Semi-Supervised Learning Guide

## What is Semi-Supervised Learning?

\*\*Definition\*\*: Semi-Supervised Learning is a machine learning approach that uses a small amount of labeled data along with a large amount of unlabeled data during training.

\*\*Analogy\*\*: Like learning in a classroom where the teacher gives a few examples, and students figure out the rest by discussion and pattern recognition.

```python  
# Example: Self-training using labeled and unlabeled data  
from sklearn.semi\_supervised import SelfTrainingClassifier  
base\_model = LogisticRegression()  
model = SelfTrainingClassifier(base\_model)  
model.fit(X\_combined, y\_combined)  
```

## Why Semi-Supervised Learning?

Reduces the need for costly labeled data.

Improves accuracy by leveraging large unlabeled datasets.

Balances supervised and unsupervised approaches.

## Key Concepts

- \*\*Labeled Data\*\*: Data with known outputs.

- \*\*Unlabeled Data\*\*: Data with unknown outputs.

- \*\*Confidence Thresholds\*\*: Determines when to accept predictions as pseudo-labels.

- \*\*Pseudo-labeling\*\*: Assigning labels to unlabeled data based on model predictions.

## Common Techniques

- \*\*Self-training\*\*: Iteratively label and retrain using high-confidence predictions.

- \*\*Co-training\*\*: Train multiple models on different views and label each other's data.

- \*\*Graph-based Models\*\*: Use data similarities to propagate labels.

- \*\*Semi-Supervised SVM\*\*: Extends SVMs using both labeled and unlabeled data.

## Popular Algorithms & Tools

- SelfTrainingClassifier (scikit-learn)

- Label Spreading / Label Propagation

- Semi-supervised K-Means

- MixMatch, FixMatch (Deep Learning)

```python  
from sklearn.semi\_supervised import LabelPropagation  
model = LabelPropagation()  
model.fit(X\_combined, y\_combined)  
```

## Applications of Semi-Supervised Learning

- Text classification with few labeled documents.

- Image classification with partial labeling.

- Fraud detection with limited confirmed fraud labels.

- Speech and audio recognition tasks.

- Medical diagnosis with annotated patient subsets.

## Model Evaluation Strategies

- Use a labeled validation/test set for evaluation.

- Compare with fully supervised baseline.

- Track pseudo-label accuracy over iterations.

## Challenges

- Incorrect pseudo-labels can mislead the model.

- Hard to tune confidence thresholds.

- Requires careful monitoring of training performance.

## Best Practices

- Start with high-quality labeled data.

- Regularly validate pseudo-labeled samples.

- Use thresholding to avoid noisy labels.

- Combine with active learning when possible.

## Common Interview Questions

- What is the difference between semi-supervised and weak supervision?

- How does self-training work?

- What are the risks of using pseudo-labels?

- When would you choose semi-supervised over supervised learning?