Title: Learnable function class

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

PageID: 48827727

Categories: Category:Machine learning

In statistical learning theory , a learnable function class is a set of functions for which an algorithm can be devised to asymptotically minimize the expected risk , uniformly over all probability distributions. The concept of learnable classes are closely related to regularization in machine learning , and provides large sample justifications for certain learning algorithms.

Definition

Background

Let Ω = X × Y = { (x , y) } {\displaystyle \Omega ={\mathcal {X}}\times {\mathcal {Y}}=\{(x,y)\}} be the sample space, where y {\displaystyle y} are the labels and x {\displaystyle x} are the covariates (predictors). F = { f : X ■ Y } {\displaystyle {\mathcal {F}}=\{f:{\mathcal {X}}\mapsto {\mathcal {Y}}\}} is a collection of mappings (functions) under consideration to link x {\displaystyle x} to y {\displaystyle y} . L : Y × Y ■ R {\displaystyle L:{\mathcal {Y}}\times {\mathcal {Y}}\mapsto \mathbb {R} } is a pre-given loss function (usually non-negative). Given a probability distribution P (x , y) {\displaystyle P(x,y)} on Ω {\displaystyle \Omega } , define the expected risk I P (f) {\displaystyle I_{P}(f)} to be:

The general goal in statistical learning is to find the function in F {\displaystyle {\mathcal {F}}} that minimizes the expected risk. That is, to find solutions to the following problem: [1]

But in practice the distribution P {\displaystyle P} is unknown, and any learning task can only be based on finite samples. Thus we seek instead to find an algorithm that asymptotically minimizes the empirical risk, i.e., to find a sequence of functions { f ^ n } n = 1 ∞ {\displaystyle \{{\hat {f}}_{n}\}_{n=1}^{\infty }} that satisfies

One usual algorithm to find such a sequence is through empirical risk minimization .

Learnable function class

We can make the condition given in the above equation stronger by requiring that the convergence is uniform for all probability distributions. That is:

The intuition behind the more strict requirement is as such: the rate at which sequence { f ^ n } {\displaystyle \{{\hat {f}}_{n}\}} converges to the minimizer of the expected risk can be very different for different P (x , y) {\displaystyle P(x,y)} . Because in real world the true distribution P {\displaystyle P} is always unknown, we would want to select a sequence that performs well under all cases.

However, by the no free lunch theorem , such a sequence that satisfies (1) does not exist if F {\displaystyle {\mathcal {F}}} is too complex. This means we need to be careful and not allow too "many" functions in F {\displaystyle {\mathcal {F}}} if we want (1) to be a meaningful requirement. Specifically, function classes that ensure the existence of a sequence { f ^ n } {\displaystyle \{{\hat {f}}_{n}\}} that satisfies (1) are known as learnable classes . [1]

It is worth noting that at least for supervised classification and regression problems, if a function class is learnable, then the empirical risk minimization automatically satisfies (1). [2] Thus in these settings not only do we know that the problem posed by (1) is solvable, we also immediately have an algorithm that gives the solution.

Interpretations

If the true relationship between y {\displaystyle y} and x {\displaystyle x} is y ~ f ∗ (x) {\displaystyle y\sim f^{*}(x)} , then by selecting the appropriate loss function, f ∗ {\displaystyle f^{*}} can always be expressed as the minimizer of the expected loss across all possible functions. That is,

Here we let F ∗ {\displaystyle {\mathcal {F}}^{*}} be the collection of all possible functions mapping X {\displaystyle {\mathcal {X}}} onto Y {\displaystyle {\mathcal {Y}}} . f ∗ {\displaystyle f^{*}} can be interpreted as the actual data generating mechanism. However, the no free lunch theorem tells us that in practice, with finite samples we cannot hope to search for the expected risk minimizer over F ∗ {\displaystyle {\mathcal {F}}^{*}} . Thus we often consider a subset of F ∗ {\displaystyle {\mathcal {F}}^{*}} , F {\displaystyle {\mathcal {F}}} , to carry out searches on. By doing so, we risk that f ∗ {\displaystyle f^{*}} might not be an element of F {\displaystyle {\mathcal {F}}} . This tradeoff can be mathematically expressed as

In the above decomposition, part (b) {\displaystyle (b)} does not depend on the data and is non-stochastic. It describes how far away our assumptions (F {\displaystyle {\mathcal {F}}}) are from the truth (F ∗ {\displaystyle {\mathcal {F}}^{*}}). (b) {\displaystyle (b)} will be strictly greater than 0 if we make assumptions that are too strong (F {\displaystyle {\mathcal {F}}} too small). On the other hand, failing to put enough restrictions on F {\displaystyle {\mathcal {F}}} will cause it to be not learnable, and part (a) {\displaystyle (a)} will not stochastically converge to 0. This is the well-known overfitting problem in statistics and machine learning literature.

Example: Tikhonov regularization

A good example where learnable classes are used is the so-called Tikhonov regularization in reproducing kernel Hilbert space (RKHS). Specifically, let F ∗ {\displaystyle {\mathcal {F^{*}}}} be an RKHS, and | | ⋅ | | 2 {\displaystyle ||\cdot ||_{2}} be the norm on F ∗ {\displaystyle {\mathcal {F^{*}}}} given by its inner product. It is shown in [3] that F = { f : | | f | | 2 ≤ γ } {\displaystyle {\mathcal {F}}=\{f:||f||_{2}\leq \gamma \}} is a learnable class for any finite, positive γ {\displaystyle \gamma } . The empirical minimization algorithm to the dual form of this problem is

This was first introduced by Tikhonov [4] to solve ill-posed problems. Many statistical learning algorithms can be expressed in such a form (for example, the well-known ridge regression).

The tradeoff between (a) {\displaystyle (a)} and (b) {\displaystyle (b)} in (2) is geometrically more intuitive with Tikhonov regularization in RKHS. We can consider a sequence of { F γ } {\displaystyle \{{\mathcal {F}}_{\gamma }\}} , which are essentially balls in F ∗ {\displaystyle {\mathcal {F^{*}}}} with centers at 0. As γ {\displaystyle \gamma } gets larger, F γ {\displaystyle {\mathcal {F}}_{\gamma }} gets closer to the entire space, and (b) {\displaystyle (b)} is likely to become smaller. However we will also suffer smaller convergence rates in (a) {\displaystyle (a)} . The way to choose an optimal γ {\displaystyle \gamma } in finite sample settings is usually through cross-validation .

Relationship to empirical process theory

Part (a) {\displaystyle (a)} in (2) is closely linked to empirical process theory in statistics, where the empirical risk { ∑ i = 1 n L (y i , f (x i)) , f ∈ F } {\displaystyle \{\sum _{i=1}^{n}L(y_{i},f(x_{i})),f\in {\mathcal {F}}\}} are known as empirical processes. [5] In this field, the function class F {\displaystyle {\mathcal {F}}} that satisfies the stochastic convergence

are known as uniform Glivenko–Cantelli classes . It has been shown that under certain regularity conditions, learnable classes and uniformly Glivenko-Cantelli classes are equivalent. [1] Interplay between (a) {\displaystyle (a)} and (b) {\displaystyle (b)} in statistics literature is often known as the bias-variance tradeoff .

However, note that in [2] the authors gave an example of stochastic convex optimization for General Setting of Learning where learnability is not equivalent with uniform convergence.

References