Title: Active learning (machine learning)

URL: https://en.wikipedia.org/wiki/Active_learning_(machine_learning)

PageID: 28801798

Categories: Category:Machine learning

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

Active learning is a special case of machine learning in which a learning algorithm can interactively query a human user (or some other information source), to label new data points with the desired outputs. The human user must possess knowledge/expertise in the problem domain, including the ability to consult/research authoritative sources when necessary. In statistics literature, it is sometimes also called optimal experimental design . The information source is also called teacher or oracle .

There are situations in which unlabeled data is abundant but manual labeling is expensive. In such a scenario, learning algorithms can actively query the user/teacher for labels. This type of iterative supervised learning is called active learning. Since the learner chooses the examples, the number of examples to learn a concept can often be much lower than the number required in normal

supervised learning. With this approach, there is a risk that the algorithm is overwhelmed by uninformative examples. Recent developments are dedicated to multi-label active learning, hybrid active learning and active learning in a single-pass (on-line) context, combining concepts from the field of machine learning (e.g. conflict and ignorance) with adaptive, incremental learning policies in the field of online machine learning . Using active learning allows for faster development of a machine learning algorithm, when comparative updates would require a quantum or super computer.

Large-scale active learning projects may benefit from crowdsourcing frameworks such as Amazon Mechanical Turk that include many humans in the active learning loop .

Definitions

Let T be the total set of all data under consideration. For example, in a protein engineering problem, T would include all proteins that are known to have a certain interesting activity and all additional proteins that one might want to test for that activity.

During each iteration, i , T is broken up into three subsets

T K , i {\displaystyle \mathbf {T} _{K,i}} : Data points where the label is known .

T U , i {\displaystyle \mathbf {T} _{U,i}} : Data points where the label is unknown .

T C , i {\displaystyle \mathbf {T} _{C,i}} : A subset of T U,i that is chosen to be labeled.

Most of the current research in active learning involves the best method to choose the data points for T C,i .

Scenarios

Pool-based sampling : In this approach, which is the most well known scenario, the learning algorithm attempts to evaluate the entire dataset before selecting data points (instances) for labeling. It is often initially trained on a fully labeled subset of the data using a machine-learning method such as logistic regression or SVM that yields class-membership probabilities for individual data instances. The candidate instances are those for which the prediction is most ambiguous. Instances are drawn from the entire data pool and assigned a confidence score, a measurement of how well the learner "understands" the data. The system then selects the instances for which it is the least confident and queries the teacher for the labels. The theoretical drawback of pool-based sampling is that it is memory-intensive and is therefore limited in its capacity to handle enormous datasets, but in practice, the rate-limiting factor is that the teacher is typically a (fatiguable) human expert who must be paid for their effort, rather than computer memory.

Stream-based selective sampling : Here, each consecutive unlabeled instance is examined one at a time with the machine evaluating the informativeness of each item against its query parameters. The learner decides for itself whether to assign a label or query the teacher for each datapoint. As contrasted with Pool-based sampling, the obvious drawback of stream-based methods is that the learning algorithm does not have sufficient information, early in the process, to make a sound assign-label-vs ask-teacher decision, and it does not capitalize as efficiently on the presence of already labeled data. Therefore, the teacher is likely to spend more effort in supplying labels than with the pool-based approach.

Membership query synthesis : This is where the learner generates synthetic data from an underlying natural distribution. For example, if the dataset are pictures of humans and animals, the learner could send a clipped image of a leg to the teacher and query if this appendage belongs to an animal or human. This is particularly useful if the dataset is small. The challenge here, as with all synthetic-data-generation efforts, is in ensuring that the synthetic data is consistent in terms of meeting the constraints on real data. As the number of variables/features in the input data increase, and strong dependencies between variables exist, it becomes increasingly difficult to generate synthetic data with sufficient fidelity. For example, to create a synthetic data set for human laboratory-test values, the sum of the various white blood cell (WBC) components in a white blood cell differential must equal 100, since the component numbers are really percentages. Similarly, the enzymes alanine transaminase (ALT) and aspartate transaminase (AST) measure liver function (though AST is also produced by other tissues, e.g., lung, pancreas) A synthetic data point with

AST at the lower limit of normal range (8–33 units/L) with an ALT several times above normal range (4–35 units/L) in a simulated chronically ill patient would be physiologically impossible.

Query strategies

Algorithms for determining which data points should be labeled can be organized into a number of different categories, based upon their purpose:

Balance exploration and exploitation : the choice of examples to label is seen as a dilemma between the exploration and the exploitation over the data space representation. This strategy manages this compromise by modelling the active learning problem as a contextual bandit problem. For example, Bouneffouf et al. propose a sequential algorithm named Active Thompson Sampling (ATS), which, in each round, assigns a sampling distribution on the pool, samples one point from this distribution, and queries the oracle for this sample point label.

Expected model change : label those points that would most change the current model.

Expected error reduction : label those points that would most reduce the model's generalization error .

Exponentiated Gradient Exploration for Active Learning : In this paper, the author proposes a sequential algorithm named exponentiated gradient (EG)-active that can improve any active learning algorithm by an optimal random exploration.

Uncertainty sampling : label those points for which the current model is least certain as to what the correct output should be.

Query by committee : a variety of models are trained on the current labeled data, and vote on the output for unlabeled data; label those points for which the "committee" disagrees the most

Querying from diverse subspaces or partitions : When the underlying model is a forest of trees, the leaf nodes might represent (overlapping) partitions of the original feature space . This offers the possibility of selecting instances from non-overlapping or minimally overlapping partitions for labeling.

Variance reduction : label those points that would minimize output variance, which is one of the components of error.

Conformal prediction : predicts that a new data point will have a label similar to old data points in some specified way and degree of the similarity within the old examples is used to estimate the confidence in the prediction.

Mismatch-first farthest-traversal : The primary selection criterion is the prediction mismatch between the current model and nearest-neighbour prediction. It targets on wrongly predicted data points. The second selection criterion is the distance to previously selected data, the farthest first. It aims at optimizing the diversity of selected data.

User-centered labeling strategies: Learning is accomplished by applying dimensionality reduction to graphs and figures like scatter plots. Then the user is asked to label the compiled data (categorical, numerical, relevance scores, relation between two instances.

A wide variety of algorithms have been studied that fall into these categories. While the traditional AL strategies can achieve remarkable performance, it is often challenging to predict in advance which strategy is the most suitable in aparticular situation. In recent years, meta-learning algorithms have been gaining in popularity. Some of them have been proposed to tackle the problem of learning AL strategies instead of relying on manually designed strategies. A benchmark which compares 'meta-learning approaches to active learning' to 'traditional heuristic-based Active Learning' may give intuitions if 'Learning active learning' is at the crossroads

Minimum marginal hyperplane

Some active learning algorithms are built upon support-vector machines (SVMs) and exploit the structure of the SVM to determine which data points to label. Such methods usually calculate the margin , W , of each unlabeled datum in T U,i and treat W as an n -dimensional distance from that datum to the separating hyperplane.

Minimum Marginal Hyperplane methods assume that the data with the smallest W are those that the SVM is most uncertain about and therefore should be placed in T C,i to be labeled. Other similar methods, such as Maximum Marginal Hyperplane, choose data with the largest W . Tradeoff methods choose a mix of the smallest and largest W s.

See also

List of datasets for machine learning research

Sample complexity

Bayesian Optimization

Reinforcement learning

Literature

Improving Generalization with Active Learning, David Cohn, Les Atlas & Richard Ladner, Machine Learning 15, 201–221 (1994). https://doi.org/10.1007/BF00993277

Balcan, Maria-Florina & Hanneke, Steve & Wortman, Jennifer. (2008). The True Sample Complexity of Active Learning.. 45-56. https://link.springer.com/article/10.1007/s10994-010-5174-y

Active Learning and Bayesian Optimization : a Unified Perspective to Learn with a Goal, Francesco Di Fiore, Michela Nardelli, Laura Mainini, https://arxiv.org/abs/2303.01560v2

Learning how to Active Learn: A Deep Reinforcement Learning Approach, Meng Fang, Yuan Li, Trevor Cohn, https://arxiv.org/abs/1708.02383v1

References