Title: Empirical risk minimization

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

PageID: 1455062

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

In statistical learning theory , the principle of empirical risk minimization defines a family of learning algorithms based on evaluating performance over a known and fixed dataset. The core idea is based on an application of the law of large numbers ; more specifically, we cannot know exactly how well a predictive algorithm will work in practice (i.e. the "true risk") because we do not know the true distribution of the data, but we can instead estimate and optimize the performance of the algorithm on a known set of training data. The performance over the known set of training data is referred to as the "empirical risk".

Background

The following situation is a general setting of many supervised learning problems. There are two spaces of objects X {\displaystyle X} and Y {\displaystyle Y} and we would like to learn a function h :

X → Y {\displaystyle \ h:X\to Y} (often called hypothesis) which outputs an object y ∈ Y {\displaystyle y\in Y} , given x ∈ X {\displaystyle x\in X} . To do so, there is a training set of n {\displaystyle n} examples (x 1 , y 1) , … , (x n , y n) {\displaystyle \ (x_{1},y_{1}),\ldots ,(x_{n},y_{n})} where x i ∈ X {\displaystyle x_{i}\in X} is an input and y i ∈ Y {\displaystyle y_{i}\in Y} is the corresponding response that is desired from h (x i) {\displaystyle h(x_{i})} .

To put it more formally, assuming that there is a joint probability distribution P (x , y) {\displaystyle P(x,y)} over X {\displaystyle X} and Y {\displaystyle Y} , and that the training set consists of n {\displaystyle n} instances (x 1 , y 1) , … , (x n , y n) {\displaystyle \ (x_{1},y_{1}),\ldots ,(x_{n},y_{n})} drawn i.i.d. from P (x , y) {\displaystyle P(x,y)} . The assumption of a joint probability distribution allows for the modelling of uncertainty in predictions (e.g. from noise in data) because y {\displaystyle y} is not a deterministic function of x {\displaystyle x} , but rather a random variable with conditional distribution P (y | x) {\displaystyle P(y|x)} for a fixed x {\displaystyle x} .

It is also assumed that there is a non-negative real-valued loss function L (y ^ , y) {\displaystyle L({\hat {y}},y)} which measures how different the prediction y ^ {\displaystyle {\hat {y}}} of a hypothesis is from the true outcome y {\displaystyle y} . For classification tasks, these loss functions can be scoring rules .

The risk associated with hypothesis h (x) {\displaystyle h(x)} is then defined as the expectation of the loss function:

A loss function commonly used in theory is the 0-1 loss function : L (y ^ , y) = { 1 if y ^ ≠ y 0 if y ^ = y {\displaystyle L({\hat {y}},y)={\begin{cases}1&{\mbox{ if }}\quad {\hat {y}}\neq y\\0&{\mbox{ if }}\quad {\hat {y}}=y\end{cases}}} .

The ultimate goal of a learning algorithm is to find a hypothesis h ∗ {\displaystyle h^{*}} among a fixed class of functions H {\displaystyle {\mathcal {H}}} for which the risk R (h) {\displaystyle R(h)} is minimal:

For classification problems, the Bayes classifier is defined to be the classifier minimizing the risk defined with the 0–1 loss function.

Formal definition

In general, the risk R (h) {\displaystyle R(h)} cannot be computed because the distribution P (x , y) {\displaystyle P(x,y)} is unknown to the learning algorithm. However, given a sample of iid training data points, we can compute an estimate , called the empirical risk , by computing the average of the loss function over the training set; more formally, computing the expectation with respect to the empirical measure :

The empirical risk minimization principle states that the learning algorithm should choose a hypothesis h ^ {\displaystyle {\hat {h}}} which minimizes the empirical risk over the hypothesis class H {\displaystyle {\mathcal {H}}} :

Thus, the learning algorithm defined by the empirical risk minimization principle consists in solving the above optimization problem.

Properties

Guarantees for the performance of empirical risk minimization depend strongly on the function class selected as well as the distributional assumptions made. In general, distribution-free methods are too coarse, and do not lead to practical bounds. However, they are still useful in deriving asymptotic properties of learning algorithms, such as consistency . In particular, distribution-free bounds on the performance of empirical risk minimization given a fixed function class can be derived using bounds on the VC complexity of the function class.

For simplicity, considering the case of binary classification tasks, it is possible to bound the probability of the selected classifier, φ n {\displaystyle \phi _{n}} being much worse than the best possible classifier φ ∗ {\displaystyle \phi ^{*}} . Consider the risk L {\displaystyle L} defined over the hypothesis class C {\displaystyle {\mathcal {C}}} with growth function S (C , n) {\displaystyle {\mathcal {S}}({\mathcal {C}},n)} given a dataset of size n {\displaystyle n} . Then, for every ■ > 0 {\displaystyle \epsilon >0} :

P (L (φ n) – L (φ ∗) > ■) ≤ 8 S (C , n) exp ■ { – n ■ 2 / 32 } {\displaystyle \mathbb {P} \left(L(\phi _{n})-L(\phi ^{*})>\epsilon \right)\leq {\mathcal {8}}S({\mathcal {C}},n)\exp\{-n\epsilon ^{2}/32\}}

Similar results hold for regression tasks. These results are often based on uniform laws of large numbers , which control the deviation of the empirical risk from the true risk, uniformly over the hypothesis class.

Impossibility results

It is also possible to show lower bounds on algorithm performance if no distributional assumptions are made. This is sometimes referred to as the No free lunch theorem . Even though a specific learning algorithm may provide the asymptotically optimal performance for any distribution, the finite sample performance is always poor for at least one data distribution. This means that no classifier can improve on the error for a given sample size for all distributions.

Specifically, let ■ > 0 {\displaystyle \epsilon >0} and consider a sample size n {\displaystyle n} and classification rule φ n {\displaystyle \phi _{n}} , there exists a distribution of (X , Y) {\displaystyle (X,Y)} with risk L ∗ = 0 {\displaystyle L^{*}=0} (meaning that perfect prediction is possible) such that: E L n ≥ 1 / 2 – ■ . {\displaystyle \mathbb {E} L_{n}\geq 1/2-\epsilon .}

It is further possible to show that the convergence rate of a learning algorithm is poor for some distributions. Specifically, given a sequence of decreasing positive numbers a i {\displaystyle a_{i}} converging to zero, it is possible to find a distribution such that:

E L n ≥ a i {\displaystyle \mathbb {E} L_{n}\geq a_{i}}

for all n {\displaystyle n} . This result shows that universally good classification rules do not exist, in the sense that the rule must be low quality for at least one distribution.

Computational complexity

Empirical risk minimization for a classification problem with a 0-1 loss function is known to be an NP-hard problem even for a relatively simple class of functions such as linear classifiers . Nevertheless, it can be solved efficiently when the minimal empirical risk is zero, i.e., data is linearly separable . [citation needed]

In practice, machine learning algorithms cope with this issue either by employing a convex approximation to the 0–1 loss function (like hinge loss for SVM), which is easier to optimize, or by imposing assumptions on the distribution P (x , y) {\displaystyle P(x,y)} (and thus stop being agnostic learning algorithms to which the above result applies).

In the case of convexification, Zhang's lemma majors the excess risk of the original problem using the excess risk of the convexified problem. Minimizing the latter using convex optimization also allow to control the former.

Tilted empirical risk minimization

Tilted empirical risk minimization is a machine learning technique used to modify standard loss functions like squared error, by introducing a tilt parameter. This parameter dynamically adjusts the weight of data points during training, allowing the algorithm to focus on specific regions or characteristics of the data distribution. Tilted empirical risk minimization is particularly useful in scenarios with imbalanced data or when there is a need to emphasize errors in certain parts of the prediction space.

See also

M-estimator

Maximum likelihood estimation

References

Further reading

Vapnik, V. (2000). The Nature of Statistical Learning Theory . Information Science and Statistics. Springer-Verlag . ISBN 978-0-387-98780-4 .