Stage 2

Behind the Abstraction: Visualizing the Inner Workings of Machine Learning Models

Tudor-Ioan Gălățan

Table of Contents

Chapter Sketch:

1.1. General Context and Topic description

This subsection provides a general overview of machine learning in industry, focusing on its impact on solving complex problems. The need for transparency and interpretability of industrial models is briefly introduced.

1.2. Relevance and Thesis Motivation

This subsection outlines the importance of machine learning in contemporary software systems, emphasizing its contribution to intelligent and automated features. The focus is on its widespread adoption in everyday applications.

1.3. Purpose and Objectives of the Thesis

This subsection introduces interactive educational applications as a means to make complex topics more accessible. Particular attention is given to their usefulness in teaching machine learning concepts.

1.4. Original Contributions

This subsection covers the original contributions of the thesis, focusing on the design and development of an interactive educational web application. It highlights the integration of visualizations and automated explanations of machine learning models.

CHAPTER 2 - THEORETICAL BACKGROUND AND STATE OF THE ART

2.1. Fundamental Concepts of Machine Learning

This subsection provides an overview of the core concepts in machine learning, introducing key elements such as data, features, labels, models, training, and prediction. It emphasizes the importance of understanding these concepts as a foundation for exploring different learning approaches and their practical applications.

2.2. Stages of a Machine Learning Process

This subsection presents the main stages of a machine learning process, from data collection to model interpretation. It highlights the sequential steps required to build and evaluate effective models.

2.3. Machine Learning models used in the application

This subsection presents the machine learning models integrated into the application, highlighting their operation and suitability for educational visualization. It emphasizes how each model's training process can be interactively explored on the web interface.

2.4. Evaluation of Machine Learning Model Performance

This subsection describes the process of evaluating machine learning models using training, validation, and test sets. It highlights the visualization of metrics and learning curves for educational purposes.

2.5. Visualization of the Machine Learning Process

This subsection describes how visualizations can reveal the inner workings of machine learning models. The web interface allows users to explore model behavior step by step, improving comprehension.

2.6. Interpretability and Explanation of Results

This subsection presents methods for making model results understandable, highlighting approaches for explaining predictions. The application provides explanations adapted to different user expertise levels.

2.7. Web Applications for Machine Learning

This subsection presents the role of frontend and backend separation in ML web applications. The focus is on enabling real-time interaction and visualization of model training and predictions.

2.8. Analysis of existing solutions

This subsection reviews existing educational tools for machine learning, such as Google Teachable Machine and scikit-learn demos. It highlights their strengths and limitations in visualizing model training.

CHAPTER 3 - CASE STUDY: DESIGN AND IMPLEMENTATION OF A WEB APPLICATION

3.1. Application Requirements Specification

This subsection outlines the functional and non-functional requirements of the web application. It emphasizes features that enable interactive visualization and step-by-step exploration of ML models.

3.2. Application Architecture and design

This subsection presents the overall architecture and design of the web application. It emphasizes the separation of frontend and backend to support interactive ML visualizations.

3.3. Application implementation

This subsection describes the coding and development process, focusing on integrating model training, data processing, and graphical representations. Special attention is given to user interaction and step-by-step visual feedback.

3.4. Application Deployment and Execution without an IDE

This subsection describes the steps required for application deployment and standalone execution. Focus is placed on ensuring users can access all interactive features without development tools.

3.5. Application Testing

This subsection presents the testing procedures for the application, highlighting evaluation of both backend computations and frontend visualizations. It stresses the importance of reliable and educational user experiences

3.6. Experimental Results and Their Evaluation

This subsection reports on experiments conducted with the application, highlighting model outputs and step-by-step learning processes. It stresses interpreting results to assess both functionality and accuracy..

CHAPTER 4 - CONCLUSIONS AND FUTURE WORK

4.1. General Conclusions

This subsection summarizes the main findings of the thesis, highlighting the achievement of objectives and the effectiveness of interactive ML visualizations.

4.2. Limitations of the Proposed Solution

This subsection discusses the limitations of the proposed application, including scalability, model diversity, and potential user interface constraints.

4.3. Future Research and Development Directions

This subsection outlines opportunities to improve and extend the application, focusing on scalability, usability, and advanced ML explanations.

Main documentation for the thesis

1.1 Introduction to the Field of the Thesis

A general definition of machine learning

The role of machine learning in modern software applications

The importance of interactive educational applications

Recent research emphasizes that while machine learning models achieve high predictive performance, their lack of transparency can reduce trust and understanding among users. Doshi-Velez and Kim (2017) argue for the necessity of interpretable machine learning, especially in educational and decision-support contexts. Visualization plays a key role in bridging this gap by transforming abstract numerical processes into intuitive graphical representations.

1.2 Context and Motivation for Choosing the Topic

Based on the content presented in Stage 1

1.3 Purpose and Objectives of the Thesis

Description of the general purpose (based on Stage 1)

The proposed application aligns with human-centered artificial intelligence principles, aiming to support users with varying levels of expertise. According to Amershi et al. (2019), effective human–AI interaction requires systems to provide clear feedback, transparency, and understandable explanations. Therefore, the application automatically generates descriptive explanations of model behavior and predictions, tailored to entry-level users.

CSV dataset upload

Training machine learning models

Visualization of the learning process

Automatic explanation of results

Model comparison

Definition of the objectives set through the proposed application

1.4 Fundamental Concepts of Machine Learning

1.4.1 Definition and Core Concepts of Machine Learning

Many modern machine learning models are considered “black boxes” due to the difficulty of understanding their internal decision-making processes. Molnar (2022) highlights that interpretability is essential for learning, debugging, and trusting machine learning systems, particularly in educational environments.

Definition of machine learning

Data, features, labels

Model, training, prediction

Definition of the above-mentioned terms

1.4.2 Types of Machine Learning

Clear categorization of machine learning paradigms supports conceptual understanding. Géron (2022) highlights that simple, visual examples of supervised and unsupervised learning help beginners grasp abstract learning mechanisms more effectively.

Supervised learning

Unsupervised learning

Simple illustrative examples

1.4.3 Stages of a Machine Learning Process

Data collection

Data preprocessing

Model training

Model evaluation

Model interpretation

1.5 Machine Learning Models Used in the Application

For educational purposes, models with transparent internal structures are preferred. Domingos (2012) argues that simpler models often provide sufficient performance while offering superior interpretability, making them suitable for instructional environments.

For each of the following models, the following aspects are addressed:

Theoretical description

Operating principle

Advantages and disadvantages

Reasons why the model is suitable for an educational application

1.5.1 Regression

- Cost function
- Visual interpretation

1.5.2 K-Nearest Neighbors Algorithm

- Distance metrics
- Neighbors
- Influence of the value of K
- Clear decision boundaries

1.5.3 Support Vector Machines

- Hyperplane
- Margin
- Kernel functions

1.5.4 Decision Trees

- Tree structure
- If–then rules
- High interpretability

1.6 Evaluation of Machine Learning Model Performance

Evaluation metrics must be interpreted in context to avoid misleading conclusions. Powers (2011) explains that metrics such as accuracy, precision, and recall provide complementary perspectives on model performance, reinforcing the need for visual comparison and explanation.

1.6.1 Datasets: Training, Validation, and Testing

- Training, validation, and test sets
- Importance of data separation

1.6.2 Evaluation Metrics

- Accuracy
- Precision
- Recall
- Mean Squared Error (MSE)

1.6.3 Overfitting and Underfitting of Models

Visualization of learning curves is a widely used technique for diagnosing overfitting and underfitting. Hohman et al. (2019) emphasize that visual analytics allows users to detect performance issues early by observing the evolution of training and validation metrics over time.

- Definitions
- Impact on model performance
- Visualization of learning curves

1.7 Visualization of the Machine Learning Process

Effective visualization is not only a technical challenge but also a design problem. Tufte (2001) and Knaflic (2020) stress that clear, minimal, and well-structured visual representations improve comprehension and reduce cognitive load. These principles guide the visual design choices of the proposed application.

Visualization techniques are central to visual analytics, a field that combines data analysis and interactive visualization. Keim et al. (2008) state that visual analytics enables users to gain insight into complex computational processes, including machine learning training and prediction.

1.7.1 Importance of Visualization in Understanding ML Models

1.7.2 Visualization of Decision Boundaries

- Definition and role in model interpretation
- Examples for K-Nearest Neighbors and Support Vector Machines

1.7.3 Visualization of Learning Curves and Loss Functions

- Performance evolution
- Detection of overfitting

1.8 Interpretability and Explanation of Results

Explanation methods such as local, instance-based explanations have been shown to improve user trust. Ribeiro et al. (2016) demonstrate that explaining individual predictions helps users understand why a model produced a specific output, an approach adopted by the application through step-by-step explanatory messages.

Trust in machine learning systems is strongly correlated with the quality of explanations provided. Lipton (2018) argues that interpretability is a prerequisite for meaningful human–machine interaction, particularly when models are used by non-experts.

1.8.1 The Need for Interpretability in Machine Learning

1.8.2 Methods for Explaining Predictions

1.8.3 Automatic Explanation of Results for Non-Technical Users

- Generation of descriptive messages
- User-adapted explanations
- Educational role

1.9 Web Applications for Machine Learning

Human-centered AI systems require architectures that support real-time interaction and feedback. Shneiderman (2020) highlights that responsive interfaces and transparent computation processes are essential for trustworthy AI systems, reinforcing the choice of a frontend–backend separation.

Web-based machine learning applications enable accessibility and rapid experimentation. Kandel et al. (2012) emphasize that interactive systems support exploratory learning by allowing users to manipulate data and observe immediate effects.

1.9.1 Modern Web Architectures for ML Applications

- Client–server architecture
- REST APIs

1.9.2 Frontend–Backend Separation in ML Applications

- Frontend: user interface and visualization
- Backend: machine learning and data processing

1.9.3 Technologies Used in the Development of ML Web Applications

- Web frameworks
- Machine learning libraries

1.10 Existing Tools and Platforms for Machine Learning Visualization

1.10.1 General Presentation of Existing Solutions

- Google Teachable Machine
- scikit-learn demos
- Other educational tools

1.10.2 Limitations of Current Solutions

- Lack of personalization
- Limited explanations
- Focus on demos rather than learning

1.10.3 Positioning of the Proposed Application Relative to Existing Solutions

Unlike existing platforms that primarily demonstrate outcomes, the proposed application focuses on explaining internal mechanisms. According to Fails and Olsen (2003), interactive exploratory tools are most effective when users can observe cause–effect relationships, a principle central to this thesis.

- Novel contributions of the proposed application
- Clear comparison with existing tools

Bibliografy

Spinner, T., Schlegel, U., Schäfer, H., & El-Assady, M. (2019). **explAIner: A visual analytics framework for interactive and explainable machine learning**. *arXiv preprint arXiv:1908.00087*. https://arxiv.org/abs/1908.00087

Hohman, F., Park, H., Robinson, C., & Chau, D. H. (2023). **WizMap: Scalable interactive visualization for exploring large machine learning embeddings**. *arXiv preprint arXiv:2306.09328*. https://arxiv.org/abs/2306.09328

Chatzimparmpas, A., Martins, R. M., Jusufi, I., & Kerren, A. (2022). **The state of the art in enhancing trust in machine learning models with the use of visualizations**. *arXiv preprint arXiv:2212.11737*. https://arxiv.org/abs/2212.11737

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. https://arxiv.org/abs/1702.08608

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, 1135–1144. https://arxiv.org/abs/1602.04938

Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable. *arXiv preprint arXiv:2203.05614*. https://arxiv.org/abs/2203.05614

Tufte, E. R. (2001). The visual display of quantitative information. *Graphics Press*.

Biecek, P., & Burzykowski, T. (2021). Explanatory model analysis: Explore, explain, and examine predictive models. *CRC Press*.

Knaflic, S. N. (2020). Storytelling with data: A data visualization guide for business professionals. *Wiley*.

Hohman, F., Kahng, M., Pienta, R., & Chau, D. H. (2019). Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25(8), 2674–2693. https://arxiv.org/abs/1801.06889

Amershi, S., et al. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. https://arxiv.org/abs/1803.07228

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4). https://onlinelibrary.wiley.com/doi/10.1002/widm.1312

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4).

Woolf, B. P. (2010). Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning. *Morgan Kaufmann*. https://www.sciencedirect.com/book/9780123735942/building-intelligent-interactive-tutors

Géron, A. (2022). Hands-on machine learning with scikit-learn, Keras, and TensorFlow. *O'Reilly Media*. https://www.oreilly.com/library/view/hands-on-machine-learning/9781098125967/

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87. https://doi.org/10.1145/2347736.2347755

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63. https://arxiv.org/abs/2010.16061

Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., & Ziegler, H. (2008). Visual analytics: Scope and challenges. *Springer Lecture Notes in Computer Science*, 4950, 76–90. https://link.springer.com/chapter/10.1007/978-3-540-70956-5_7

Liu, S., Wang, X., Liu, M., & Zhu, J. (2017). Towards better analysis of machine learning models: A visual analytics perspective. *IEEE Computer Graphics and Applications*, 37(4), 84–93. https://ieeexplore.ieee.org/document/8019837

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43. https://arxiv.org/abs/1606.03490

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., & Van Ham, F. (2012). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 11(4), 271–288. https://doi.org/10.1177/1473871612451444

Fails, J. A., & Olsen, D. R. (2003). Interactive machine learning. *Proceedings of the 8th International Conference on Intelligent User Interfaces*, 39–45. https://dl.acm.org/doi/10.1145/604045.604056