Title: Kernel embedding of distributions

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

PageID: 41370976

Categories: Category:Machine learning, Category:Theory of probability distributions

---

In machine learning , the kernel embedding of distributions (also called the kernel mean or mean map) comprises a class of nonparametric methods in which a probability distribution is represented as an element of a reproducing kernel Hilbert space (RKHS). [1] A generalization of the individual data-point feature mapping done in classical kernel methods , the embedding of distributions into infinite-dimensional feature spaces can preserve all of the statistical features of arbitrary distributions, while allowing one to compare and manipulate distributions using Hilbert space operations such as inner products , distances, projections , linear transformations , and spectral analysis . [2] This learning framework is very general and can be applied to distributions over any space Ω {\displaystyle \Omega } on which a sensible kernel function (measuring similarity between elements of Ω {\displaystyle \Omega }) may be defined. For example, various kernels have been proposed for learning from data which are: vectors in R d {\displaystyle \mathbb {R} ^{d}} , discrete classes/categories, strings , graphs / networks , images, time series , manifolds , dynamical systems , and other structured objects. [3] [4] The theory behind kernel embeddings of distributions has been primarily developed by Alex Smola , Le Song Archived 2021-04-12 at the Wayback Machine , Arthur Gretton , and Bernhard Schölkopf . A review of recent works on kernel embedding of distributions can be found in. [5]

The analysis of distributions is fundamental in machine learning and statistics , and many algorithms in these fields rely on information theoretic approaches such as entropy , mutual information , or Kullback–Leibler divergence . However, to estimate these quantities, one must first either perform density estimation, or employ sophisticated space-partitioning/bias-correction strategies which are typically infeasible for high-dimensional data. [6] Commonly, methods for modeling complex distributions rely on parametric assumptions that may be unfounded or computationally challenging (e.g. Gaussian mixture models), while nonparametric methods like kernel density estimation (Note: the smoothing kernels in this context have a different interpretation than the kernels discussed here) or characteristic function representation (via the Fourier transform of the distribution) break down in high-dimensional settings. [2]

Methods based on the kernel embedding of distributions sidestep these problems and also possess the following advantages: [6]

Data may be modeled without restrictive assumptions about the form of the distributions and relationships between variables

Intermediate density estimation is not needed

Practitioners may specify the properties of a distribution most relevant for their problem (incorporating prior knowledge via choice of the kernel)

If a characteristic kernel is used, then the embedding can uniquely preserve all information about a distribution, while thanks to the kernel trick , computations on the potentially infinite-dimensional RKHS can be implemented in practice as simple Gram matrix operations

Dimensionality-independent rates of convergence for the empirical kernel mean (estimated using samples from the distribution) to the kernel embedding of the true underlying distribution can be proven.

Learning algorithms based on this framework exhibit good generalization ability and finite sample convergence, while often being simpler and more effective than information theoretic methods

Thus, learning via the kernel embedding of distributions offers a principled drop-in replacement for information theoretic approaches and is a framework which not only subsumes many popular methods in machine learning and statistics as special cases, but also can lead to entirely new learning algorithms.

Definitions

Let X {\displaystyle X} denote a random variable with domain Ω {\displaystyle \Omega } and distribution P {\displaystyle P} . Given a symmetric, positive-definite kernel k : Ω × Ω → R {\displaystyle k:\Omega \times \Omega \rightarrow \mathbb {R} } the Moore–Aronszajn theorem asserts the existence of a unique RKHS H {\displaystyle {\mathcal {H}}} on Ω {\displaystyle \Omega } (a Hilbert space of functions f : Ω → R {\displaystyle f:\Omega \to \mathbb {R} } equipped with an inner product ■ ⋅ , ⋅ ■ H {\displaystyle \langle \cdot ,\cdot \rangle _{\mathcal {H}}} and a norm ■ ⋅ ■ H {\displaystyle \|\cdot \|_{\mathcal {H}}}) for which k {\displaystyle k} is a reproducing kernel, i.e., in which the element k (x , ⋅) {\displaystyle k(x,\cdot)} satisfies the reproducing property

One may alternatively consider x ■ k (x , ⋅) {\displaystyle x\mapsto k(x,\cdot)} as an implicit feature mapping ϕ : Ω → H {\displaystyle \varphi :\Omega \rightarrow {\mathcal {H}}} (which is therefore also called the feature space), so that k (x , x ′) = ■ ϕ (x) , ϕ (x ′) ■ H {\displaystyle k(x,x')=\langle \varphi (x),\varphi (x')\rangle _{\mathcal {H}}} can be viewed as a measure of similarity between points x , x ′ ∈ Ω . {\displaystyle x,x'\in \Omega .} While the similarity measure is linear in the feature space, it may be highly nonlinear in the original space depending on the choice of kernel.

Kernel embedding

The kernel embedding of the distribution P {\displaystyle P} in H {\displaystyle {\mathcal {H}}} (also called the kernel mean or mean map) is given by: [1]

If P {\displaystyle P} allows a square integrable density p {\displaystyle p} , then μ X = E k p {\displaystyle \mu _{X}={\mathcal {E}}_{k}p} , where E k {\displaystyle {\mathcal {E}}_{k}} is the Hilbert–Schmidt integral operator . A kernel is characteristic if the mean embedding μ : { family of distributions over Ω } → H {\displaystyle \mu :\{{\text{family of distributions over }}\Omega \}\to {\mathcal {H}}} is injective. [7] Each distribution can thus be uniquely represented in the RKHS and all statistical features of distributions are preserved by the kernel embedding if a characteristic kernel is used.

Empirical kernel embedding

Given n {\displaystyle n} training examples { x 1 , … , x n } {\displaystyle \{x_{1},\ldots ,x_{n}\}} drawn independently and identically distributed (i.i.d.) from P , {\displaystyle P,} the kernel embedding of P {\displaystyle P} can be empirically estimated as

Joint distribution embedding

If Y {\displaystyle Y} denotes another random variable (for simplicity, assume the co-domain of Y {\displaystyle Y} is also Ω {\displaystyle \Omega } with the same kernel k {\displaystyle k} which satisfies ■ ϕ (x) ⊗ ϕ (y) , ϕ (x ′) ⊗ ϕ (y ′) ■ = k (x , x ′) k (y , y ′) {\displaystyle \langle \varphi (x)\otimes \varphi (y),\varphi (x')\otimes \varphi (y')\rangle =k(x,x')k(y,y')}), then the joint distribution P (x , y)) {\displaystyle P(x,y))} can be mapped into a tensor product feature space H ⊗ H {\displaystyle {\mathcal {H}}\otimes {\mathcal {H}}} via [2]

By the equivalence between a tensor and a linear map , this joint embedding may be interpreted as an uncentered cross-covariance operator C X Y : H → H {\displaystyle {\mathcal {C}}_{XY}:{\mathcal {H}}\to {\mathcal {H}}} from which the cross-covariance of functions f , g ∈ H {\displaystyle f,g\in {\mathcal {H}}} can be computed as [8]

Given n {\displaystyle n} pairs of training examples { (x 1 , y 1) , … , (x n , y n) } {\displaystyle \{(x_{1},y_{1}),\dots ,(x_{n},y_{n})\}} drawn i.i.d. from P {\displaystyle P} , we can also empirically estimate the joint distribution kernel embedding via

Conditional distribution embedding

Given a conditional distribution P (y ■ x) , {\displaystyle P(y\mid x),} one can define the corresponding RKHS embedding as [2]

Note that the embedding of P (y ■ x) {\displaystyle P(y\mid x)} thus defines a family of points in the RKHS indexed by the values x {\displaystyle x} taken by conditioning variable X {\displaystyle X} . By fixing X {\displaystyle X} to a particular value, we obtain a single element in H {\displaystyle {\mathcal {H}}} , and thus it is natural to define the operator

which given the feature mapping of x {\displaystyle x} outputs the conditional embedding of Y {\displaystyle Y} given X = x . {\displaystyle X=x.} Assuming that for all g ∈ H : E [g (Y) ■ X] ∈ H , {\displaystyle g\in {\mathcal {H}}:\mathbb {E} [g(Y)\mid X]\in {\mathcal {H}},} it can be shown that [8]

This assumption is always true for finite domains with characteristic kernels, but may not necessarily hold for continuous domains. [2] Nevertheless, even in cases where the assumption fails, C Y ■ X ϕ (x) {\displaystyle {\mathcal {C}}_{Y\mid X}\varphi (x)} may still be used to approximate the conditional kernel embedding μ Y ■ x , {\displaystyle \mu _{Y\mid x},} and in practice, the inversion operator is replaced with a regularized version of itself (C X X + λ I) − 1 {\displaystyle ({\mathcal {C}}_{XX}+\lambda \mathbf {I})^{-1}} (where I {\displaystyle \mathbf {I} } denotes the identity matrix).

Given training examples { (x 1 , y 1) , … , (x n , y n) } , {\displaystyle \{(x_{1},y_{1}),\dots ,(x_{n},y_{n})\},} the empirical kernel conditional embedding operator may be estimated as [2]

where Φ = (ϕ (y 1) , … , ϕ (y n)) , Υ = (ϕ (x 1) , … , ϕ (x n)) {\displaystyle {\boldsymbol {\Phi }}=\left(\varphi (y_{1}),\dots ,\varphi (y_{n})\right),{\boldsymbol {\Upsilon }}=\left(\varphi (x_{1}),\dots ,\varphi (x_{n})\right)} are implicitly formed feature matrices, K = Υ T Υ {\displaystyle \mathbf {K} ={\boldsymbol {\Upsilon }}^{T}{\boldsymbol {\Upsilon }}} is the Gram matrix for samples of X {\displaystyle X} , and λ {\displaystyle \lambda } is a regularization parameter needed to avoid overfitting .

Thus, the empirical estimate of the kernel conditional embedding is given by a weighted sum of samples of Y {\displaystyle Y} in the feature space:

where β (x) = (K + λ I) − 1 K x {\displaystyle {\boldsymbol {\beta }}(x)=(\mathbf {K} +\lambda \mathbf {I})^{-1}\mathbf {K} _{x}} and K x = (k (x 1 , x) , … , k (x n , x)) T {\displaystyle \mathbf {K} _{x}=\left(k(x_{1},x),\dots ,k(x_{n},x)\right)^{T}}

Properties

The expectation of any function f {\displaystyle f} in the RKHS can be computed as an inner product with the kernel embedding:

In the presence of large sample sizes, manipulations of the n × n {\displaystyle n\times n} Gram matrix may be computationally demanding. Through use of a low-rank approximation of the Gram matrix (such as the incomplete Cholesky factorization), running time and memory requirements of kernel-embedding-based learning algorithms can be drastically reduced without suffering much loss in approximation accuracy. [2]

Convergence of empirical kernel mean to the true distribution embedding

If k {\displaystyle k} is defined such that f {\displaystyle f} takes values in [0 , 1] {\displaystyle [0,1]} for all f ∈ H {\displaystyle f\in {\mathcal {H}}} with ■ f ■ H ≤ 1 {\displaystyle \|f\|_{\mathcal {H}}\leq 1} (as is the case for the widely used radial basis function kernels), then with probability at least 1 − δ {\displaystyle 1-\delta } : [6]

The rate of convergence (in RKHS norm) of the empirical kernel embedding to its distribution counterpart is O (n − 1 / 2) {\displaystyle O(n^{-1/2})} and does not depend on the dimension of X {\displaystyle X} .

Statistics based on kernel embeddings thus avoid the curse of dimensionality , and though the true underlying distribution is unknown in practice, one can (with high probability) obtain an approximation within O (n − 1 / 2) {\displaystyle O(n^{-1/2})} of the true kernel embedding based on a finite sample of size n {\displaystyle n} .

For the embedding of conditional distributions, the empirical estimate can be seen as a weighted average of feature mappings (where the weights β i (x) {\displaystyle \beta _{i}(x)} depend on the value of the conditioning variable and capture the effect of the conditioning on the kernel embedding). In this case, the empirical estimate converges to the conditional distribution RKHS embedding with rate O (n – 1 / 4) {\displaystyle O\left(n^{-1/4}\right)} if the regularization parameter λ {\displaystyle \lambda } is decreased as O (n – 1 / 2) , {\displaystyle O\left(n^{-1/2}\right),} though faster rates of convergence may be achieved by placing additional assumptions on the joint distribution. [2]

Universal kernels

Let X ⊆ R b {\displaystyle {\mathcal {X}}\subseteq \mathbb {R} ^{b}} be a compact metric space and C (X) {\displaystyle C({\mathcal {X}})} the set of continuous functions . The reproducing kernel k : X × X → R {\displaystyle k:{\mathcal {X}}\times {\mathcal {X}}\rightarrow \mathbb {R} } is called universal if and only if the RKHS H {\displaystyle {\mathcal {H}}} of k {\displaystyle k} is dense in C (X) {\displaystyle C({\mathcal {X}})} , i.e., for any g ∈ C (X) {\displaystyle g\in C({\mathcal {X}})} and all ε > 0 {\displaystyle \varepsilon >0} there exists an f ∈ H {\displaystyle f\in {\mathcal {H}}} such that ■ f – g ■ ∞ ≤ ε {\displaystyle \|f-g\|_{\infty }\leq \varepsilon } . [9] All universal kernels defined on a compact space are characteristic kernels but the converse is not always true. [10]

Let k {\displaystyle k} be a continuous translation invariant kernel k (x , x ′) = h (x – x ′) {\displaystyle k(x,x')=h(x-x')} with x ∈ R b {\displaystyle x\in \mathbb {R} ^{b}} . Then Bochner's theorem guarantees the existence of a unique finite Borel measure μ {\displaystyle \mu } (called the spectral measure) on R b {\displaystyle \mathbb {R} ^{b}} such that

If k {\displaystyle k} induces a strictly positive definite kernel matrix for any set of distinct points, then it is a universal kernel. [6] For example, the widely used Gaussian RBF kernel

Parameter selection for conditional distribution kernel embeddings

The empirical kernel conditional distribution embedding operator C ^ Y | X {\displaystyle {\widehat {\mathcal {C}}}_{Y|X}} can alternatively be viewed as the solution of the following regularized least squares (function-valued) regression problem [14]

One can thus select the regularization parameter λ {\displaystyle \lambda } by performing cross-validation based on the squared loss function of the regression problem.

Rules of probability as operations in the RKHS

This section illustrates how basic probabilistic rules may be reformulated as (multi)linear algebraic operations in the kernel embedding framework and is primarily based on the work of Song et al. [2] [8] The following notation is adopted:

P (X , Y) = {\displaystyle P(X,Y)=} joint distribution over random variables X , Y {\displaystyle X,Y}

P (X) = ∫ Ω P (X , d y) = {\displaystyle P(X)=\int _{\Omega }P(X,\mathrm {d} y)=} marginal distribution of X {\displaystyle X} ; P (Y) = {\displaystyle P(Y)=} marginal distribution of Y {\displaystyle Y}

P (Y ■ X) = P (X , Y) P (X) = {\displaystyle P(Y\mid X)={\frac {P(X,Y)}{P(X)}}=} conditional distribution of Y {\displaystyle Y} given X {\displaystyle X} with corresponding conditional embedding operator C Y ■ X {\displaystyle {\mathcal {C}}_{Y\mid X}}

π (Y) = {\displaystyle \pi (Y)=} prior distribution over Y {\displaystyle Y}

Q {\displaystyle Q} is used to distinguish distributions which incorporate the prior from distributions P {\displaystyle P} which do not rely on the prior

In practice, all embeddings are empirically estimated from data { (x 1 , y 1) , … , (x n , y n) } {\displaystyle \{(x_{1},y_{1}),\dots ,(x_{n},y_{n})\}} and it assumed that a set of samples { y ~ 1 , … , y ~ n ~ } {\displaystyle \{{\widetilde {y}}_{1},\ldots ,{\widetilde {y}}_{\widetilde {n}}\}} may be used to estimate the kernel embedding of the prior distribution π (Y) {\displaystyle \pi (Y)} .

Kernel sum rule

In probability theory, the marginal distribution of X {\displaystyle X} can be computed by integrating out Y {\displaystyle Y} from the joint density (including the prior distribution on Y {\displaystyle Y})

The analog of this rule in the kernel embedding framework states that μ X π , {\displaystyle \mu _{X}^{\pi },} the RKHS embedding of Q (X) {\displaystyle Q(X)} , can be computed via

where μ Y π {\displaystyle \mu _{Y}^{\pi }} is the kernel embedding of π (Y) . {\displaystyle \pi (Y).} In practical implementations, the kernel sum rule takes the following form

where

is the empirical kernel embedding of the prior distribution, α = (α 1 , … , α n ~) T , {\displaystyle {\boldsymbol {\alpha }}=(\alpha _{1},\ldots ,\alpha _{\widetilde {n}})^{T},} Υ = (ϕ (x 1) , … , ϕ (x n)) {\displaystyle {\boldsymbol {\Upsilon }}=\left(\varphi (x_{1}),\ldots ,\varphi (x_{n})\right)} , and G , G ~ {\displaystyle \mathbf {G} ,{\widetilde {\mathbf {G} }}} are Gram matrices with entries G i j = k (y i , y j) , G ~ i j = k (y i , y ~ j) {\displaystyle \mathbf {G} _{ij}=k(y_{i},y_{j}),{\widetilde {\mathbf {G} }}_{ij}=k(y_{i},{\widetilde {y}}_{j})} respectively.

Kernel chain rule

In probability theory, a joint distribution can be factorized into a product between conditional and marginal distributions

The analog of this rule in the kernel embedding framework states that C X Y π , {\displaystyle {\mathcal {C}}_{XY}^{\pi },} the joint embedding of Q (X , Y) , {\displaystyle Q(X,Y),} can be factorized as a composition of conditional embedding operator with the auto-covariance operator associated with π (Y) {\displaystyle \pi (Y)}

where

In practical implementations, the kernel chain rule takes the following form

Kernel Bayes' rule

In probability theory, a posterior distribution can be expressed in terms of a prior distribution and a likelihood function as

The analog of this rule in the kernel embedding framework expresses the kernel embedding of the conditional distribution in terms of conditional embedding operators which are modified by the prior distribution

where from the chain rule:

In practical implementations, the kernel Bayes' rule takes the following form

where

Two regularization parameters are used in this framework: λ {\displaystyle \lambda } for the estimation of C ^ Y X π , C ^ X X π = Υ D Υ T {\displaystyle {\widehat {\mathcal {C}}}_{YX}^{\pi },{\widehat {\mathcal {C}}}_{XX}^{\pi }={\boldsymbol {\Upsilon }}\mathbf {D} {\boldsymbol {\Upsilon }}^{T}} and λ ~ {\displaystyle {\widetilde {\lambda }}} for the estimation of the final conditional embedding operator

The latter regularization is done on square of C ^ X X π {\displaystyle {\widehat {\mathcal {C}}}_{XX}^{\pi }} because D {\displaystyle D} may not be positive definite .

Applications

Measuring distance between distributions

The maximum mean discrepancy (MMD) is a distance-measure between distributions P (X) {\displaystyle P(X)} and Q (Y) {\displaystyle Q(Y)} which is defined as the distance between their embeddings in the RKHS [6]

While most distance-measures between distributions such as the widely used Kullback–Leibler divergence either require density estimation (either parametrically or nonparametrically) or space partitioning/bias correction strategies, [6] the MMD is easily estimated as an empirical mean which

is concentrated around the true value of the MMD. The characterization of this distance as the maximum mean discrepancy refers to the fact that computing the MMD is equivalent to finding the RKHS function that maximizes the difference in expectations between the two probability distributions

a form of integral probability metric .

Kernel two-sample test

Given n training examples from P (X) {\displaystyle P(X)} and m samples from Q (Y) {\displaystyle Q(Y)} , one can formulate a test statistic based on the empirical estimate of the MMD

to obtain a two-sample test [15] of the null hypothesis that both samples stem from the same distribution (i.e. P = Q {\displaystyle P=Q}) against the broad alternative P ≠ Q {\displaystyle P\neq Q} .

Density estimation via kernel embeddings

Although learning algorithms in the kernel embedding framework circumvent the need for intermediate density estimation, one may nonetheless use the empirical embedding to perform density estimation based on n samples drawn from an underlying distribution P X ∗ {\displaystyle P_{X}^{*}} . This can be done by solving the following optimization problem [6] [16]

where the maximization is done over the entire space of distributions on Ω . {\displaystyle \Omega .} Here, μ X [P X] {\displaystyle \mu _{X}[P_{X}]} is the kernel embedding of the proposed density P X {\displaystyle P_{X}} and H {\displaystyle H} is an entropy-like quantity (e.g. Entropy , KL divergence , Bregman divergence). The distribution which solves this optimization may be interpreted as a compromise between fitting the empirical kernel means of the samples well, while still allocating a substantial portion of the probability mass to all regions of the probability space (much of which may not be represented in the training examples). In practice, a good approximate solution of the difficult optimization may be found by restricting the space of candidate densities to a mixture of M candidate distributions with regularized mixing proportions. Connections between the ideas underlying Gaussian processes and conditional random fields may be drawn with the estimation of conditional probability distributions in this fashion, if one views the feature mappings associated with the kernel as sufficient statistics in generalized (possibly infinite-dimensional) exponential families . [6]

Measuring dependence of random variables

A measure of the statistical dependence between random variables X {\displaystyle X} and Y {\displaystyle Y} (from any domains on which sensible kernels can be defined) can be formulated based on the Hilbert–Schmidt Independence Criterion [17]

and can be used as a principled replacement for mutual information , Pearson correlation or any other dependence measure used in learning algorithms. Most notably, HSIC can detect arbitrary dependencies (when a characteristic kernel is used in the embeddings, HSIC is zero if and only if the variables are independent), and can be used to measure dependence between different types of data (e.g. images and text captions). Given n i.i.d. samples of each random variable, a simple parameter-free unbiased estimator of HSIC which exhibits concentration about the true value can be computed in O (n (d f 2 + d g 2)) {\displaystyle O(n(d_{f}^{2}+d_{g}^{2}))} time, [6] where the Gram matrices of the two datasets are approximated using A A T , B B T {\displaystyle \mathbf {A} \mathbf {A} ^{T},\mathbf {B} \mathbf {B} ^{T}} with A ∈ R n × d f , B ∈ R n × d g {\displaystyle \mathbf {A} \in \mathbb {R} ^{n\times d_{f}},\mathbf {B} \in \mathbb {R} ^{n\times d_{g}}} . The desirable properties of HSIC have led to the formulation of numerous algorithms which utilize this dependence measure for a variety of common machine learning tasks such as: feature selection (BAHSIC [18]), clustering (CLUHSIC [19]), and dimensionality reduction (MUHSIC [20]).

HSIC can be extended to measure the dependence of multiple random variables. The question of when HSIC captures independence in this case has recently been studied: [21] for

more than two variables

on R d {\displaystyle \mathbb {R} ^{d}} : the characteristic property of the individual kernels remains an equivalent condition.

on general domains: the characteristic property of the kernel components is necessary but not sufficient .

Kernel belief propagation

Belief propagation is a fundamental algorithm for inference in graphical models in which nodes repeatedly pass and receive messages corresponding to the evaluation of conditional expectations. In the kernel embedding framework, the messages may be represented as RKHS functions and the conditional distribution embeddings can be applied to efficiently compute message updates. Given n samples of random variables represented by nodes in a Markov random field , the incoming message to node t from node u can be expressed as

if it assumed to lie in the RKHS. The kernel belief propagation update message from t to node s is then given by [2]

where ■ {\displaystyle \odot } denotes the element-wise vector product, N (t) ■ s {\displaystyle N(t)\backslash s} is the set of nodes connected to t excluding node s , β u t = (β u t 1 , … , β u t n) {\displaystyle {\boldsymbol {\beta }}_{ut}=\left(\beta _{ut}^{1},\dots ,\beta _{ut}^{n}\right)} , K t , K s {\displaystyle \mathbf {K} _{t},\mathbf {K} _{s}} are the Gram matrices of the samples from variables X t , X s {\displaystyle X_{t},X_{s}} , respectively, and Υ s = (ϕ (x s 1) , … , ϕ (x s n)) {\displaystyle {\boldsymbol {\Upsilon }}_{s}=\left(\varphi (x_{s}^{1}),\dots ,\varphi (x_{s}^{n})\right)} is the feature matrix for the samples from X s {\displaystyle X_{s}} .

Thus, if the incoming messages to node t are linear combinations of feature mapped samples from X t {\displaystyle X_{t}} , then the outgoing message from this node is also a linear combination of feature mapped samples from X s {\displaystyle X_{s}} . This RKHS function representation of message-passing updates therefore produces an efficient belief propagation algorithm in which the potentials are nonparametric functions inferred from the data so that arbitrary statistical relationships may be modeled. [2]

Nonparametric filtering in hidden Markov models

In the hidden Markov model (HMM), two key quantities of interest are the transition probabilities between hidden states P (S t ■ S t – 1) {\displaystyle P(S^{t}\mid S^{t-1})} and the emission probabilities P (O t ■ S t) {\displaystyle P(O^{t}\mid S^{t})} for observations. Using the kernel conditional distribution embedding framework, these quantities may be expressed in terms of samples from the HMM. A serious limitation of the embedding methods in this domain is the need for training samples containing hidden states, as otherwise inference with arbitrary distributions in the HMM is not possible.

One common use of HMMs is filtering in which the goal is to estimate posterior distribution over the hidden state s t {\displaystyle s^{t}} at time step t given a history of previous observations h t = (o 1 , … , o t) {\displaystyle h^{t}=(o^{1},\dots ,o^{t})} from the system. In filtering, a belief state P (S t + 1 ■ h t + 1) {\displaystyle P(S^{t+1}\mid h^{t+1})} is recursively maintained via a prediction step (where updates P (S t + 1 ■ h t) = E [P (S t + 1 ■ S t) ■ h t] {\displaystyle P(S^{t+1}\mid h^{t})=\mathbb {E} [P(S^{t+1}\mid S^{t})\mid h^{t}]} are computed by marginalizing out the previous hidden state) followed by a conditioning step (where updates P (S t + 1 ■ h t , o t + 1) ∝ P (o t + 1 ■ S t + 1) P (S t + 1 ■ h t) {\displaystyle P(S^{t+1}\mid h^{t},o^{t+1})\propto P(o^{t+1}\mid S^{t+1})P(S^{t+1}\mid h^{t})} are computed by applying Bayes' rule to condition on a new observation). [2] The RKHS embedding of the belief state at time t+1 can be recursively expressed as

by computing the embeddings of the prediction step via the kernel sum rule and the embedding of the conditioning step via kernel Bayes' rule . Assuming a training sample (s ~ 1 , … , s ~ T , o ~ 1 , … , o ~ T) {\displaystyle ({\widetilde {s}}^{1},\dots ,{\widetilde {s}}^{T},{\widetilde {o}}^{1},\dots ,{\widetilde {o}}^{T})} is given, one can in practice estimate

and filtering with kernel embeddings is thus implemented recursively using the following updates for the weights α = (α 1 , … , α T) {\displaystyle {\boldsymbol {\alpha }}=(\alpha _{1},\dots ,\alpha

_{T})} [2]

where G , K {\displaystyle \mathbf {G} ,\mathbf {K} } denote the Gram matrices of s ~ 1 , … , s ~ T {\displaystyle {\widetilde {s}}^{1},\dots ,{\widetilde {s}}^{T}} and o ~ 1 , … , o ~ T {\displaystyle {\widetilde {o}}^{1},\dots ,{\widetilde {o}}^{T}} respectively, G ~ {\displaystyle {\widetilde {\mathbf {G} }}} is a transfer Gram matrix defined as G ~ i j = k (s ~ i , s ~ j + 1) , {\displaystyle {\widetilde {\mathbf {G} }}_{ij}=k({\widetilde {s}}_{i},{\widetilde {s}}_{j+1}),} and K o t + 1 = (k (o ~ 1 , o t + 1) , … , k (o ~ T , o t + 1)) T . {\displaystyle \mathbf {K} _{o^{t+1}}=(k({\widetilde {o}}^{1},o^{t+1}),\dots ,k({\widetilde {o}}^{T},o^{t+1}))^{T}.}

Support measure machines

The support measure machine (SMM) is a generalization of the support vector machine (SVM) in which the training examples are probability distributions paired with labels { P i , y i } i = 1 n , y i ∈ { + 1 , − 1 } {\displaystyle \{P_{i},y_{i}\}_{i=1}^{n},\ y_{i}\in \{+1,-1\}} . [22] SMMs solve the standard SVM dual optimization problem using the following expected kernel

which is computable in closed form for many common specific distributions P i {\displaystyle P_{i}} (such as the Gaussian distribution) combined with popular embedding kernels k {\displaystyle k} (e.g. the Gaussian kernel or polynomial kernel), or can be accurately empirically estimated from i.i.d. samples { x i } i = 1 n ∼ P (X) , { z j } j = 1 m ∼ Q (Z) {\displaystyle \{x_{i}\}_{i=1}^{n}\sim P(X),\{z_{j}\}_{j=1}^{m}\sim Q(Z)} via

Under certain choices of the embedding kernel k {\displaystyle k} , the SMM applied to training examples { P i , y i } i = 1 n {\displaystyle \{P_{i},y_{i}\}_{i=1}^{n}} is equivalent to a SVM trained on samples { x i , y i } i = 1 n {\displaystyle \{x_{i},y_{i}\}_{i=1}^{n}} , and thus the SMM can be viewed as a flexible SVM in which a different data-dependent kernel (specified by the assumed form of the distribution P i {\displaystyle P_{i}}) may be placed on each training point. [22]

Domain adaptation under covariate, target, and conditional shift

The goal of domain adaptation is the formulation of learning algorithms which generalize well when the training and test data have different distributions. Given training examples { (x i tr , y i tr) } i = 1 n {\displaystyle \{(x_{i}^{\text{tr}},y_{i}^{\text{tr}})\}_{i=1}^{n}} and a test set { (x j te , y j te) } j = 1 m {\displaystyle \{(x_{j}^{\text{te}},y_{j}^{\text{te}})\}_{j=1}^{m}} where the y j te {\displaystyle y_{j}^{\text{te}}} are unknown, three types of differences are commonly assumed between the distribution of the training examples P tr (X , Y) {\displaystyle P^{\text{tr}}(X,Y)} and the test distribution P te (X , Y) {\displaystyle P^{\text{te}}(X,Y)} : [23] [24]

Covariate shift in which the marginal distribution of the covariates changes across domains: P tr (X) ≠ P te (X) {\displaystyle P^{\text{tr}}(X)\neq P^{\text{te}}(X)}

Target shift in which the marginal distribution of the outputs changes across domains: P tr (Y) ≠ P te (Y) {\displaystyle P^{\text{tr}}(Y)\neq P^{\text{te}}(Y)}

Conditional shift in which P (Y) {\displaystyle P(Y)} remains the same across domains, but the conditional distributions differ: P tr (X ■ Y) ≠ P te (X ■ Y) {\displaystyle P^{\text{tr}}(X\mid Y)\neq P^{\text{te}}(X\mid Y)} . In general, the presence of conditional shift leads to an ill-posed problem, and the additional assumption that P (X ■ Y) {\displaystyle P(X\mid Y)} changes only under location - scale (LS) transformations on X {\displaystyle X} is commonly imposed to make the problem tractable.

By utilizing the kernel embedding of marginal and conditional distributions, practical approaches to deal with the presence of these types of differences between training and test domains can be formulated. Covariate shift may be accounted for by reweighting examples via estimates of the ratio P te (X) / P tr (X) {\displaystyle P^{\text{te}}(X)/P^{\text{tr}}(X)} obtained directly from the kernel embeddings of the marginal distributions of X {\displaystyle X} in each domain without any need for explicit estimation of the distributions. [24] Target shift, which cannot be similarly dealt with since no samples from Y {\displaystyle Y} are available in the test domain, is accounted for by weighting training examples using the vector β ∗ (y tr) {\displaystyle {\boldsymbol {\beta }}^{*}(\mathbf {y} ^{\text{tr}})} which solves the following optimization problem (where in practice, empirical approximations must be used) [23]

To deal with location scale conditional shift, one can perform a LS transformation of the training points to obtain new transformed training data X new = X tr ■ W + B {\displaystyle \mathbf {X} ^{\text{new}}=\mathbf {X} ^{\text{tr}}\odot \mathbf {W} +\mathbf {B} } (where ■ {\displaystyle \odot } denotes the element-wise vector product). To ensure similar distributions between the new transformed training samples and the test data, W , B {\displaystyle \mathbf {W} ,\mathbf {B} } are estimated by minimizing the following empirical kernel embedding distance [23]

In general, the kernel embedding methods for dealing with LS conditional shift and target shift may be combined to find a reweighted transformation of the training data which mimics the test distribution, and these methods may perform well even in the presence of conditional shifts other than location-scale changes. [23]

Domain generalization via invariant feature representation

Given N sets of training examples sampled i.i.d. from distributions P (1) (X , Y) , P (2) (X , Y) , … , P (N) (X , Y) {\displaystyle P^{(1)}(X,Y),P^{(2)}(X,Y),\ldots ,P^{(N)}(X,Y)} , the goal of domain generalization is to formulate learning algorithms which perform well on test examples sampled from a previously unseen domain P ∗ (X , Y) {\displaystyle P^{*}(X,Y)} where no data from the test domain is available at training time. If conditional distributions P (Y ■ X) {\displaystyle P(Y\mid X)} are assumed to be relatively similar across all domains, then a learner capable of domain generalization must estimate a functional relationship between the variables which is robust to changes in the marginals P (X) {\displaystyle P(X)} . Based on kernel embeddings of these distributions, Domain Invariant Component Analysis (DICA) is a method which determines the transformation of the training data that minimizes the difference between marginal distributions while preserving a common conditional distribution shared between all training domains. [25] DICA thus extracts invariants , features that transfer across domains, and may be viewed as a generalization of many popular dimension-reduction methods such as kernel principal component analysis , transfer component analysis, and covariance operator inverse regression. [25]

Defining a probability distribution P {\displaystyle {\mathcal {P}}} on the RKHS H {\displaystyle {\mathcal {H}}} with

DICA measures dissimilarity between domains via distributional variance which is computed as

where

so G {\displaystyle \mathbf {G} } is a N × N {\displaystyle N\times N} Gram matrix over the distributions from which the training data are sampled. Finding an orthogonal transform onto a low-dimensional subspace B (in the feature space) which minimizes the distributional variance, DICA simultaneously ensures that B aligns with the bases of a central subspace C for which Y {\displaystyle Y} becomes independent of X {\displaystyle X} given C T X {\displaystyle C^{T}X} across all domains. In the absence of target values Y {\displaystyle Y} , an unsupervised version of DICA may be formulated which finds a low-dimensional subspace that minimizes distributional variance while simultaneously maximizing the variance of X {\displaystyle X} (in the feature space) across all domains (rather than preserving a central subspace). [25]

Distribution regression

In distribution regression, the goal is to regress from probability distributions to reals (or vectors). Many important machine learning and statistical tasks fit into this framework, including multi-instance learning , and point estimation problems without analytical solution (such as hyperparameter or entropy estimation). In practice only samples from sampled distributions are observable, and the estimates have to rely on similarities computed between sets of points . Distribution regression has been successfully applied for example in supervised entropy learning, and aerosol prediction using multispectral satellite images. [26]

Given ({ X i , n } n = 1 N i , y i) i = 1 ■ {\displaystyle {\left(\{X_{i,n}\}_{n=1}^{N_{i}},y_{i}\right)}_{i=1}^{\ell }} training data, where the X i ^ := { X i , n } n = 1 N i {\displaystyle {\hat {X_{i}}}:=\{X_{i,n}\}_{n=1}^{N_{i}}} bag contains samples from a probability distribution X i {\displaystyle X_{i}} and the i th {\displaystyle i^{\text{th}}} output label is y i ∈ R {\displaystyle y_{i}\in \mathbb {R} } , one can tackle the distribution regression task by taking the embeddings of the distributions, and learning the regressor from the embeddings to the outputs. In

other words, one can consider the following kernel ridge regression problem (λ > 0) {\displaystyle (\lambda >0)}

where

with a k {\displaystyle k} kernel on the domain of X i {\displaystyle X_{i}} -s (k : Ω × Ω → R) {\displaystyle (k:\Omega \times \Omega \to \mathbb {R})} , K {\displaystyle K} is a kernel on the embedded distributions, and H (K) {\displaystyle {\mathcal {H}}(K)} is the RKHS determined by K {\displaystyle K} . Examples for K {\displaystyle K} include the linear kernel [K (μ P , μ Q) = ■ μ P , μ Q ■ H (k)] {\displaystyle \left[K(\mu _{P},\mu _{Q})=\langle \mu _{P},\mu _{Q}\rangle _{{\mathcal {H}}(k)}\right]} , the Gaussian kernel [K (μ P , μ Q) = e − ■ μ P − μ Q ■ H (k) 2 / (2 σ 2)] {\displaystyle \left[K(\mu _{P},\mu _{Q})=e^{-\left\|\mu _{P}-\mu _{Q}\right\|_{H(k)}^{2}/(2\sigma ^{2})}\right]} , the exponential kernel [K (μ P , μ Q) = e − ■ μ P − μ Q ■ H (k) / (2 σ 2)] {\displaystyle \left[K(\mu _{P},\mu _{Q})=e^{-\left\|\mu _{P}-\mu _{Q}\right\|_{H(k)}/(2\sigma ^{2})}\right]} , the Cauchy kernel [K (μ P , μ Q) = (1 + ■ μ P − μ Q ■ H (k) 2 / σ 2) − 1] {\displaystyle \left[K(\mu _{P},\mu _{Q})=\left(1+\left\|\mu _{P}-\mu _{Q}\right\|_{H(k)}^{2}/\sigma ^{2}\right)^{-1}\right]} , the generalized t-student kernel [K (μ P , μ Q) = (1 + ■ μ P − μ Q ■ H (k) σ) − 1 , (σ ≤ 2)] {\displaystyle \left[K(\mu _{P},\mu _{Q})=\left(1+\left\|\mu _{P}-\mu _{Q}\right\|_{H(k)}^{\sigma }\right)^{-1},(\sigma \leq 2)\right]} , or the inverse multiquadrics kernel [K (μ P , μ Q) = (■ μ P − μ Q ■ H (k) 2 + σ 2) − 1 2] {\displaystyle \left[K(\mu _{P},\mu _{Q})=\left(\left\|\mu _{P}-\mu _{Q}\right\|_{H(k)}^{2}+\sigma ^{2}\right)^{-{\frac {1}{2}}}\right]} .

The prediction on a new distribution (X ^) {\displaystyle ({\hat {X}})} takes the simple, analytical form

where k = [K (μ X ^ i , μ X ^)] ∈ R 1 × ■ {\displaystyle \mathbf {k} =\big []K\big (\mu _{{\hat {X}}_{i}},\mu _{\hat {X}}\big)\big]\in \mathbb {R} ^{1\times \ell }} , G = [G i j] ∈ R ■ × ■ {\displaystyle \mathbf {G} =[G_{ij}]\in \mathbb {R} ^{\ell \times \ell }} , G i j = K (μ X ^ i , μ X ^ j) ∈ R {\displaystyle G_{ij}=K\big (\mu _{{\hat {X}}_{i}},\mu _{{\hat {X}}_{j}}\big)\in \mathbb {R} } , y = [y 1 ; … ; y ■] ∈ R ■ {\displaystyle \mathbf {y} =[y_{1};\ldots ;y_{\ell }]\in \mathbb {R} ^{\ell }} . Under mild regularity conditions this estimator can be shown to be consistent and it can achieve the one-stage sampled (as if one had access to the true X i {\displaystyle X_{i}} -s) minimax optimal rate. [26] In the J {\displaystyle J} objective function y i {\displaystyle y_{i}} -s are real numbers; the results can also be extended to the case when y i {\displaystyle y_{i}} -s are d {\displaystyle d} -dimensional vectors, or more generally elements of a separable Hilbert space using operator-valued K {\displaystyle K} kernels.

Example

In this simple example, which is taken from Song et al., [2] X , Y {\displaystyle X,Y} are assumed to be discrete random variables which take values in the set { 1 , … , K } {\displaystyle \{1,\ldots ,K\}} and the kernel is chosen to be the Kronecker delta function, so k (x , x ′) = δ (x , x ′) {\displaystyle k(x,x')=\delta (x,x')} . The feature map corresponding to this kernel is the standard basis vector ϕ (x) = e x {\displaystyle \varphi (x)=\mathbf {e} _{x}} . The kernel embeddings of such a distributions are thus vectors of marginal probabilities while the embeddings of joint distributions in this setting are K × K {\displaystyle K\times K} matrices specifying joint probability tables, and the explicit form of these embeddings is

When P (X = s) > 0 {\displaystyle P(X=s)>0} , for all s ∈ { 1 , … , K } {\displaystyle s\in \{1,\ldots ,K\}} , the conditional distribution embedding operator,

is in this setting a conditional probability table

and

Thus, the embeddings of the conditional distribution under a fixed value of X {\displaystyle X} may be computed as

In this discrete-valued setting with the Kronecker delta kernel, the kernel sum rule becomes

The kernel chain rule in this case is given by

References

External links

Information Theoretical Estimators toolbox (distribution regression demonstration).