Title: Weak supervision

URL: https://en.wikipedia.org/wiki/Weak_supervision

PageID: 60968880

Categories: Category:Machine learning

Weak supervision (also known as semi-supervised learning) is a paradigm in machine learning , the relevance and notability of which increased with the advent of large language models due to large amount of data required to train them. It is characterized by using a combination of a small amount of human- labeled data (exclusively used in more expensive and time-consuming supervised learning paradigm), followed by a large amount of unlabeled data (used exclusively in unsupervised learning paradigm). In other words, the desired output values are provided only for a subset of the training data. The remaining data is unlabeled or imprecisely labeled. Intuitively, it can be seen as an exam and labeled data as sample problems that the teacher solves for the class as an aid in solving another set of problems. In the transductive setting, these unsolved problems act as exam questions. In the inductive setting, they become practice problems of the sort that will make up the exam.

Problem

Supervised learning

Unsupervised learning

Semi-supervised learning

Self-supervised learning

Reinforcement learning

Meta-learning

Online learning

Batch learning

Curriculum learning

Rule-based learning

Neuro-symbolic AI

Neuromorphic engineering

Quantum machine learning

Classification

Generative modeling

Regression

Clustering

Dimensionality reduction

Density estimation

Anomaly detection

Data cleaning

AutoML

Association rules

Semantic analysis
Structured prediction
Feature engineering
Feature learning
Learning to rank
Grammar induction
Ontology learning
Multimodal learning
Apprenticeship learning
Decision trees
Ensembles Bagging Boosting Random forest
Bagging
Boosting
Random forest
k -NN
Linear regression
Naive Bayes
Artificial neural networks
Logistic regression
Perceptron
Relevance vector machine (RVM)
Support vector machine (SVM)
BIRCH
CURE
Hierarchical
k -means
Fuzzy
Expectation–maximization (EM)
DBSCAN
OPTICS
Mean shift
Factor analysis
CCA
ICA
LDA
NMF
PCA
PGD

t-SNE

SDL

Graphical models Bayes net Conditional random field Hidden Markov

Bayes net

Conditional random field

Hidden Markov

RANSAC

k -NN

Local outlier factor

Isolation forest

Autoencoder

Deep learning

Feedforward neural network

Recurrent neural network LSTM GRU ESN reservoir computing

LSTM

GRU

ESN

reservoir computing

Boltzmann machine Restricted

Restricted

GAN

Diffusion model

SOM

Convolutional neural network U-Net LeNet AlexNet DeepDream

U-Net

LeNet

AlexNet

DeepDream

Neural field Neural radiance field Physics-informed neural networks

Neural radiance field

Physics-informed neural networks

Transformer Vision

Vision

Mamba

Spiking neural network

Memtransistor

Electrochemical RAM (ECRAM)

Q-learning

Policy gradient

SARSA

Temporal difference (TD)

Multi-agent Self-play

Self-play

Active learning

Crowdsourcing

Human-in-the-loop

Mechanistic interpretability

RLHF

Coefficient of determination

Confusion matrix

Learning curve

ROC curve

Kernel machines

Bias–variance tradeoff

Computational learning theory

Empirical risk minimization

Occam learning

PAC learning

Statistical learning

VC theory

Topological deep learning

AAAI

ECML PKDD

NeurIPS

ICML

ICLR

IJCAI

ML

JMLR

Glossary of artificial intelligence

List of datasets for machine-learning research List of datasets in computer vision and image processing

List of datasets in computer vision and image processing

Outline of machine learning

v

t

e

The acquisition of labeled data for a learning problem often requires a skilled human agent (e.g. to transcribe an audio segment) or a physical experiment (e.g. determining the 3D structure of a protein or determining whether there is oil at a particular location). The cost associated with the labeling process thus may render large, fully labeled training sets infeasible, whereas acquisition of unlabeled data is relatively inexpensive. In such situations, semi-supervised learning can be of great practical value. Semi-supervised learning is also of theoretical interest in machine learning and as a model for human learning.

Technique

More formally, semi-supervised learning assumes a set of l independently identically distributed examples $x_1, \dots, x_l \in X$ with corresponding labels $y_1, \dots, y_l \in Y$ and u unlabeled examples $x_{l+1}, \dots, x_{l+u} \in X$ are processed. Semi-supervised learning combines this information to surpass the classification performance that can be obtained either by discarding the unlabeled data and doing supervised learning or by discarding the labels and doing unsupervised learning.

Semi-supervised learning may refer to either transductive learning or inductive learning . [1] The goal of transductive learning is to infer the correct labels for the given unlabeled data $x_{l+1}, \dots, x_{l+u}$ only. The goal of inductive learning is to infer the correct mapping from X to Y .

It is unnecessary (and, according to Vapnik's principle , imprudent) to perform transductive learning by way of inferring a classification rule over the entire input space; however, in practice, algorithms formally designed for transduction or induction are often used interchangeably.

Assumptions

In order to make any use of unlabeled data, some relationship to the underlying distribution of data must exist. Semi-supervised learning algorithms make use of at least one of the following assumptions: [2]

Continuity / smoothness assumption

Points that are close to each other are more likely to share a label. This is also generally assumed in supervised learning and yields a preference for geometrically simple decision boundaries . In the case of semi-supervised learning, the smoothness assumption additionally yields a preference for decision boundaries in low-density regions, so few points are close to each other but in different classes. [3]

Cluster assumption

The data tend to form discrete clusters, and points in the same cluster are more likely to share a label (although data that shares a label may spread across multiple clusters). This is a special case of the smoothness assumption and gives rise to feature learning with clustering algorithms.

Manifold assumption

The data lie approximately on a manifold of much lower dimension than the input space. In this case learning the manifold using both the labeled and unlabeled data can avoid the curse of dimensionality . Then learning can proceed using distances and densities defined on the manifold.

The manifold assumption is practical when high-dimensional data are generated by some process that may be hard to model directly, but which has only a few degrees of freedom. For instance, human voice is controlled by a few vocal folds, [4] and images of various facial expressions are controlled by a few muscles. In these cases, it is better to consider distances and smoothness in the natural space of the generating problem, rather than in the space of all possible acoustic waves or images, respectively.

History

The heuristic approach of self-training (also known as self-learning or self-labeling) is historically the oldest approach to semi-supervised learning, [2] with examples of applications starting in the 1960s. [5]

The transductive learning framework was formally introduced by Vladimir Vapnik in the 1970s. [6] Interest in inductive learning using generative models also began in the 1970s. A probably approximately correct learning bound for semi-supervised learning of a Gaussian mixture was demonstrated by Ratsaby and Venkatesh in 1995. [7]

Methods

Generative models

Generative approaches to statistical learning first seek to estimate p (x | y) {\displaystyle p(x|y)} , [disputed – discuss] the distribution of data points belonging to each class. The probability p (y | x) {\displaystyle p(y|x)} that a given point x {\displaystyle x} has label y {\displaystyle y} is then proportional to p (x | y) p (y) {\displaystyle p(x|y)p(y)} by Bayes' rule . Semi-supervised learning with generative models can be viewed either as an extension of supervised learning (classification plus information about p (x) {\displaystyle p(x)}) or as an extension of unsupervised learning (clustering plus some labels).

Generative models assume that the distributions take some particular form p (x | y , θ) {\displaystyle p(x|y,\theta)} parameterized by the vector θ {\displaystyle \theta } . If these assumptions are incorrect, the unlabeled data may actually decrease the accuracy of the solution relative to what would have been obtained from labeled data alone. [8] However, if the assumptions are correct, then the unlabeled data necessarily improves performance. [7]

The unlabeled data are distributed according to a mixture of individual-class distributions. In order to learn the mixture distribution from the unlabeled data, it must be identifiable, that is, different parameters must yield different summed distributions. Gaussian mixture distributions are identifiable and commonly used for generative models.

The parameterized joint distribution can be written as p (x , y | θ) = p (y | θ) p (x | y , θ) {\displaystyle p(x,y|\theta)=p(y|\theta)p(x|y,\theta)} by using the chain rule . Each parameter vector θ {\displaystyle \theta } is associated with a decision function f θ (x) = argmax y p (y | x , θ) {\displaystyle f_{\theta }(x)={\underset {y}{\operatorname {argmax} }}\ p(y|x,\theta)} .

The parameter is then chosen based on fit to both the labeled and unlabeled data, weighted by λ {\displaystyle \lambda } :

Low-density separation

Another major class of methods attempts to place boundaries in regions with few data points (labeled or unlabeled). One of the most commonly used algorithms is the transductive support vector machine , or TSVM (which, despite its name, may be used for inductive learning as well). Whereas support vector machines for supervised learning seek a decision boundary with maximal margin over the labeled data, the goal of TSVM is a labeling of the unlabeled data such that the decision boundary has maximal margin over all of the data. In addition to the standard hinge loss (1 − y f (x)) + {\displaystyle (1-yf(x))_{+}} for labeled data, a loss function (1 − | f (x) |) + {\displaystyle (1-|f(x)|)_{+}} is introduced over the unlabeled data by letting y = sign ⁡ f (x) {\displaystyle y=\operatorname {sign} {f(x)}} . TSVM then selects f ∗ (x) = h ∗ (x) + b {\displaystyle f^{*}(x)=h^{*}(x)+b} from a reproducing kernel Hilbert space H {\displaystyle {\mathcal {H}}} by minimizing the regularized empirical risk :

An exact solution is intractable due to the non- convex term (1 − | f (x) |) + {\displaystyle (1-|f(x)|)_{+}} , so research focuses on useful approximations. [9]

Other approaches that implement low-density separation include Gaussian process models, information regularization, and entropy minimization (of which TSVM is a special case).

Laplacian regularization

Laplacian regularization has been historically approached through graph-Laplacian.

Graph-based methods for semi-supervised learning use a graph representation of the data, with a node for each labeled and unlabeled example. The graph may be constructed using domain knowledge or similarity of examples; two common methods are to connect each data point to its k {\displaystyle k} nearest neighbors or to examples within some distance ■ {\displaystyle \epsilon } . The weight W i j {\displaystyle W_{ij}} of an edge between x i {\displaystyle x_{i}} and x j {\displaystyle x_{j}} is then set to e − ■ x i − x j ■ 2 / ■ 2 {\displaystyle e^{-\|x_{i}-x_{j}\|^{2}/\epsilon ^{2}}} .

Within the framework of manifold regularization , [10] [11] the graph serves as a proxy for the manifold. A term is added to the standard Tikhonov regularization problem to enforce smoothness of the solution relative to the manifold (in the intrinsic space of the problem) as well as relative to the ambient input space. The minimization problem becomes

where H {\displaystyle {\mathcal {H}}} is a reproducing kernel Hilbert space and M {\displaystyle {\mathcal {M}}} is the manifold on which the data lie. The regularization parameters λ A {\displaystyle \lambda _{A}} and λ I {\displaystyle \lambda _{I}} control smoothness in the ambient and intrinsic spaces respectively. The graph is used to approximate the intrinsic regularization term. Defining the graph Laplacian L = D − W {\displaystyle L=D-W} where D i i = ∑ j = 1 l + u W i j {\displaystyle D_{ii}=\sum _{j=1}^{l+u}W_{ij}} and f {\displaystyle \mathbf {f} } is the vector [f (x 1) … f (x l + u)] {\displaystyle [f(x_{1})\dots f(x_{l+u})]} , we have

The graph-based approach to Laplacian regularization is to put in relation with finite difference method . [clarification needed] [citation needed]

The Laplacian can also be used to extend the supervised learning algorithms: regularized least squares and support vector machines (SVM) to semi-supervised versions Laplacian regularized least squares and Laplacian SVM.

Heuristic approaches

Some methods for semi-supervised learning are not intrinsically geared to learning from both unlabeled and labeled data, but instead make use of unlabeled data within a supervised learning framework. For instance, the labeled and unlabeled examples x 1 , … , x l + u {\displaystyle x_{1},\dots ,x_{l+u}} may inform a choice of representation, distance metric , or kernel for the data in an unsupervised first step. Then supervised learning proceeds from only the labeled examples. In this vein, some methods learn a low-dimensional representation using the supervised data and then apply either low-density separation or graph-based methods to the learned representation. [12] [13] Iteratively refining the representation and then performing semi-supervised learning on said representation may further improve performance.

Self-training is a wrapper method for semi-supervised learning. [14] First a supervised learning algorithm is trained based on the labeled data only. This classifier is then applied to the unlabeled data to generate more labeled examples as input for the supervised learning algorithm. Generally only the labels the classifier is most confident in are added at each step. [15] In natural language processing, a common self-training algorithm is the Yarowsky algorithm for problems like word sense disambiguation, accent restoration, and spelling correction. [16]

Co-training is an extension of self-training in which multiple classifiers are trained on different (ideally disjoint) sets of features and generate labeled examples for one another. [17]

In human cognition

Human responses to formal semi-supervised learning problems have yielded varying conclusions about the degree of influence of the unlabeled data. [18] More natural learning problems may also be viewed as instances of semi-supervised learning. Much of human concept learning involves a small amount of direct instruction (e.g. parental labeling of objects during childhood) combined with large amounts of unlabeled experience (e.g. observation of objects without naming or counting them, or at least without feedback).

Human infants are sensitive to the structure of unlabeled natural categories such as images of dogs and cats or male and female faces. [19] Infants and children take into account not only unlabeled examples, but the sampling process from which labeled examples arise. [20] [21]

Weak Supervision in Predictive Maintenance

Weak supervision is an emerging machine learning approach in predictive maintenance that addresses the challenge of limited or imprecise labeled data. Traditional predictive maintenance models often rely on large volumes of accurately labeled failure and operational data, which can be costly or impractical to obtain. Weak supervision mitigates this by leveraging imperfect sources of supervision—such as noisy labels, heuristics, domain expert rules, or partially labeled datasets—to train predictive models. By incorporating techniques such as data programming, label modeling, and semi-supervised learning, weak supervision enables the development of robust predictive maintenance systems capable of identifying equipment failures or anomalies with reduced reliance on high-quality labeled data. This approach is particularly valuable in industrial settings where machine failures are rare and labeled fault data is scarce [22] .

See also

PU learning

References

Sources

Chapelle, Olivier; Schölkopf, Bernhard; Zien, Alexander (2006). Semi-supervised learning . Cambridge, Mass.: MIT Press. ISBN 978-0-262-03358-9 .

External links

Manifold Regularization A freely available MATLAB implementation of the graph-based semi-supervised algorithms Laplacian support vector machines and Laplacian regularized least squares.

KEEL: A software tool to assess evolutionary algorithms for Data Mining problems (regression, classification, clustering, pattern mining and so on) KEEL module for semi-supervised learning.

Semi-Supervised Learning Software

Semi-Supervised learning — scikit-learn documentation Semi-supervised learning in scikit-learn .

v

t

e

Differentiable programming

Information geometry

Statistical manifold

Automatic differentiation

Neuromorphic computing

Pattern recognition

Ricci calculus

Computational learning theory

Inductive bias

IPU

TPU

VPU

Memristor

SpiNNaker

TensorFlow

PyTorch

Keras

scikit-learn

Theano

JAX

Flux.jl

MindSpore

Portals Computer programming Technology

Computer programming

Technology