PH 240C Semi-supervised Learning

Jingshen Wang

October 13, 2021

1 Semi-supervised Learning

1.1 Overview

As the name suggests, semi-supervised learning is a machine learning method that stands between unsupervised and supervised learning, and can also be called as classification with labelled and unlabelled data (or partially labelled data). Semi-supervised learning has tremendous practical value. In many tasks, there is a paucity of labelled data. The labels Y may be difficult to obtain because they require human annotators, special devices, or expensive and slow experiments. For example,

- In speech recognition, an instance X is a speech utterance, and the label Y is the corresponding transcript. For example, here are some detailed phonetic transcript of words as they are spoken:

  $$\text{film} \Rightarrow \text{f ih\_n uh\_gl\_n m.}$$

  Accurate transcription by human expert annotators can be extremely time consuming: it took as long as 400 hours to transcribe 1 hour of speech at the phonetic level for a recordings of randomly paired participants discussing various topics such as social, economic, political, and environmental issues Godfrey et al. (1992).

- In spam filtering, an instance X is an email, and the label Y is the user's judgment (spam or ham). In this situation, the bottleneck is an average user's patience to label a large number of emails.

- In healthcare, given the vast accessibility of electronic health record data, labeled disease status are rather sparse. How to use the labelled disease status to infer the unlabelled patients is of vital interest (Ford et al., 2016).

While labelled data pair is difficult to obtain in these examples, unlabelled data X are available in large quantity and easy to collect: speech utterance can be recorded from radio broadcasts. Semi-supervised learning is attractive because it can potentially utilize both labeled and un labeled data to achieve better performance than supervised learning. From a different perspective, semi-supervised learning may achieve the same level of performance as supervised learning, but with fewer labelled instances. This reduces the annotation effort, which leads to reduced cost.

Semi-supervised learning also provides a computational model of how humans learn from labelled and unlabeled data. Consider the task of concept learning in children, which is similar to classification: an instance x is an object (e.g., an animal), and the label y is the corresponding concept (e.g., dog). Young

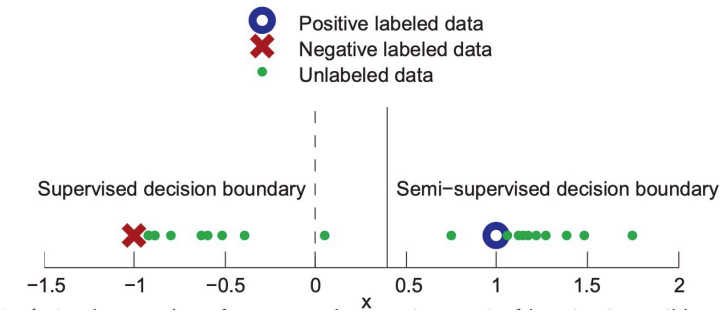

Figure 1: A simple example to demonstrate how semi-supervised learning is possible.

children receive labeled data from teachers (e.g., Daddy points to a brown animal and says "dog!"). But more often they observe various animals by themselves without receiving explicit labels. It seems self-evident that children are able to combine labeled and unlabeled data to facilitate concept learning. The study of semi-supervised learning is therefore an opportunity to bridge machine learning and human learning.

1.2 Why semi-supervised learning can be helpful?

At first glance, it might seem paradoxical that one can learn anything about a classifier from unlabeled data. After all, a classifier f is about the mapping from the instance X to the label Y, yet unlabeled data do not provide any example of such a mapping. The answer lies in the assumptions one makes about the *link between the distribution of unlabeled data and the target label*.

Figure 1 shows a simple example of semi-supervised learning. Let each instance be represented by a one-dimensional attribute $X \in \mathbb{R}$. They are two classes $Y \in \{-1, 1\}$. Consider the following two scenarios:

1. In supervised learning, we are only given two labeled training instance $(X_1, Y_1) = (-1, -1)$ and $(X_2, Y_2) = (1, 1)$. The best decision boundary is at $x = 0$.

2. In addition, we are also given a large number of unlabeled instances, shown as green dots in the figure. The correct class labels for these unlabeled examples are unknown. However, we observe that they form two groups. If we **assume** that subjects in each class form a coherent group (for example the conditional distribution $\mathbb{P}(X|Y = 1)$ has finite variance), this unlabeled data gives us more information. Specifically, it seems that the two labeled instances are not the most prototypical examples for the two considered classes. The semi-supervised estimate of the decision boundary should be between the two groups instead at $x \approx 0.4$.

If our assumption is true, then using both labeled and unlabeled data gives us a more reliable estimate of the decision boundary. Intuitively, the distribution of unlabeled data helps to find a good division of the attribute space, and then the few labeled data then provide actual labels. Of course, this is merely one assumption considered in an ocean of literature on semi-supervised learning.

1.3 Inductive and transductive semi-supervised learning

We now move forward with introducing two slightly different semi-supervised leaning setups, namely inductive and transductive semi-supervised learning. Recalled that in supervised learning, the training sample is

fully labeled, so one is always interested in the performance on future test data. In semi-supervised classification, however, the training sample contains some unlabeled data. Therefore, there are two distinct goals:

1. Inductive semi-supervised learning. Given a training sample with labeled part $\{(X_i, Y_i)\}_{i=1}^n$, and unlabeled part $\{X_j\}_{j=n+1}^N$, inductive semi-supervised learning learns a classifier that predicts well on **future** data, beyond $\{X_j\}_{j=n+1}^N$.

2. Transductive learning. Given a training sample with labeled part $\{(X_i, Y_i)\}_{i=1}^n$, and unlabeled part $\{X_j\}_{j=n+1}^N$, transductive semi-supervised learning learns a classifier that predicts well on **unlabeled** data $\{X_j\}_{j=n+1}^N$.

There is an interesting analogy: inductive semi-supervised learning is like an in-class exam, where the questions are not known in advance, and a student needs to prepare for all possible questions; in contrast, transductive learning is like a take-home exam, where the student knows the exam questions and needs not prepare beyond those.

1.4 Self-training models

Self-training is characterized by the fact that the learning process uses its own predictions to teach itself. For this reason, it is also called self-teaching or bootstrapping (not to be confused with the statistical procedure with the same name). Self-training can be either inductive or transductive, depending on the problem:

1. Input labeled data $\{(X_i, Y_i)\}_{i=1}^n$ and unlabeled data $\{X_j\}_{j=n+1}^N$
2. Define the initial labeled set as $L = \{1, \ldots, n\}$ and the initial unlabeled set as U
3. Repeat the following steps that iteratively update the labeled set and the unlabeled set:
   (a) Train f from the data points in L using supervised learning
   (b) Apply f to the unlabeled instances in U
   (c) Remove a subset S that is confidently labeled from U
   (d) Add $\{(X_i, f(X_i))|X_i \in S\}$ to L

The main idea is to first train a classifier on labeled data. The function f is then used to predict the labels for the unlabeled data. A subset S of the unlabeled data, together with their predicted labels, are then selected to augment the labeled data. Typically, S consists of a few unlabeled instances with the most confidence f predictions. The function f is then re-trained on the now larger set of labeled data, and the procedure repeats. This procedure looks quite intuitive, but it also relies on the **assumption** that the labeled dataset prediction results tend to be correct, at least for the high confidence ones. This is likely to be the case when the classes from well-separated clusters.

The major advantages of self-training are its simplicity, and the choice of the classifier for f in Step 3(b) is left completely open. For example, the classifier can be a simple kNN algorithm, or a very complicated classifier. The self-training procedure "wraps" around the classifier without changing its inner workings. The disadvantage is also clear: the early mistake an algorithm makes can be reinforced by keep generating incorrectly labeled data. Re-training with wrongly classified data will lead to an even worse f in the next iteration.

A concrete example of self-training, we now introduce an algorithm with a simple example:

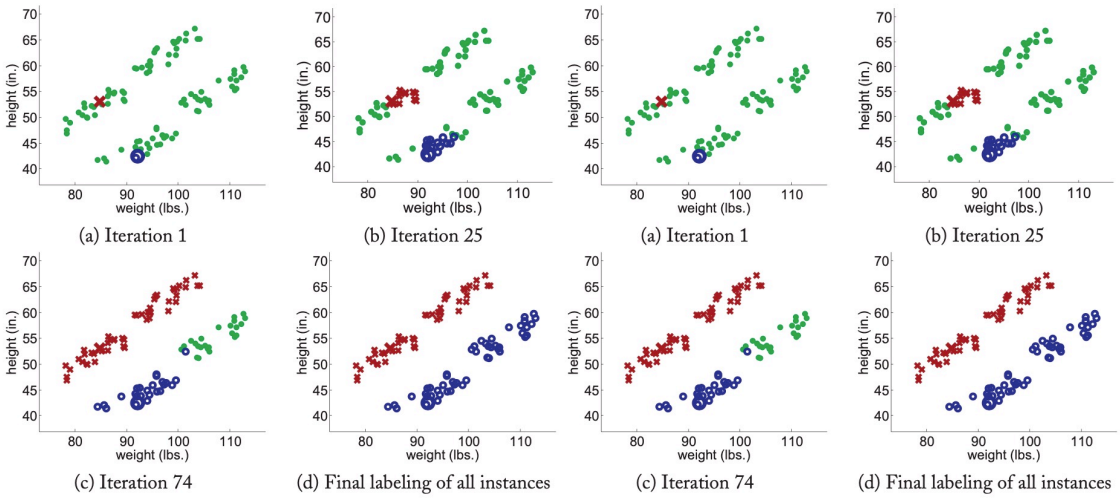

Figure 2: Semi-supervised learning algorithm with a 1-nearest-neighbour classifier.

1. Input labeled data $\{(X_i, Y_i)\}_{i=1}^n$, unlabeled data $\{X_j\}_{j=n+1}^N$ and a distance function $d(\cdot)$
2. Define the initial labeled set as $L = \{1, \ldots, n\}$ and the initial unlabeled set as U
3. Repeat the following steps until U is empty:
   (a) Select $X = \arg\min_{X \in U} \min_{X' \in L} d(X, X')$
   (b) Set $f(X)$ to be the labels of X's nearest instance in L
   (c) Remove X from U, and add $\{X, f(X)\}$ to L.

See a simple illustrative example in Figure 2.

1.5 Semi-supervised learning in healthcare

Suppose we have patient cohorts from two large hospitals. Our goal is to predict the status for bipolar disorder, a heritable mental disorder characterized by mood swings between mania and depression. Similar attributes are collected at both sites including age, gender, race, count of the diagnostic code for bipolar disorder and count of the major depressive disorder or depression. At one site, clinical investigators have manually labelled the binary disease status for n patients, and the other site, we have access to $N - n$ additional patient attributes. In this example, we focus on transductive learning as its goal is more aligned with healthcare at large.

To have a concrete statistical model, suppose the labeled data are denoted as $\{(X_i, Y_i)\}_{i=1}^n$, and the unlabeled data are denoted as $\{X_j\}_{j=n+1}^N$. Here $X_i \in \mathbb{R}^p$ captures the attributes and $Y_i \in \{-1, 1\}$ is the corresponding label. We use S_i to denote whether a data point i is labeled:

$$S_i = \begin{cases} 1 & i\text{th data point is labeled} \\ 0 & i\text{th data point is unlabeled,} \end{cases} \qquad i = 1, \ldots, N.$$

Again, semi-supervised learning only improves the prediction accuracy under certain assumption. From our motivating example, we assume that given a set of attributes $X_i = x$ either from labeled or unlabeled

dataset, the probability of $Y_i = 1$ does not change, i.e.,

$$\mathbb{P}\big(Y_i = 1|X_i, S_i = 1\big) = \mathbb{P}\big(Y_i = 1|X_i, S_i = 0\big), \quad i = 1, \ldots, N.$$

In other words, we assume that there is an underlying true status of patients disease profile regardless of whether his/her label is missing (when will this assumption be violated?). Nevertheless, given the patient information is collected across different locations, their baseline profile may differ accrediting to whether they are labeled or not. Thus, it is reasonable to expect that there is a covariate shift between two sites:

$$\mathbb{P}(X_i = x|S_i = 1) \neq \mathbb{P}(X_i = x|S_i = 0), \quad i = 1, \ldots, N.$$

Our goal here is to accurate predict the disease status for unlabeled patients, i.e., we aim to find a classifier $f_0(\cdot)$ that minimizes the empirical risk conditional on the data are unlabelled:

$$f_0 = \arg\min_{f \in \mathcal{F}} \mathbb{E}\big[l(Y_j, f(X_j))|S_j = 0\big].$$

We can further start from the class of linear classifiers, then the problem can be simplified into

$$\mathrm{w}_0 = \arg\min_{\mathrm{W} \in \mathbb{R}^p} \mathbb{E}\big[l_{\mathrm{W}}(Y_j, h_{\mathrm{W}}(X_j))|S_j = 0\big].$$

Note that different from supervised learning approaches used in previous lectures, Y_j is unobserved and the expectation function cannot be replaced by its sample analogue. Fortunately, under covariate shift, we have

$$\begin{aligned}\text{unobserevd} \to \mathbb{E}\big[l_{\mathrm{W}}(Y_j, h_{\mathrm{W}}(X_j))|S_j = 0\big] &= \mathbb{E}\Big[\mathbb{E}\big[l_{\mathrm{W}}(Y_j, h_{\mathrm{W}}(X_j))|X_j, S_j = 0\big]\big|S_j = 0\Big] \\ &= \mathbb{E}\Big[\mathbb{E}\big[l_{\mathrm{W}}(Y_j, h_{\mathrm{W}}(X_j))|X_j, S_j = 1\big]\big|S_j = 0\Big] \leftarrow \text{observed}\end{aligned}$$

Hence, as long as we can approximate the covariate shift between the labeled and unlabeled data, we may try to find w_0 by minimizing the following objective function

$$\frac{1}{n}\sum_{i=1}^{n} l_{\mathrm{W}}(Y_i, h_{\mathrm{W}}(X_i)) \cdot \frac{\mathbb{P}(X_i|S_i = 0)}{\mathbb{P}(X_i|S_i = 1)}.$$

Now the question remains is how to approximate the covariates shift especially when we have access to many attributes. What ideas do you have?

Multiple researchers have informally noted that semi-supervised learning does not always help. Little is written about it, except a few papers like Cozman et al. (2003); ELWORTHY (1994). This is presumably due to "publication bias," that negative results tend not to be published. A deeper understanding of when semi-supervised learning works merits further study.

References

Fabio Gagliardi Cozman, Ira Cohen, Marcelo Cesar Cirelo, et al. Semi-supervised learning of mixture models. In *ICML*, volume 4, page 24, 2003.

David ELWORTHY. Does baum-welch re-estimation help taggers? In *Proceedings of the Fourth Conference on Applied Natural Language Processing, 1994*, 1994.

Elizabeth Ford, John A Carroll, Helen E Smith, Donia Scott, and Jackie A Cassell. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, 23(5):1007–1015, 2016.

John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society, 1992.