Title: Bayesian interpretation of kernel regularization

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

PageID: 35867897

Categories: Category:Bayesian statistics, Category:Machine learning

-----

Bayesian interpretation of kernel regularization examines how kernel methods in machine learning can be understood through the lens of Bayesian statistics , a framework that uses probability to model uncertainty. Kernel methods are founded on the concept of similarity between inputs within a structured space. While techniques like support vector machines (SVMs) and their regularization (a technique to make a model more generalizable and transferable) were not originally formulated using Bayesian principles, analyzing them from a Bayesian perspective provides valuable insights.

In the Bayesian framework, kernel methods serve as a fundamental component of Gaussian processes , where the kernel function operates as a covariance function that defines relationships between inputs. Traditionally, these methods have been applied to supervised learning problems where inputs are represented as vectors and outputs as scalars. Recent developments have extended kernel methods to handle multiple outputs , as seen in multi-task learning .

The mathematical framework for kernel methods typically involves reproducing kernel Hilbert spaces (RKHS). Not all kernels form inner product spaces, as they may not always be positive semidefinite (a property ensuring non-negative similarity measures), but they still operate within these more general RKHS. A mathematical equivalence between regularization approaches and Bayesian methods can be established, particularly in cases where the reproducing kernel Hilbert space is finite-dimensional. This equivalence demonstrates how both perspectives converge to essentially the same estimators , revealing the underlying connection between these seemingly different approaches.

The supervised learning problem

The classical supervised learning problem requires estimating the output for some new input point x ′ {\displaystyle \mathbf {x} '} by learning a scalar-valued estimator f ^ (x ′) {\displaystyle {\hat {f}}(\mathbf {x} ')} on the basis of a training set S {\displaystyle S} consisting of n {\displaystyle n} input-output pairs, S = (X , Y) = (x 1 , y 1) , … , (x n , y n) {\displaystyle S=(\mathbf {X} ,\mathbf {Y})=(\mathbf {x} _{1},y_{1}),\ldots ,(\mathbf {x} _{n},y_{n})} . Given a symmetric and positive bivariate function k (⋅ , ⋅) {\displaystyle k(\cdot ,\cdot)} called a kernel , one of the most popular estimators in machine learning is given by

where K ≡ k (X , X) {\displaystyle \mathbf {K} \equiv k(\mathbf {X} ,\mathbf {X})} is the kernel matrix with entries K i j = k (x i , x j) {\displaystyle \mathbf {K} _{ij}=k(\mathbf {x} _{i},\mathbf {x} _{j})} , k = [k (x 1 , x ′) , … , k (x n , x ′)] ■ {\displaystyle \mathbf {k} =[k(\mathbf {x} _{1},\mathbf {x} '),\ldots ,k(\mathbf {x} _{n},\mathbf {x} ')]^{\top }} , and Y = [y 1 , … , y n] ■ {\displaystyle \mathbf {Y} =[y_{1},\ldots ,y_{n}]^{\top }} . We will see how this estimator can be derived both from a regularization and a Bayesian perspective.

A regularization perspective

The main assumption in the regularization perspective is that the set of functions F {\displaystyle {\mathcal {F}}} is assumed to belong to a reproducing kernel Hilbert space H k {\displaystyle {\mathcal {H}}_{k}} .

Reproducing kernel Hilbert space

A reproducing kernel Hilbert space (RKHS) H k {\displaystyle {\mathcal {H}}_{k}} is a Hilbert space of functions defined by a symmetric , positive-definite function k : X × X → R {\displaystyle k:{\mathcal {X}}\times {\mathcal {X}}\rightarrow \mathbb {R} } called the reproducing kernel such that the function k (x , ⋅) {\displaystyle k(\mathbf {x} ,\cdot)} belongs to H k {\displaystyle {\mathcal

{H}}_{k}} for all x ∈ X {\displaystyle \mathbf {x} \in {\mathcal {X}}} . There are three main properties that make an RKHS appealing:

1. The reproducing property , after which the RKHS is named,

where ■ · , · ■ k {\displaystyle \langle \cdot ,\cdot \rangle _{k}} is the inner product in H k {\displaystyle {\mathcal {H}}_{k}} .

2. Functions in an RKHS are in the closure of the linear combination of the kernel at given points,

This allows the construction in a unified framework of both linear and generalized linear models.

3. The squared norm in an RKHS can be written as

and could be viewed as measuring the complexity of the function.

The regularized functional

The estimator is derived as the minimizer of the regularized functional

where f ∈ H k {\displaystyle f\in {\mathcal {H}}_{k}} and ■ · ■ k {\displaystyle \|\cdot \|_{k}} is the norm in H k {\displaystyle {\mathcal {H}}_{k}} . The first term in this functional, which measures the average of the squares of the errors between the f (x i) {\displaystyle f(\mathbf {x} _{i})} and the y i {\displaystyle y_{i}} , is called the empirical risk and represents the cost we pay by predicting f (x i) {\displaystyle f(\mathbf {x} _{i})} for the true value y i {\displaystyle y_{i}} . The second term in the functional is the squared norm in a RKHS multiplied by a weight λ {\displaystyle \lambda } and serves the purpose of stabilizing the problem as well as of adding a trade-off between fitting and complexity of the estimator. The weight λ {\displaystyle \lambda } , called the regularizer , determines the degree to which instability and complexity of the estimator should be penalized (higher penalty for increasing value of λ {\displaystyle \lambda }).

Derivation of the estimator

The explicit form of the estimator in equation (1) is derived in two steps. First, the representer theorem states that the minimizer of the functional (2) can always be written as a linear combination of the kernels centered at the training-set points,

for some c ∈ R n {\displaystyle \mathbf {c} \in \mathbb {R} ^{n}} . The explicit form of the coefficients c = [c 1 , … , c n] ■ {\displaystyle \mathbf {c} =[c_{1},\ldots ,c_{n}]^{\top }} can be found by substituting for f (·) {\displaystyle f(\cdot)} in the functional (2). For a function of the form in equation (3), we have that

We can rewrite the functional (2) as

This functional is convex in c {\displaystyle \mathbf {c} } and therefore we can find its minimum by setting the gradient with respect to c {\displaystyle \mathbf {c} } to zero,

Substituting this expression for the coefficients in equation (3), we obtain the estimator stated previously in equation (1),

A Bayesian perspective

The notion of a kernel plays a crucial role in Bayesian probability as the covariance function of a stochastic process called the Gaussian process .

A review of Bayesian probability

As part of the Bayesian framework, the Gaussian process specifies the prior distribution that describes the prior beliefs about the properties of the function being modeled. These beliefs are updated after taking into account observational data by means of a likelihood function that relates the prior beliefs to the observations. Taken together, the prior and likelihood lead to an updated distribution called the posterior distribution that is customarily used for predicting test cases.

The Gaussian process

A Gaussian process (GP) is a stochastic process in which any finite number of random variables that are sampled follow a joint Normal distribution . The mean vector and covariance matrix of the

Gaussian distribution completely specify the GP. GPs are usually used as a priori distribution for functions, and as such the mean vector and covariance matrix can be viewed as functions, where the covariance function is also called the kernel of the GP. Let a function f {\displaystyle f} follow a Gaussian process with mean function m {\displaystyle m} and kernel function k {\displaystyle k} ,

In terms of the underlying Gaussian distribution, we have that for any finite set X = { x i } i = 1 n {\displaystyle \mathbf {X} =\{\mathbf {x} _{i}\}_{i=1}^{n}} if we let f (X) = [f (x 1) , … , f (x n)] ■ {\displaystyle f(\mathbf {X})=[f(\mathbf {x} _{1}),\ldots ,f(\mathbf {x} _{n})]^{\top }} then

where m = m (X) = [m (x 1) , … , m (x N)] ■ {\displaystyle \mathbf {m} =m(\mathbf {X})=[m(\mathbf {x} _{1}),\ldots ,m(\mathbf {x} _{N})]^{\top }} is the mean vector and K = k (X , X) {\displaystyle \mathbf {K} =k(\mathbf {X} ,\mathbf {X})} is the covariance matrix of the multivariate Gaussian distribution.

Derivation of the estimator

In a regression context, the likelihood function is usually assumed to be a Gaussian distribution and the observations to be independent and identically distributed (iid),

This assumption corresponds to the observations being corrupted with zero-mean Gaussian noise with variance σ 2 {\displaystyle \sigma ^{2}} . The iid assumption makes it possible to factorize the likelihood function over the data points given the set of inputs X {\displaystyle \mathbf {X} } and the variance of the noise σ 2 {\displaystyle \sigma ^{2}} , and thus the posterior distribution can be computed analytically. For a test input vector x ′ {\displaystyle \mathbf {x} '} , given the training data S = { X , Y } {\displaystyle S=\{\mathbf {X} ,\mathbf {Y} \}} , the posterior distribution is given by

where φ {\displaystyle {\boldsymbol {\phi }}} denotes the set of parameters which include the variance of the noise σ 2 {\displaystyle \sigma ^{2}} and any parameters from the covariance function k {\displaystyle k} and where

The connection between regularization and Bayes

A connection between regularization theory and Bayesian theory can only be achieved in the case of finite dimensional RKHS . Under this assumption, regularization theory and Bayesian theory are connected through Gaussian process prediction.

In the finite dimensional case, every RKHS can be described in terms of a feature map Φ : X → R p {\displaystyle \Phi :{\mathcal {X}}\rightarrow \mathbb {R} ^{p}} such that

Functions in the RKHS with kernel K {\displaystyle \mathbf {K} } can then be written as

and we also have that

We can now build a Gaussian process by assuming w = [w 1 , … , w p] ■ {\displaystyle \mathbf {w} =[w^{1},\ldots ,w^{p}]^{\top }} to be distributed according to a multivariate Gaussian distribution with zero mean and identity covariance matrix,

If we assume a Gaussian likelihood we have

where f w (X) = (■ w , Φ (x 1) ■ , … , ■ w , Φ (x n ■) {\displaystyle f_{\mathbf {w} }(\mathbf {X})=(\langle \mathbf {w} ,\Phi (\mathbf {x} _{1})\rangle ,\ldots ,\langle \mathbf {w} ,\Phi (\mathbf {x} _{n}\rangle)} . The resulting posterior distribution is then given by

We can see that a maximum posterior (MAP) estimate is equivalent to the minimization problem defining Tikhonov regularization , where in the Bayesian case the regularization parameter is related to the noise variance.

From a philosophical perspective, the loss function in a regularization setting plays a different role than the likelihood function in the Bayesian setting. Whereas the loss function measures the error that is incurred when predicting f (x) {\displaystyle f(\mathbf {x})} in place of y {\displaystyle y} , the likelihood function measures how likely the observations are from the model that was assumed to be true in the generative process. From a mathematical perspective, however, the formulations of the regularization and Bayesian frameworks make the loss function and the likelihood function to have the same mathematical role of promoting the inference of functions f {\displaystyle f} that approximate the labels y {\displaystyle y} as much as possible.

See also

Regularized least squares

Bayesian linear regression

Bayesian interpretation of Tikhonov regularization

References