

7.4: Semi-supervised learning

There is no deep distinction between supervised and non-supervised methods, some of non-supervised (like SOM or PCA) could use training whereas some supervised (LDA, Random Forest, recursive partitioning) are useful directly as visualizations.

And there is a in-between semi-supervised learning. It takes into account both data features and data labeling (Figure 7.4.2).

One of the most important features of SSL is an ability to work with the very small training sample. Many really bright ideas are embedded in SSL, here we illustrate two of them. Self-learning is when classification is developed in multiple cycles. On each cycle, testing points which are most confident, are labeled and added to the training set:

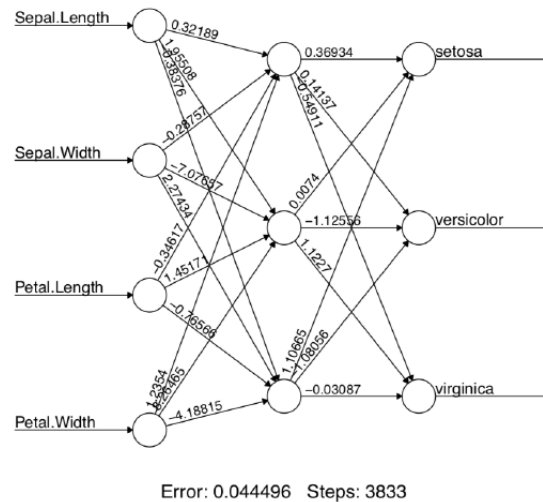


Figure 7.4.1 The neural network.

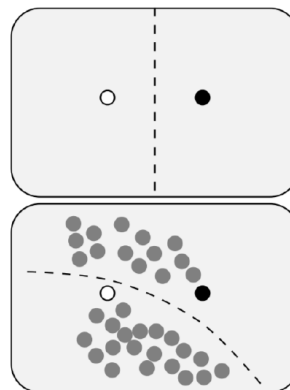


Figure 7.4.2 How semi-supervised learning can improve learning results. If only labeled data used, then the most logical split is between labeled points. However, if we look on the testing set, it become apparent that training points are parts of more complicated structures, and the actual split goes in the other direction.

As you see, with only 5 data points (approximately 3% of data vs. 33% of data in [iris.train](#)), semi-supervised self-learning (based on gradient boosting in this case) reached 73% of accuracy.

Another semi-supervised approach is based on graph theory and uses graph label propagation:

The idea of this algorithm is similar to what was shown on the illustration (Figure 7.4.2) above. Label propagation with 10 points outperforms Random Forest (see above) which used 30 points.

This page titled 7.4: Semi-supervised learning is shared under a [Public Domain](#) license and was authored, remixed, and/or curated by Alexey Shipunov via source content that was edited to the style and standards of the LibreTexts platform.