Lab 2

Task B: Supervised Learning

Alejandro Silva Rodríguez

Marta Cuevas Rodríguez

Aprendizaje Computacional
Universidad de Málaga

October 2024

Contents

1 Introduction

This project explores both fundamental and advanced concepts in supervised learning, placing a particular emphasis on the significant potential of Deep Learning in biomedical fields. Supervised learning, a technique where models learn to classify or predict outcomes based on labeled data, is foundational in machine learning, particularly for applications involving complex classification tasks. However, with the advent of Deep Learning, models now have the capacity to process and analyze high-dimensional data with enhanced accuracy and flexibility, making them highly effective in medical and health-related applications.

In this project, we begin by examining essential concepts, such as partitioning datasets into training and testing sets to evaluate model generalizability, and the use of cross-validation to optimize model performance. We then delve into the structure of neural networks, differentiating traditional feed-forward networks from more advanced architectures used in Deep Learning. These modern architectures—characterized by multiple hidden layers, complex connections, and sophisticated activation functions—enable models to learn hierarchical representations of data, capturing intricate patterns and dependencies.

Moreover, this project addresses critical challenges that arise when applying Deep Learning to biomedical contexts. These include the risk of overfitting, where models may perform well on training data but fail to generalize to new data, and the common issue of data scarcity, given the limited availability of labeled biomedical datasets. We will explore techniques to alleviate these challenges, such as regularization, data augmentation, and synthetic data generation, which expand the training dataset and improve model robustness.

Finally, we will review recent advancements in Deep Learning applied to healthcare, showcasing examples where such models have significantly improved diagnostics, disease prediction, and patient outcomes. This project aims to build a comprehensive understanding of how supervised learning, and Deep Learning in particular, can drive impactful changes in the healthcare sector.

1.1 Objectives

- **Develop a foundational understanding of supervised learning techniques**: Explore how supervised learning methods work, emphasizing their role in solving classification tasks in healthcare.
- **Examine Deep Learning concepts and architectures**: Differentiate traditional feed-forward networks from advanced Deep Learning architectures, analyzing the benefits of additional hidden layers, complex connections, and hierarchical data representation.
- **Understand data partitioning and evaluation**: Learn the importance of splitting datasets into training and test sets to ensure model performance generalizes to new data, and explore cross-validation as a technique to further enhance reliability and robustness.
- **Implement overfitting prevention techniques**: Identify strategies to reduce overfitting, such as regularization, dropout, and early stopping, ensuring that models perform well not only on training data but also on new, unseen data.
- **Address data scarcity through data augmentation and synthetic data generation**: Explore how data scarcity impacts biomedical applications and implement data augmentation and synthetic data generation techniques to expand limited datasets and enhance model accuracy.
- **Analyze real-world applications of Deep Learning in healthcare**: Review case studies where Deep Learning has been successfully applied in the medical field, such as diagnostics and predictive models for disease progression, highlighting its potential to transform patient care.

2 Supervised Learning

Supervised learning is a foundational approach within machine learning, widely researched and applied across various fields, including multimedia content processing. The core feature that distinguishes supervised

learning is its use of annotated training data, where each example is paired with a known label, often representing a class in classification tasks. This setup is akin to having a "supervisor" guiding the learning system by providing correct associations for each training sample [3].

Through supervised learning algorithms, models are derived from the training data to enable the classification of new, unlabeled examples. These models are trained with the aim of minimizing prediction errors—a concept underpinned by the theory of risk minimization. Techniques such as support vector machines and nearest neighbor classifiers, two highly popular algorithms in multimedia and other domains, exemplify the methods utilized to implement supervised learning, as they effectively leverage labeled data to create accurate and adaptable models [3].

2.1 Importance of Dataset Splitting in Supervised Learning

Dividing a dataset into training and test sets is a critical step in developing supervised learning models, especially in fields such as biomedical classification. This separation ensures that the model's performance is evaluated on data it has not encountered before, providing an honest assessment of its generalization ability [4]. Careful consideration of sampling methods during the training and testing stages directly impacts the system's stability and performance, as inappropriate sample selection can lead to misleading evaluations and potential overfitting.

2.2 Cross-Validation in Model Evaluation and Selection

Cross-validation is a widely-used data resampling technique in model evaluation and selection, serving multiple purposes, including hyperparameter tuning, preventing overfitting, and estimating the generalization error of predictive models. By partitioning the dataset into multiple subsets or "folds," cross-validation allows a model to train on various splits and validate on the remaining portions, leading to a more reliable performance estimate [1].

Among the most common cross-validation methods are k-fold cross-validation, leave-one-out cross-validation (LOOCV), and nested cross-validation. In particular, nested cross-validation is valuable for tuning hyperparameters, as it involves two layers of cross-validation: the outer loop for testing and the inner loop for training and validation. This approach is especially beneficial in complex models where a single validation loop might underestimate the generalization error.

Cross-validation is crucial for enhancing models, particularly by mitigating overfitting and enabling robust comparisons among learning algorithms. In deep learning, however, standard cross-validation is less common due to the high computational cost. Instead, methods such as single hold-out validation or using a dedicated validation set within an iterative training process are more frequently used, allowing large models to be evaluated without prohibitive resource requirements [1].Cross-validation is less commonly used in deep learning due to the high computational costs, especially with large datasets and deep networks. Instead, deep learning models often use a fixed validation set, allowing for efficient hyperparameter tuning and performance evaluation without the resource demands of full cross-validation.

2.3 Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational models inspired by the complex structure and functioning of the human brain, where billions of interconnected neurons process information simultaneously. Just as the brain can recognize patterns, make decisions, and process language, ANNs have been developed to perform tasks like language translation, image recognition, and time series forecasting. ANNs operate as universal function approximators, meaning they can map input data to outputs in a highly flexible manner, making them suitable for a wide variety of applications [6].

An ANN consists of multiple layers of artificial neurons or nodes, which are organized in an architecture of interconnected layers. These layers typically include an input layer, one or more hidden layers, and an output layer. Each neuron in a given layer processes the input it receives, applies a weighted sum and an

activation function, and passes the output to the next layer. Through training, these weights are adjusted to minimize the difference between the network's prediction and the true output, allowing the ANN to "learn" patterns in data.

3 Deep learning

- What is Deep Learning (DL)? lo mismo pero con muchos datos y a lo mejor modelos mas buenos tipo transformers y esas cosas

- What is new in DL models with respect to traditional feedforward neural networks? supongo que los modelos son mas complejos pero ni idea

tienes q hablar de transformers si o si de este trabajo:[5] y ya si quieres de cnn o rnn nose q mas

- Google (among others) has produced astonishing results in the application of DL models in different domains. Mention two of these cases describing shortly the problem solved. **Alphafold Gemini** inserta textaco y referenxias

- What is overfitting and how DL models avoid it? es cuando tu modelo deja de generalizar y se aprende los datos concretos con los que se alimenta y hace que no sirva para nada. se soluciona con muchos datos diferentes, o capas de normalizacion y esa vaina

- Lack of data is a big limitation regarding the application of DL models to biomedical problems. What techniques can be applied to alleviate this problem.

data augmentation techniques. pues cosas de estadistica o darle la vuelta a las imagenes o generar datos con otra ia . por ejemplo la adversative neural network q dijo el profe que creo que enfrenta a dos nn a ver cual inventa cosas mejor

4 Repository Access

All additional information, including the source code and full documentation of this project, is available in the GitHub repository [2].

References

[1] Daniel Berrar et al. Cross-validation., 2019.

[2] Marta Cuevas. Lab2_computational_learning. `https://github.com/martacuevasr/Lab2_Computational_learning`, 2024. Último acceso: 1 noviembre 2024.

[3] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. *Supervised Learning*, pages 21–49. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[4] Muhammed Kürşad Uçar, Majid Nour, Hatem Sindi, and Kemal Polat. The effect of training and testing process on machine learning in biomedical datasets. *Mathematical Problems in Engineering*, 2020(1):2836236, 2020.

[5] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[6] Sun-Chong Wang and Sun-Chong Wang. Artificial neural network. *Interdisciplinary computing in java programming*, pages 81–100, 2003.