
    {
      "chunk": "Springer Nature 2021 LATEX template Learning Curves for Decision Making in Supervised Machine Learning: A Survey Felix Mohr1* and Jan N. van Rijn2 1Universidad de La Sabana, Ch´ıa, Cundinamarca, Colombia. 2Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands. *Corresponding author(s). E-mail(s): felix.mohr@unisabana.edu.co; Contributing authors: j.n.van.rijn@liacs.leidenuniv.nl; Abstract Learning curves are a concept from social sciences that has been adopted in the context of machine learning to assess the performance of a learn- ing algorithm with respect to a certain resource, e.g., the number of training examples or the number of training iterations. Learning curves have important applications in several machine learning contexts, most notably in data acquisition, early stopping of model training, and model selection. For instance, learning curves can be used to model the per- formance of the combination of an algorithm and its hyperparameter configuration, providing insights into their potential suitability at an early stage and often expediting the algorithm selection process. Var- ious learning curve models have been proposed to use learning curves for decision making. Some of these models answer the binary deci- sion question of whether a given algorithm at a certain budget will outperform a certain reference performance, whereas more complex mod- els predict the entire learning curve of an algorithm. We contribute a framework that categorises learning curve approaches using three cri- teria: the decision-making situation they address, the intrinsic learning curve question they answer and the type of resources they use. We survey papers from the literature and classify them into this framework. Keywords: learning curves, supervised machine learning 1 arXiv:2201.12150v2 [cs.LG] 28 Jan 2025 Springer Nature 2021 LATEX template 2 Learning Curves for Decision Making 1 Introduction Learning curves describe a system’s performance on a task as a function of some resource to solve that task. There can be a pre-defined budget of that resource, limiting the amount of resources that can be spent. In other cases, the goal can be to obtain reasonable results while minimising the spent budget of that resource. Typical types of budgets are the number of examples the learner has observed before performing the task or the number of iterations or time the learner spends in an environment. The performance measure expresses the quality of the obtained model, e.g., error rate or F1 measure. Learning curves are an important source of information for making decisions on the following matters in machine learning: • Data Acquisition determines how many data points should reasonably be acquired to obtain a desired performance. The top right plot in Fig. 1 visu- alises a scenario where we have already observed performance up to a certain amount of data (the blue learning curve). We can extrapolate this and make a prediction of what the performance would be if more data was available, i.e., the value of the orange extrapolation at different vertical pink lines in the figure (see, e.g., Last, 2009; Weiss and Tian, 2008). • Early Stopping of training a model. If we are committed to some specific learner (a learning algorithm and its hyperparameters), we might want to minimise the",
      "metadata": {
        "paper_id": "2201.12150v2",
        "chunk_index": 0,
        "total_chunks": 64
      },
      "distance": 1.0289952754974365,
      "rank": 1
```

```
    },
    {
      "chunk": "may be difficult or impossible for humans to discern. The typical process
      of designing a machine learning algorithm consists of two main phases: training and
      inference. During the training phase, the parameters of a machine learning model are
      tuned to solve a specific task. For this, a wide variety of data sources can be used,
      or even outputs from existing or pre-trained machine learning models. After the model
      is trained, its parameters are kept fixed so that the model can be used to predict
      outcomes or identify patterns in new or previously unseen data. This process is
      called inference. It is important to keep in mind the distinction between training
      and inference. After training, the model remains unchanged, and it remains unchanged
      until it is retrained again. Later, in Section 3, we will focus on the disparities
      that this can cause w.r.t. the inference phase and, in consequence, official
      statistics production. Supervised learning is one of the most common types of machine
      learning used for official statistics. This method involves training a model on a
      labeled dataset, where each data point has a known outcome or target variable. The
      model learns to associate features in the data with the target variable, enabling it
      to make predictions on new data with similar features. In the context of official
      statistics, supervised learning can for example be used to predict the happiness of
      an individual based on their Twitter profile [8]. Unsupervised learning, on the other
      hand, is used when the target variable is unknown or the goal is to identify patterns
      or relationships within the data. In this approach, the machine learning model learns
      to recognize similarities and differences among input data without explicit guidance
      from labeled data. In the context of official statistics, unsupervised learning can
      for example be used to identify citizens, companies or events that are similar to
      each other on one or more aspects that could be hidden from plain sight [9]. Machine
      learning can be used to complement or even replace official statistics, and its
      ability to nowcast and forecast is an extremely valuable addition. Modern machine
      learning models, tools and hardware can analyze vast amounts of data in real-time or
      near-real-time, providing more up-to-date and precise estimates of e.g. economic and
      social trends. By 1Although originally (and technically) they imply different methods
      and techniques, the terms machine learning, data science, artificial intelligence,
      deep learning... are nowadays considered interchangeable. In this paper we
      consistently use the term machine learning to denote the scientific discipline
      concerned with learning the most optimal model parameters based on data. It is a
      subdiscipline of artificial intelligence, while deep learning is a subdiscipline of
      machine learning. Data science encompasses both machine learning as well as data
      preparation, analytics and visualization. 2 incorporating machine learning into
      official statistical production, one can benefit from the strengths of both
      approaches and make more informed decisions based on the most current and accurate
      data [10]. 2.2 External data sources Let’s focus on the data sources that will power
      such machine learning models. Their nature, size, structure, frequency... can",
      "metadata": {
        "paper_id": "2306.04338v1",
        "chunk_index": 2,
        "total_chunks": 13
      },
      "distance": 1.04233980178833,
      "rank": 2
    },
    {
      "chunk": "Wierstra, D. Matching networks for one shot learning. in 3630–3638 (2016).
      31. Mehta, P. et al. A high-bias, low-variance introduction to machine learning for
```

```
        physicists. Phys. Rep. (2019). 32. Guyon, I. & Elisseeff, A. An introduction to
        variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182 (2003). 33. He, J.
        et al. The practical implementation of artificial intelligence technologies in
        medicine. Nat. Med. 25, 30–36 (2019). 34. Rudin, C. Stop explaining black box machine
        learning models for high stakes decisions and use interpretable models instead. Nat.
        Mach. Intell. 1, 206–215 (2019). 35. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.
        A. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification:
        an overview. Bioinforma. Oxf. Engl. 16, 412–424 (2000). 36. Goecks, J., Nekrutenko,
        A., Taylor, J., & Galaxy Team. Galaxy: a comprehensive approach for supporting
        accessible, reproducible, and transparent computational research in the life
        sciences. Genome Biol. 11, R86 (2010). 37. Di Tommaso, P. et al. Nextflow enables
        reproducible computational workflows. Nat. Biotechnol. 35, 316–319 (2017). 38.
        Arrieta, A. B. et al. Explainable Artificial Intelligence (XAI): Concepts,
        taxonomies, opportunities and 15 challenges toward responsible AI. Inf. Fusion 58,
        82–115 (2020). 39. Guidotti, R. et al. A survey of methods for explaining black box
        models. ACM Comput. Surv. CSUR 51, 1–42 (2018). 40. Adadi, A. & Berrada, M. Peeking
        inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE
        Access 6, 52138–52160 (2018). 41. Holm, E. A. In defense of the black box. Science
        364, 26–27 (2019). 42. O'Mahony, S. The governance of open source initiatives: what
        does it mean to be community managed? J. Manag. Gov. 11, 139–150 (2007). 43. Brazma,
        A. et al. Minimum information about a microarray experiment (MIAME)—toward standards
        for microarray data. Nat. Genet. 29, 365–371 (2001). 44. Hermjakob, H. et al. The
        HUPO PSI's Molecular Interaction format—a community standard for the representation
        of protein interaction data. Nat. Biotechnol. 22, 177–183 (2004). 16 Supplementary
        Material Machine learning summary table: 17 DOME Version 1.0 Data Provenance Source
        of data, data points (positive, Npos / negative, Nneg). Used by previous papers
        and/or community. Dataset splits Size of Npos and Nneg of training set, validation
        set (if present), test set. Distribution of Npos and Nneg across sets. (section to be
        repeated for each dataset) Redundancy between data splits Independence between sets.
        Strategy used to make examples representative (e.g. eliminating data points more
        similar than X%). Comparison relative to other datasets. Availability of data Yes/no
        for datasets. If yes: Supporting information, website URL, license. Optimization
        Algorithm ML class (e.g. neural network, random forest, SVM). If novel approach,
        reason is it not previously published. (section to be repeated for each trained
        model) Meta-predictions Yes/No. If yes: how other methods are used and whether the
        datasets are clearly independent. Data encoding How input data is transformed (e.g.
        global features, sliding window on sequence). Parameters Number of ML model
        parameters (p), e.g. tunable weights in neural networks. Protocol used to select p.
        Features Number of ML input features (f), i.e. encoding of data points. In case of
        feature selection: Protocol used, indicating whether",
      "metadata": {
        "paper_id": "2006.16189v4",
        "chunk_index": 12,
        "total_chunks": 16
      },
      "distance": 1.0742677450180054,
      "rank": 3
    }
  ]
}
```