Semi-Supervised Learning

Bridging the Gap Between Labeled and Unlabeled Data

Introduction to Semi-Supervised Learning

Imagine teaching a classroom where only a few students have taken practice quizzes (labeled data), while most have not (unlabeled). You use the quiz results to understand basic concepts, then leverage group discussions to help those without quiz experience. Semi-supervised learning works similarly: models use a small amount of labeled data combined with large amounts of unlabeled data to improve performance.

Why it matters:

- **Label scarcity**: Obtaining labeled data is expensive and time-consuming.
- **Data abundance**: Unlabeled data is often plentiful and cheap to collect.
- **Improved performance**: Leveraging unlabeled data often yields better models than using labeled data alone.

Semi-supervised learning sits between supervised (all labeled) and unsupervised learning (all unlabeled).

Key Approaches in Semi-Supervised Learning

1. Self-Training (Pseudo-Labeling)

Concept: Train a model on labeled data, predict labels for unlabeled data, then retrain using both real and high-confidence pseudo-labels.

Example Flow:

1. Train initial classifier on labeled dataset.
2. Use classifier to predict labels for unlabeled data.
3. Select predictions with confidence above threshold (e.g., 0.9) as pseudo-labels.
4. Retrain model on combined labeled + pseudo-labeled data.

Reflection:

- **Key:** Simple to implement, leverages model’s own predictions.
- **Pitfall:** Errors in pseudo-labels reinforce mistakes.
- **Application:** Text classification, image recognition.

2. Consistency Regularization

Concept: A good model should make consistent predictions for an example under small perturbations (e.g., noise, augmentations).

Example Flow:

1. Apply data augmentation (e.g., synonym replacement in text).
2. Enforce model outputs before and after augmentation to be similar, adding a consistency loss term:

$$L_{consistency} = \|f(x) - f(\text{augment}(x))\|^2$$

3. Combine with supervised loss on labeled data.

Reflection:

- **Key:** Exploits structure of data manifold.
- **Pitfall:** Augmentations must preserve semantics.
- **Application:** SSL for image and text tasks.

3. Graph-Based Methods

Concept: Represent data points as nodes in a graph, edges encode similarity. Propagate labels along edges to unlabeled nodes.

Example Flow:

1. Build graph where nodes are examples and edge weights are similarity measures.
2. Perform label propagation: each unlabeled node's label is the weighted average of its neighbors.

Reflection:

- **Key:** Captures global data relationships.
- **Pitfall:** Construction of graph can be computationally expensive.
- **Application:** Social network analysis, recommendation systems.

Short Example: Pseudo-Labeling in Text Classification

```
# Assume we have small labeled dataset (X_l, y_l) and large unlabeled X_u
from sklearn.linear_model import LogisticRegression

# Step 1: Train initial model
model = LogisticRegression()
model.fit(X_l, y_l)

# Step 2: Predict on unlabeled data
probs = model.predict_proba(X_u)
high_confidence = probs.max(axis=1) &gt; 0.9
pseudo_X = X_u[high_confidence]
pseudo_y = model.predict(X_u[high_confidence])
```

```
# Step 3: Retrain on combined data
X_combined = np.vstack([X_l, pseudo_X])
y_combined = np.concatenate([y_l, pseudo_y])
model.fit(X_combined, y_combined)
```

Output Discussion: The model now leverages high-confidence predictions, often boosting accuracy on test data by incorporating pseudo-labeled examples.

Reflection and Best Practices

Key Takeaways:

- **Semi-supervised learning** leverages both labeled and unlabeled data to improve performance.
- **Approaches** include self-training, consistency regularization, and graph-based propagation.
- **Balance** supervised and unsupervised losses carefully.

Common Pitfalls:

- **Error amplification** in pseudo-labeling.
- **Inappropriate perturbations** breaking semantics in consistency regularization.
- **Scalability issues** in graph construction.

Applications:

- **NLP**: Text classification, sentiment analysis with limited labeled data.
- **CV**: Image classification where labels are scarce.
- **Medical AI**: Utilizing vast unlabeled imaging data with few annotations.

Download the PDF above for a beginner-friendly, ready-to-publish chapter on Semi-Supervised Learning.