SEMI-SUPERVISED LEARNING

Semi-Supervised Learning:

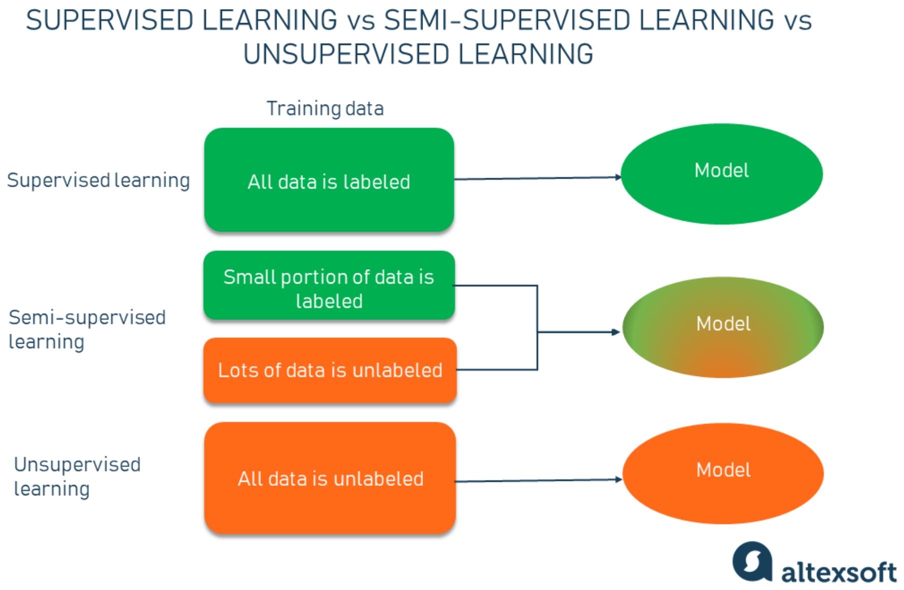
It is an approach to machine learning that combines a small amount of labeled data with a large amount of unlabeled data during training.

Semi-supervised learning falls between unsupervised learning (with no labeled training data) and supervised learning (with only labeled training data).

Unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy.

It refers to a learning problem (and algorithms designed for the learning problem) that involves a small portion of labeled examples and a large number of unlabeled examples from which a model must learn and make predictions on new examples.

- Unlike unsupervised learning, SSL works for a variety of problems from classification and regression to clustering and association.
- Unlike supervised learning, the method uses small amounts of labeled data and also large amounts of unlabeled data, which reduces expenses on manual annotation and cuts data preparation time.



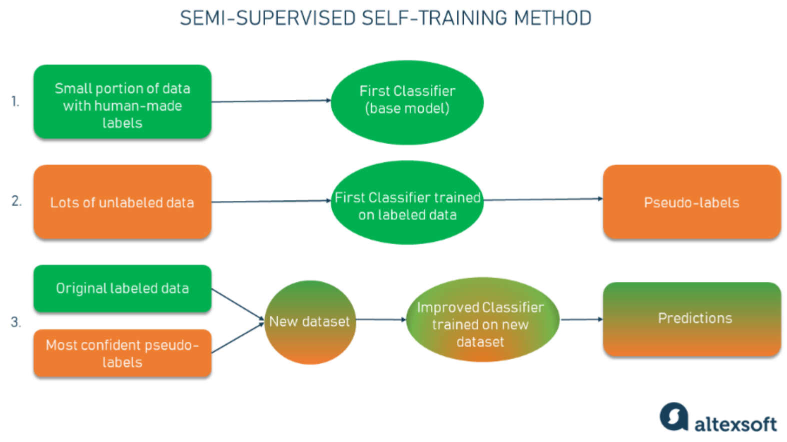
How semi-supervised learning works?

Imagine, you have collected a large set of unlabeled data that you want to train a model on. Manual labeling of all this information will probably cost you a fortune, besides taking months to complete the annotations. That's when the semi-supervised machine learning method comes to the rescue.

The working principle is quite simple. Instead of adding tags to the entire dataset, you go through and hand-label just a small part of the data and use it to train a model, which then is applied to the ocean of unlabeled data.

One of the simplest examples of semi-supervised learning, in general, is Self-Training.

Self-training is the procedure in which you can take any supervised method for classification or regression and modify it to work in a semi-supervised manner, taking advantage of labeled and unlabeled data.


SEMI-SUPERVISED SELF-TRAINING METHOD
1.
Small portion of data with human-made labels
First Classifier (base model)
2.
Lots of unlabeled data
First Classifier trained on labeled data
Pseudo-labels
3.
Original labeled data
Most confident pseudo-labels
New dataset
Improved Classifier trained on new dataset
Predictions
altexsoft

- You pick a small amount of labeled data, e.g., images showing cats and dogs with their respective tags, and you use this dataset to train a base model with the help of ordinary supervised methods.

- Then you apply the process known as pseudo-labeling — when you take the partially trained model and use it to make predictions for the rest of the database which is yet unlabeled. The labels generated thereafter are called pseudo as they are produced based on the originally labeled data that has limitations (say, there may be an uneven representation of classes in the set resulting in bias — more dogs than cats).

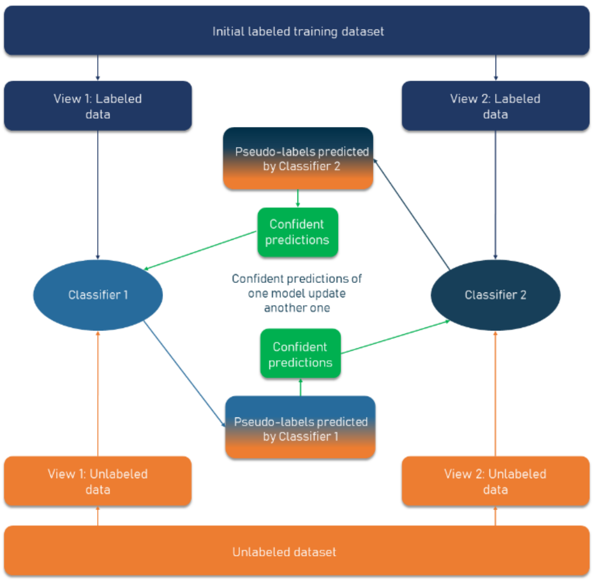
- From this point, you take the most confident predictions made with your model.

Co-training

Derived from the self-training approach and being its improved version, co-training is another semi-supervised learning technique used when only a small portion of labeled data is available. Unlike the typical process, co-training trains two individual classifiers based on two views of data.

The views are basically different sets of features that provide additional information about each instance, meaning they are independent given the class. Also, each view is sufficient — the class of sample data can be accurately predicted from each set of features alone.

SEMI-SUPERVISED CO-TRAINING METHOD



altexsoft

Here is how co-training works in simple terms:

- First, you train a separate classifier (model) for each view with the help of a small amount of labeled data.
- Then the bigger pool of unlabeled data is added to receive pseudo-labels.
- Classifiers co-train one another using pseudo-labels with the highest confidence level. If the first classifier confidently predicts the genuine label for a data sample while the other one makes a prediction error, then the data with the confident pseudo-labels assigned by the first classifier updates the second classifier and vice-versa.
- The final step involves the combining of the predictions from the two updated classifiers to get one classification result.

As with self-training, co-training goes through many iterations to construct an additional training labeled dataset from the vast amounts of unlabeled data.

Semi-supervised learning examples:

- Speech recognition
- Web content classification
- Text document classification