Minor in AI

Revision

Semi Supervised Learning

May 01, 2025

1 Introduction to Semi-Supervised Learning (SSL)

Semi-Supervised Learning (SSL) is a hybrid learning paradigm that leverages both labeled and unlabeled data during the training process. In most real-world scenarios, obtaining labeled data is costly, time-consuming, and often requires domain expertise. In contrast, unlabeled data is abundant and cheap. SSL aims to bridge the gap between supervised and unsupervised learning by utilizing the strengths of both.

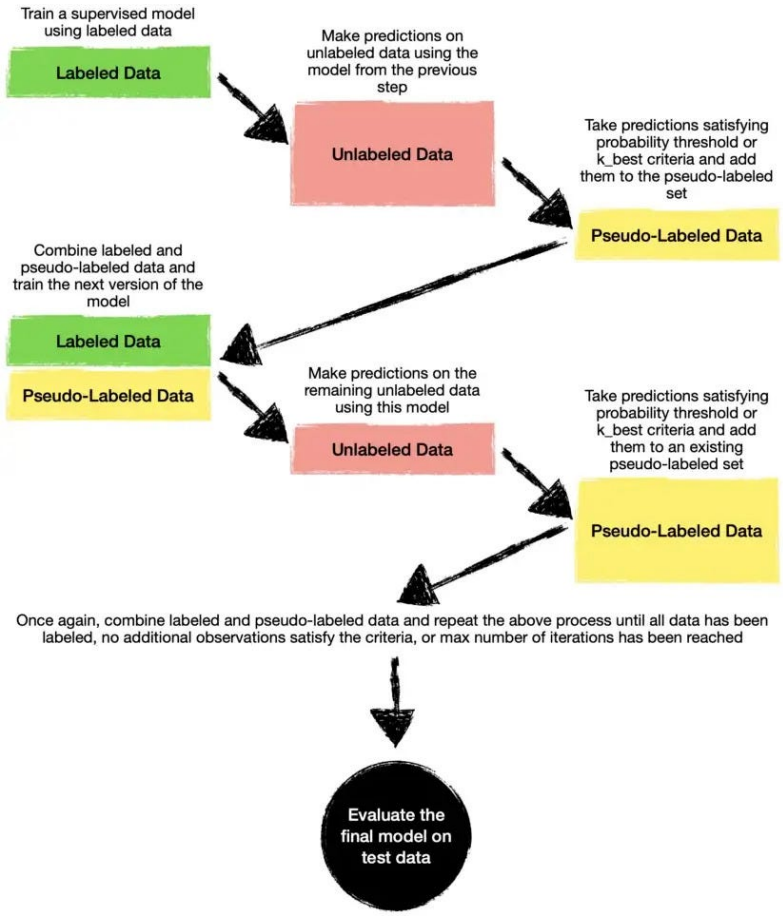

Figure 1: SSL

At its core, SSL seeks to learn a function $f(x)$ that maps inputs to outputs by using a small set of labeled examples along with a large pool of unlabeled data. This makes it especially useful in tasks where full supervision is infeasible but unlabeled data is plentiful.

2 Motivating Example: Spam Detection

We are given a dataset containing 50,000 emails. Out of these, only 500 emails are labeled as either spam or not spam.

- Training a supervised model solely on these 500 labeled samples is likely to lead to overfitting.
- The model may struggle to generalize well to new, unseen data.

- Semi-Supervised Learning (SSL) incorporates the structure of the 49,500 unlabeled emails.
- It uses patterns in the unlabeled data to improve learning.
- This results in more accurate and generalizable decision boundaries.

SSL thus showcases its effectiveness in practical settings with limited labeled data.

3 Why Semi-Supervised Learning matters?

SSL is not just a theoretical convenience; it provides tangible benefits in real-world applications:

- It dramatically reduces the cost and effort of manual data annotation.
- In domains like medical imaging, speech recognition, and legal document processing—where expert labeling is expensive—SSL helps build models using relatively few labeled instances.
- It improves the robustness and generalization ability of models by enabling them to learn from the structure of unlabeled data.

Moreover, SSL facilitates learning in situations where new classes or variations emerge over time, and annotating them all would be impractical.

4 Positioning SSL Among Other Learning Paradigms

To understand the value of SSL, it is important to contrast it with supervised and unsupervised learning.

4.1 Supervised Learning – Limitations

Supervised learning relies entirely on labeled data, delivering high accuracy and interpretable models when enough data is available. However, the downside is its dependency on large, clean, and annotated datasets, which are costly to produce.

4.2 Unsupervised Learning – Limitations

Unsupervised learning, on the other hand, works purely on unlabeled data to uncover hidden patterns or groupings. While it is flexible and data-efficient, it often lacks interpretability and alignment with downstream tasks.

4.3 How SSL solves this?

SSL lies between these extremes. It uses the small labeled set to guide the interpretation of patterns learned from the larger unlabeled set, combining the clarity of supervision with the scalability of unsupervised methods.

5 How SSL Works: Intuition and Example

Imagine a basket containing apples, bananas, and oranges, where only a few bananas and oranges are labeled. Initially, a classifier trained on these labels would misclassify apples or label them as "unknown." If a few apple examples are labeled and the model is retrained, it starts to generalize better. SSL works by identifying patterns and similarities in the unlabeled data and propagating label information accordingly.

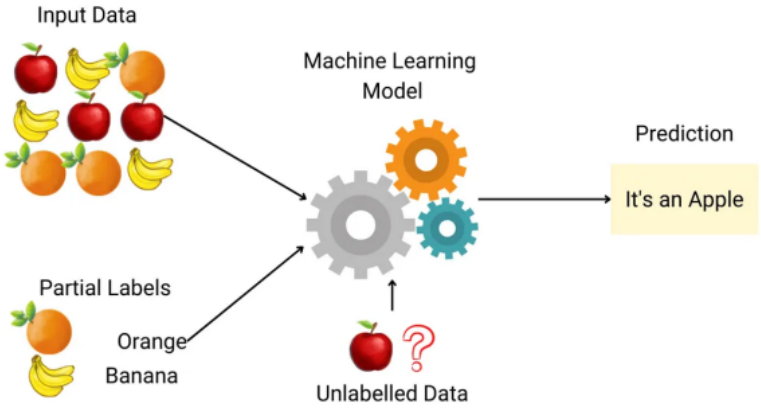

Figure 2: SSL Fruit Classification

This highlights the core strength of SSL: its ability to infer class structure and spread label information using the relationships within the data.

6 Workflow: Pseudo-Labeling

One common approach in SSL is pseudo-labeling, which operates in the following iterative process:

1. Train the model on the available labeled data.
2. Use the trained model to predict labels for the unlabeled data.
3. Combine the original labeled data with the pseudo-labeled data.
4. Retrain the model on this augmented dataset.
5. Repeat the process to further refine predictions.

This method gradually improves model performance by learning from its own high-confidence predictions, effectively expanding the labeled dataset while reducing dependency on manual annotations.

7 Core Assumptions

For SSL to succeed, it often relies on several key assumptions:

7.1 Smoothness Assumption

If two data points x and x' are close in the input space, then their corresponding outputs y and y' should also be similar. This assumption allows labels to be propagated through similar examples. For instance, slightly altered images of the digit "3" should still be classified as "3".

7.2 Cluster Assumption

Data tends to form discrete clusters, and instances within the same cluster likely share the same label. Therefore, decision boundaries should ideally lie in low-density regions between clusters. This helps in creating more robust classifiers.

7.3 Manifold Assumption

High-dimensional data often lies on a low-dimensional manifold. By learning this underlying structure, SSL can better generalize from few labeled examples. For example, images of different fruits under various lighting and angles still lie on continuous manifolds corresponding to their types.

8 Applications of SSL

SSL has demonstrated success across various domains:

- **Face Recognition:** Using Graph Convolutional Networks (GCNs) to exploit relationships between faces.
- **Handwriting Recognition:** Leveraging VAEs to adapt to individual writing styles.
- **Speech Recognition:** Improving recognition accuracy with unlabeled audio data; e.g., reducing word error rates.
- **Recommender Systems:** Enhancing personalization by learning from user behavior.
- **Document Classification:** Automatically tagging legal or scientific documents.
- **Image Classification & OCR:** Classifying massive image datasets with minimal labeling effort.

9 Challenges in Semi-Supervised Learning

While SSL offers numerous benefits, it also presents some challenges:

- **Quality of Unlabeled Data:** Noisy or irrelevant data can mislead the model.
- **Distribution Mismatch:** The labeled and unlabeled datasets may come from different distributions.
- **Model Complexity:** Advanced SSL methods can be computationally demanding.

- **Incorrect Pseudo-Labels:** Poor predictions can reinforce errors during training.
- **Privacy and Security:** Sensitive information in unlabeled data poses risks.

Effective SSL implementation requires careful handling of these issues, often by integrating domain knowledge and robust pre-processing steps.

Key Takeaways

1. Semi-Supervised Learning offers a powerful solution where labeled data is scarce and unlabeled data is abundant.
2. It combines the interpretability of supervised learning with the scalability of unsupervised learning.
3. Pseudo-labeling is a practical and commonly used approach.
4. The smoothness, cluster, and manifold assumptions are crucial for its success.
5. SSL is widely applied across computer vision, speech, and NLP domains, but care must be taken to avoid pitfalls such as poor pseudo-labeling or data mismatch.