4-1-2021

Why Semi-Supervised Learning Makes Sense: A Pedagogical Note

Olga Kosheleva
The University of Texas at El Paso, olgak@utep.edu

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Comments:
Technical Report: UTEP-CS-21-43

Why Semi-Supervised Learning Makes Sense: A Pedagogical Note

Olga Kosheleva and Vladik Kreinovich

Abstract The main idea behind semi-supervised learning is that when we do not enough human-generated labels, we train a machine learning system based on what we have, and we add the resulting labels (called *pseudo-labels*) to the training sample. Interesting, this idea works well, but why is somewhat a mystery: we did not add any new information so why is this working? There exist explanations for this empirical phenomenon, but most these explanations are based on complicated math. In this paper, we provide a simple intuitive explanation.

1 Formulation of the Problem

Usual (supervised) machine learning. In the usual machine learning, we have several (K) objects. Each object $k = 1, \ldots, K$ is characterized by parameters $x^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})$. For these objects, we also know the values $y^{(k)}$ of some characteristics y. In the simplest case, the values y come from a small finite set Y – e.g., we know which object is a cat and which is a dog. In this discrete cases, the corresponding values y are called *labels*.

Based on this information, we want to come up with an algorithm that, given a new object $x = (x_1, \ldots, x_n)$, will predict the value y corresponding to this object. This is exactly what many efficient machine learning algorithms do, including deep learning algorithms [1, 2].

Olga Kosheleva
Department of Teacher Education, University of Texas at El Paso, 500 W. University
El Paso, TX 79968, USA
e-mail: olgak@utep.edu

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso, 500 W. University
El Paso, TX 79968, USA
e-mail: vladik@utep.edu

Many such algorithms provide, for each object x, not only the corresponding label, they also provide a degree to which the system is confident in this label.

Limitations of supervised machine learning. The more patterns $(x^{(k)}, y^{(k)})$ we have, the more information we have about the desired dependence, the more accurate is the resulting algorithm $y = f(x_1, \ldots, x_n)$.

When the values y come from experiments, we can get thousands and even millions of patterns, and thus reach a high accuracy of the resulting algorithm. The problem is that in many cases, the values $y^{(k)}$ have to be provided by humans, and so it is not realistic to expect that many patterns. This is important, e.g., when we teach a computer to analyze videos or visual schemes. Because the number of patterns is limited, the resulting accuracy is not so good.

Idea of semi-supervised learning. Since we do have that many labeled patterns, when we train a machine learning algorithm on whatever labeled patterns we have, we get a not very accurate description. For some objects x, the system provides the estimate y with higher degrees of confidence, for some, lower.

The idea of semi-supervised learning is that, after setting some threshold p_0, we assume that all the labels assigned with confidence p_0 or higher are correct, and repeat the training with the correspondingly enlarged set of patterns. To distinguish the new labels from the ones provided by human experts, the newly added labels are called *pseudo-labels*.

We can stop here or, alternatively, after this second training, we can again select labels assigned with confidence greater than or equal to p_0, add them, etc.

Interestingly, this idea works very well; see, e.g., [2].

But why does this idea work? At first glance, the fact that this idea works seems like magic. We did not add any new expert information, we did not add any new knowledge about the classified objects, so why does this improve the accuracy?

What we do in this paper. There are rather complicated mathematical explanations of why this idea works. In this paper, we provide a simpler more intuitive explanation.

2 An Explanation

Simplified setting. Our explanation is based on a simple but natural idea: if we have two classes, e.g., cats and dogs, and a new object is closer to some known cats than to all known dogs, then it is natural to classify this object as a cat.

Comment. Of course, this is a very simplified version of machine learning, but it is definitely one of the main ideas behind the discrete case of machine learning.

From this viewpoint, when are we more confident. From this viewpoint, the larger the ratio between the distance between this object and the nearest dog and its distance to the nearest cat, the more confident we are that the new object is a cat.

So what new information do we add? Suppose that originally, we have one sample cat and one sample dog. Suppose, for simplicity, that all cats are largely alike, while dogs differ a lot – by size, etc. Then, when we perform a crude first approximation, most cats will be correctly classified as cats, but only dogs close in size to the original dog will be confidently classified as dogs. Let us call them *1st generation*.

When we add all cats and all 1st generation dogs as new patterns and apply training again, we will get new dogs confidently classified as dogs – namely, those that are close to dogs of 1st generation. Let us call them *2nd generation*. We then add 2nd generation dogs, etc.

At each point, we add dogs which are somewhat close to dogs from the previous generation. Any two dogs can be connected by such a sequence of not large transitions. So, at the end, we get a good classification of all the dogs.

In a nutshell: to the previous sample patterns, we added information about closeness: which objects are close to each other. Objects close to objects of known type are probable to belong to the same type.

Illustrative example. Let us have a simple 1-D example illustrating this explanation. Suppose that each object is characterize by only one parameter x_1, and that we have two groups of objects:

- the first group (the left one) is spread a lot, while
- in the second group (to the right) all the objects are so close together that they are practically indistinguishable.

Each object will be denoted by a dashed vertical segment:

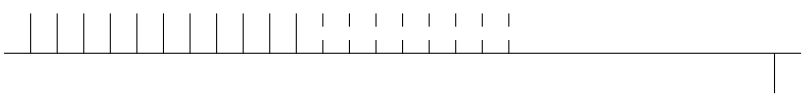
- objects from the first group correspond to segments pointed up,
- objects from the second group are marked by segments pointing down.

Objects for which we originally know the labels are marked by an interrupted segments.

Here is our original status.

Suppose that we confidently identify an object as belonging to a class if its distance to the nearest object from this class is at least twice smaller than its distance to the nearest object of another class.

In this case, after the first application of machine learning, some objects of the first class will be correctly classified as such:



However, many other objects from the first class remain unclassified. But now, we can perform the second iteration, with all newly classified objects as pseudo-labels. Now, more objects from the first class will be correctly classified:

A few objects from the first class are still unclassified, but if we apply the same procedure for the third time, all objects will be correctly classified:

Comment. If we simply classified objects based on their closeness to the original labels, we would get several objects of the first class misclassified as belonging to the second class (with one object – as exactly the same distance from both original labels – left uncertain):

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

References

1. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.
2. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Leaning*, MIT Press, Cambridge, Massachusetts, 2016.