Covid Pneumonia Prediction

By

Mohamed Hisham (20200483)

Yousef Mohamed (20200669)

Mohamed Nasser (20200480)

Amira Ashraf (20200094)

Artificial Intelligence Department
Cairo University
2023-2024

Table of Contents

1. Introduction

Machine learning (ML) is a field of artificial intelligence (AI) that allows computers to learn from data without being explicitly programmed. Instead of writing out every single instruction, ML algorithms can identify patterns and relationships in data, and then use those insights to make predictions or decisions on new data.

Imagine you are training a machine learning model to recognize handwritten digits. You would feed the model a bunch of images labeled with the corresponding digit (1, 2, 3, etc.). The model learns the underlying patterns that differentiate a seven from a nine by analyzing these examples. Then, when presented with a new, unseen image, the model can use its knowledge to predict the digit accurately.

Here are some key things to remember about machine learning:

- Learning from data: ML algorithms are about extracting knowledge from data. The more data you have, the better the model will perform.
- Different learning styles: There are various machine learning approaches, categorized based on how data is presented. In supervised learning, the data is labeled with the desired output, while unsupervised learning deals with unlabeled data and finds hidden patterns.
- Making predictions: A core function of machine learning models is to make predictions on new, unseen data. This could be anything from recognizing faces in images to predicting stock prices.

Machine learning has become a powerful tool across various industries, from healthcare and finance to manufacturing and entertainment. If you are interested in learning more, there are many resources available online and in libraries to delve deeper into this fascinating field.

2. Objectives

There are several compelling objectives for using machine learning (ML) instead of deep learning (DL) even though DL is a powerful subset of ML. Here is a breakdown of some key reasons:

- **Simpler data requirements:** Deep learning models often require vast amounts of labeled data to train effectively. For problems with limited data availability, traditional ML algorithms can be more efficient, requiring less data for reasonable performance.
- **Interpretability:** Understanding how an ML model arrives at a decision can be crucial in certain situations. Machine learning models are easier to interpret than deep learning models, whose complex architectures can make it difficult to pinpoint the reasoning behind their outputs. This interpretability can be essential in fields like finance or medicine where understanding the "why" behind a decision is critical.
- **Computational efficiency:** Training deep learning models often demands significant computational resources. Machine learning algorithms can be computationally less expensive to train and run, making them suitable for situations with limited computing power or where real-time responsiveness is essential.
- **Focus on specific tasks:** Machine learning offers a diverse range of algorithms specifically designed for various tasks like linear regression for prediction or k-nearest neighbors for classification. These specialized algorithms might outperform deep learning models for well-defined problems.
- **Clearer problem definition:** Machine learning often thrives with well-defined problems where the features (data points) relevant to the task are identified. Deep learning excels at uncovering complex patterns in data, but if the problem itself is not clearly defined, it might struggle to find the right direction.

3. Methodology

3.1. Data (**Chest X-ray Images of COVID-19 Patients)**

This dataset contains chest X-ray images from patients diagnosed with COVID-19. The images are categorized into two classes:

- COVID-19 with Pneumonia: X-ray images of patients confirmed to have both COVID-19 infection and pneumonia. Pneumonia is an inflammation of the lungs that can be caused by several factors, including viral infections like COVID-19. In these images, you might expect to see signs of lung opacities, which are areas where the lungs appear whiter than usual due to fluid buildup.
- COVID-19 without Pneumonia: X-ray images of patients confirmed to have COVID-19 infection but without signs of pneumonia. These images may show normal lung appearances or subtle abnormalities not indicative of pneumonia.

3.2. Techniques Used

3.2.1. Feature Extraction (Centroid)

The centroid of an image, in the context of feature extraction, refers to the midway point that represents the "average" location of all the pixels within the image. It is a simple but useful technique for capturing the overall distribution of mass or intensity in an image.

Advantages:

- Simplicity: Calculating the centroid is a straightforward process, making it computationally efficient.
- Interpretability: The centroid's meaning is clear - it represents the midway point.

Limitations:

- Limited Information for Complex Shapes: For complex shapes with irregular distributions of pixels, the centroid might not provide sufficient information for accurate feature extraction.
- Sensitivity to Noise: Outliers or noise in the image data can significantly influence the centroid's location.

3.2.2. Optimizer (PSO)

What it is:

- PSO is a computational optimization technique inspired by the collective movement of swarming animals like birds or fish.
- It aims to find an optimal solution to a problem by iteratively improving candidate solutions through a population-based approach.

How it works:

1. Particle Swarm: A set of virtual particles represents candidate solutions in the search space of the problem. Each particle has a position and a velocity.
2. Personal Best: Each particle keeps track of its own best position encountered so far, which represents the best solution it has found individually.
3. Global Best: The entire swarm also tracks the best position discovered by any particle within the swarm, representing the current best solution found collaboratively.
4. Movement: Particles move through the search space based on their current velocity. This velocity is influenced by two factors:
    - Attraction to Personal Best: Particles are drawn towards their own best positions, encouraging them to explore promising regions they have discovered.
    - Attraction to Global Best: Particles are attracted towards the swarm's global best position, guiding them towards the most promising region found so far by the entire search.
5. Iteration: The process of updating positions and velocities based on personal and global bests repeated iteratively. Over time, particles are expected to converge towards the optimal solution in the search space.

3.2.3. Feature Reduction (LDA)

What it is:

- LDA is a supervised learning technique used for dimensionality reduction and classification.
- It aims to find a linear transformation that projects data points from a high-dimensional space onto a lower-dimensional space while maximizing the separation between different classes.

How it works:

1. Dimensionality Reduction: LDA assumes that the data classes share an underlying linear structure. It identifies the most important directions (dimensions) in the high-dimensional space that separate the classes best.
2. Maximizing Class Separation: By projecting data points onto these key dimensions, LDA aims to maximize the distance between the means of different classes while minimizing the variance within each class. This creates a new, lower-dimensional space where the classes are well-separated.
3. Classification: Once the data is projected onto the reduced dimension space, a standard classification algorithm (e.g., logistic regression) can be applied to classify new, unseen data points.

Key characteristics:

- Supervised learning: Requires labeled data where each data point belongs to a known class.
- Linear transformation: LDA assumes a linear relationship between the features and the classes. It might not be suitable for problems with complex, non-linear relationships.
- Interpretability: LDA provides insights into the features that contribute most to the class separation. This can help understand the underlying factors driving the classification.
- Dimensionality reduction: LDA reduces the dimensionality of the data, potentially improving computational efficiency and reducing the risk of overfitting in classification models.

3.3. Used Machine Learning (Logistic Regression)

Logistic regression is a powerful statistical method widely used in machine learning for classification tasks. Unlike linear regression, which predicts continuous values, logistic regression deals with categorical outcomes, typically focusing on two categories (binary classification).

Imagine you want to classify an email as spam or not spam. Logistic regression comes into play here.

How it Works:

1. Sigmoid Function: At the heart of logistic regression lies the sigmoid function, also known as the logistic function. This S-shaped function transforms the linear combination of input features (data points) into a probability value between 0 and 1. A value closer to one indicates a higher likelihood of belonging to one class, while a value closer to zero suggests the opposite class.
2. Learning from Data: The logistic regression model learns from a training dataset containing labeled examples. Each data point has features (independent variables) and a corresponding category label (dependent variable). The model analyzes these relationships and adjusts its internal coefficients to best fit the data and predict probabilities for new, unseen instances.
3. Classification Threshold: To convert the predicted probabilities into concrete class labels, a classification threshold is often applied. A common choice is 0.5, where any probability above 0.5 is classified as class 1 and below 0.5 as class 2. This threshold can be adjusted based on the specific problem and desired trade-off between precision and recall (explained later).

Strengths of Logistic Regression:

- Interpretability: Unlike complex models like deep learning, logistic regression provides a more interpretable view of how features influence the outcome. You can understand which features have a stronger positive or negative impact on the predicted probability.
- Relatively Simple to Implement: Logistic regression has a well-defined mathematical foundation and is computationally efficient compared to some other machine learning models.

- Works Well with Limited Data: Logistic regression can be effective even with moderately sized datasets, making it suitable for scenarios where obtaining vast amounts of data might be challenging.

4. Experimental Results

4.1. Before PSO

Number of Features	12
Learning Rate	0.01
Loss	0.61
Accuracy of Train	79.1
Accuracy of Test	76.8
F1 Score	78.7
Precision	66.6
Recall	96.1

4.2. After PSO

Number of Features	9
Learning Rate	0.09
Loss	0.47
Accuracy of Train	84.5
Accuracy of Test	81.5
F1 Score	82.2
Precision	71.8
Recall	96.1

4.3. After LDA

Number of Features	1
Learning Rate	0.09
Loss	0.36
Accuracy of Train	86.3
Accuracy of Test	80.3
F1 Score	80.4
Precision	72.1
Recall	90.9