The Leave-one-out Kernel

Koji Tsuda[1] and Motoaki Kawanabe[2]

[1] AIST CBRC, 2-41-6, Aomi, Koto-ku, Tokyo, 1350064 , Japan
[2] Fraunhofer FIRST, Kekuléstr. 7, 12489 Berlin, Germany
koji.tsuda@aist.go.jp, nabe@first.fraunhofer.de

Abstract. Recently, several attempts have been made for deriving *data-dependent* kernels from distribution estimates with parametric models (e.g. the Fisher kernel). In this paper, we propose a new kernel derived from any distribution estimators, parametric or nonparametric. This kernel is called the Leave-one-out kernel (i.e. LOO kernel), because the leave-one-out process plays an important role to compute this kernel. We will show that, when applied to a parametric model, the LOO kernel converges to the Fisher kernel asymptotically as the number of samples goes to infinity.

1 Introduction

In kernel-based learning algorithms[6], a kernel function has to be defined a priori. Most algorithms require the kernel function to be positive semidefinite, and such kernels are often called "Mercer kernels"[9]. The design of Mercer kernels is an important topic in the study of kernel-based learning algorithms[6]. Here, one major direction is to derive a kernel function based on the estimated input distribution (e.g. [4, 8, 7]). The Fisher kernel[4] is constructed by a distribution estimate with a parametric model, and is applied to many tasks successfully, e.g. protein classification[4]. One important contribution of the Fisher kernel is that it enables to apply the kernel machines to discrete data such as sequences of different lengths or graphs, which used to be hard to deal with.
When the parametric model is not known for given data, nonparametric distribution estimators are often used[2]. Typical methods are e.g. kernel density estimators, k-nearest neighbor methods, orthogonal series estimators and so on. However, in principle, the Fisher kernel method cannot be applied for nonparametric estimators.
In this paper, we propose a general method to derive a Mercer kernel from any distribution estimates, parametric or nonparametric. The leave-one-out process plays an important role for obtaining this kernel. In order to compute the kernel function between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, we consider the leave-one-out density estimates $\hat{p}^{(i)}$ and $\hat{p}^{(j)}$, where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are left out, respectively. Then, the Hellinger inner product between $\hat{p}^{(i)}$ and $\hat{p}^{(j)}$ in the space of probability distributions[1] is taken as the kernel function, which is called the *Leave-one-out kernel* (LOO kernel). By constructing the correspondence between $\boldsymbol{x}_i$ and the leave-one-out estimate $\hat{p}^{(i)}$, the inner product in the space of probability distributions is imported as a kernel function in the input space $\mathcal{X}$. Intuitively, the LOO kernel reflects

the similarity of the *influences* on the density estimate when samples are left out.
When a parametric estimator is used, we will show that the LOO kernel converges to the Fisher kernel in probability as n goes to infinity. This fact shows that the Fisher kernel can be understood in terms of the influences when samples are left out. This view provides us a clear intuition about what the Fisher kernel actually measures.

2 The Leave-one-out kernel

To begin with, let us describe notations. Let $\mathcal{X}$ be a set of all possible inputs. $\mathcal{X}$ may be a finite set or an infinite set like $\mathbb{R}^d$. Also, let $\mathcal{P}$ be a space of all probability distributions defined on $\mathcal{X}$.

$$\mathcal{P} = \{p | p(\boldsymbol{x}) > 0, \forall \boldsymbol{x} \in \mathcal{X}; \int p(\boldsymbol{x}) d\boldsymbol{x} = 1\}.$$

We introduce the Hellinger inner product between the probability distributions:

$$\langle p, q \rangle = \int (\sqrt{p(\boldsymbol{x})} - \sqrt{p_0(\boldsymbol{x})})(\sqrt{q(\boldsymbol{x})} - \sqrt{p_0(\boldsymbol{x})}) d\boldsymbol{x}, \tag{1}$$

where $p_0 \in \mathcal{P}$ is the origin determined arbitrarily. The norm induced in this space is the Hellinger distance[1]. Let us define $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\} \in \mathcal{X}^n$ as the set of training samples. A distribution estimator, parametric or nonparametric, is described as a mapping $\mathcal{X}^n \to \mathcal{P}$.
Let $\hat{p} \in \mathcal{P}$ be a density estimate from the training samples. Also, let $\hat{p}^{(k)}$ be a leave-one-out density estimate where $\boldsymbol{x}_k$ is left out. Now, the LOO kernel, which we propose, is defined as the inner product between $\hat{p}^{(i)}$ and $\hat{p}^{(j)}$ where the origin is placed on $\hat{p}$:

$$K_\ell(\boldsymbol{x}_i, \boldsymbol{x}_j) = 4(n-1)^2 \int (\sqrt{\hat{p}^{(i)}(\boldsymbol{x})} - \sqrt{\hat{p}(\boldsymbol{x})})(\sqrt{\hat{p}^{(j)}(\boldsymbol{x})} - \sqrt{\hat{p}(\boldsymbol{x})}) d\boldsymbol{x}. \tag{2}$$

The normalizing factor $4(n-1)^2$ is necessary for the consistency with the Fisher kernel, which will be shown in the next section. Here, the inner product in $\mathcal{P}$ is imported to $\mathcal{X}$ through the distribution estimation operator. The value of a LOO kernel is related to the *influence* to the estimated distribution $\hat{p}$ when a sample is left out. The value becomes large when the LOO distributions $\hat{p}^{(i)}$ and $\hat{p}^{(j)}$ move a lot in the same direction. When no influence is induced by removing $\boldsymbol{x}_i$ or $\boldsymbol{x}_j$, the LOO kernel becomes zero, and the two samples are regarded as orthogonal in the feature space.

The LOO kernel from k-nearest neighbor As a example, the LOO kernel is derived from the k-nearest neighbor density estimate[2]. Assume $\mathcal{X} = \Re^d$. Let $D_k(\boldsymbol{x})$ denote the Euclidean distance from $\boldsymbol{x}$ to its k-nearest neighbor and c_d is the volume of unit sphere in d-dimensional space. The k-nn density estimate is obtained as

$$\hat{p}(\boldsymbol{x}) = \frac{k/n}{\mathrm{vol}_k(\boldsymbol{x})} \tag{3}$$

where $\mathrm{vol}_k(\boldsymbol{x}) = c_d(D_k(\boldsymbol{x}))^d$. The LOO kernel for the k-nn density estimate is straightforwardly obtained by subsuituting (3) to (2). However, since the numerical calculation of the integral in (2) is difficult, $\hat{p}(\boldsymbol{x})$ is approximated by the empirical distribution as

$$K_\ell(\boldsymbol{x}_i, \boldsymbol{x}_j) \approx \frac{4(n-1)^2}{n} \sum_{i=1}^{n} \frac{(\sqrt{\hat{p}^{(i)}(\boldsymbol{x}_i)} - \sqrt{\hat{p}(\boldsymbol{x}_i)})(\sqrt{\hat{p}^{(j)}(\boldsymbol{x}_i)} - \sqrt{\hat{p}(\boldsymbol{x}_i)})}{\hat{p}(\boldsymbol{x}_i)}.$$

3 The LOO kernel for parametric methods

In this section, we will show that, when the parametric model is available and the parameters are obtained from the maximum likelihood method, the Fisher kernel is an asymptotic approximation of the LOO kernel.

The Fisher kernel In order to define the Fisher kernel, one needs to have a parametric model $p(\boldsymbol{x}|\boldsymbol{\theta}), \boldsymbol{\theta} \in \mathbb{R}^p$ defined on $\mathcal{X}$[4]. Also, one needs a parameter estimate $\hat{\boldsymbol{\theta}}$ obtained from training data by some method, e.g. the maximum likelihood method. Then the Fisher kernel is denoted as $K_f(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{u}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})^\top G(\hat{\boldsymbol{\theta}})^{-1} \boldsymbol{u}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ where $\boldsymbol{u}$ and G are the score function and the Fisher information matrix, respectively: $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{x}|\boldsymbol{\theta}), G(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})^\top]$, where $\nabla_{\boldsymbol{\theta}}$ denote the gradient vector. The Fisher kernel is successfully applied to many application areas such as protein classification[3].

Connection to the Fisher kernel Here we derive the LOO kernel from a parametric model $p(\boldsymbol{x}|\boldsymbol{\theta})$. Let $\hat{\boldsymbol{\theta}}$ be the maximum likelihood solution from n training samples, and $\hat{\boldsymbol{\theta}}^{(i)}$ be the maximum likelihood solution where the i-th sample $\boldsymbol{x}_i$ is left out. Then, the LOO kernel is described as follows: $K_\ell(\boldsymbol{x}_i, \boldsymbol{x}_j) = 4(n-1)^2 K_H(\boldsymbol{x}_i, \boldsymbol{x}_j)$, where

$$K_H(\boldsymbol{x}_i, \boldsymbol{x}_j) = \int (p^{1/2}(\boldsymbol{x}|\hat{\boldsymbol{\theta}}^{(i)}) - p^{1/2}(\boldsymbol{x}|\hat{\boldsymbol{\theta}}))(p^{1/2}(\boldsymbol{x}|\hat{\boldsymbol{\theta}}^{(j)}) - p^{1/2}(\boldsymbol{x}|\hat{\boldsymbol{\theta}}))d\boldsymbol{x}.$$

Theorem 1. *In the limit $n \to \infty$, $K_\ell(\boldsymbol{x}_i, \boldsymbol{x}_j)$ converges to $K_f(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in probability.*

(Proof) Because $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^{(k)}$ are maximum likelihood estimators, we have the following equations: $\sum_{i=1}^n \boldsymbol{u}(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}) = \boldsymbol{0}$, $\sum_{i \neq k} \boldsymbol{u}(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^{(k)}) = \boldsymbol{0}$. By the Taylor expansion of the latter equation, we have

$$\boldsymbol{0} = \sum_{i \neq k} \boldsymbol{u}(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) + (n-1)H(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{(k)} - \hat{\boldsymbol{\theta}}) + o_p(1), \tag{4}$$

where $H_{ab}(\hat{\boldsymbol{\theta}}) = \frac{1}{n-1} \sum_{i \neq k} \partial_{\theta_a} \partial_{\theta_b} \log p(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}})$. For notational convenience, $\partial_{\theta_a} := \partial / \partial \theta_a$. It is verified that

$$H_{ab}(\hat{\boldsymbol{\theta}}) = -G_{ab}(\hat{\boldsymbol{\theta}}) + o_p(1), \tag{5}$$

because we have $H_{ab}(\hat{\boldsymbol{\theta}}) = \int \partial_{\theta_a}\partial_{\theta_b} \log p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) d\boldsymbol{x} + o_p(1)$ from the law of large numbers and its first term is rewritten as follows:

$$\begin{aligned}\int \partial_{\theta_a}\partial_{\theta_b} \log p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) d\boldsymbol{x} &= \int \partial_{\theta_a}\partial_{\theta_b} p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) d\boldsymbol{x} - \int \frac{\partial_{\theta_a} p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) \partial_{\theta_b} p(\boldsymbol{x}|\hat{\boldsymbol{\theta}})}{p(\boldsymbol{x}|\hat{\boldsymbol{\theta}})} d\boldsymbol{x} \\ &= -\int \partial_{\theta_a} \log p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) \partial_{\theta_b} \log p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) d\boldsymbol{x} \\ &= -G_{ab}(\hat{\boldsymbol{\theta}}).\end{aligned}$$

The first term of (4) is described as

$$\sum_{i \neq k} \boldsymbol{u}(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) = -\boldsymbol{u}(\boldsymbol{x}_k, \hat{\boldsymbol{\theta}}). \tag{6}$$

By substituting (5) and (6) into (4), we have

$$\hat{\boldsymbol{\theta}}^{(k)} - \hat{\boldsymbol{\theta}} = \frac{-1}{n-1} G(\hat{\boldsymbol{\theta}})^{-1} \boldsymbol{u}(\boldsymbol{x}_k|\hat{\boldsymbol{\theta}}) + o_p(n^{-1}). \tag{7}$$

By the Taylor expansion of $p^{1/2}(\boldsymbol{x}|\hat{\boldsymbol{\theta}}^{(k)})$ around $\hat{\boldsymbol{\theta}}$, we have the following:

$$p^{1/2}(\boldsymbol{x}|\hat{\boldsymbol{\theta}}^{(k)}) - p^{1/2}(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) = \frac{\nabla_{\boldsymbol{\theta}} p(\boldsymbol{x}|\hat{\boldsymbol{\theta}})}{2p^{1/2}(\boldsymbol{x}|\hat{\boldsymbol{\theta}})} (\hat{\boldsymbol{\theta}}^{(k)} - \hat{\boldsymbol{\theta}}) + o_p(n^{-1}). \tag{8}$$

By substituting (8) into K_H,

$$\begin{aligned}K_H(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \frac{1}{4}(\hat{\boldsymbol{\theta}}^{(i)} - \hat{\boldsymbol{\theta}})^\top \left[\int \left(\frac{\nabla_{\boldsymbol{\theta}} p(\boldsymbol{x}|\hat{\boldsymbol{\theta}})}{p(\boldsymbol{x}|\hat{\boldsymbol{\theta}})}\right)\left(\frac{\nabla_{\boldsymbol{\theta}} p(\boldsymbol{x}|\hat{\boldsymbol{\theta}})}{p(\boldsymbol{x}|\hat{\boldsymbol{\theta}})}\right)^\top p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) d\boldsymbol{x}\right](\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}) \\ &= \frac{1}{4}(\hat{\boldsymbol{\theta}}^{(i)} - \hat{\boldsymbol{\theta}})^\top G(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}) + o_p(n^{-2}).\end{aligned} \tag{9}$$

By substituting (7) into (9), we have

$$K_H(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1}{4(n-1)^2} \boldsymbol{u}(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})^\top G^{-1}(\hat{\boldsymbol{\theta}}) \boldsymbol{u}(\boldsymbol{x}_j, \hat{\boldsymbol{\theta}}) + o_p(n^{-2}).$$

4 Simulations

First of all, we will show the simulation result with the one-dimensional Gaussian $p(x|\mu, \eta) = \frac{\eta}{\sqrt{2\pi}} \exp(-\frac{\eta^2 (x-\mu)^2}{2})$, where μ and η are the mean and the inverse of standard deviation. Fig.1 shows an example of the kernel matrices of two methods between 30 samples drawn from the Gaussian. The difference is so small that they look almost the same. It shows that, in such a low dimensional problem, 30 samples are enough for asympotic approximation.

Secondly, we perform a clustering experiment. In most conventional partitional clustering methods[5] utilize the Euclidean or Mahalanobis distances, thus a cluster is assumed as spherical or ellipsoidal shaped. However, when the cluster shape is not actually spherical (e.g. Fig.2), these methods will fail. So, a proper distance measure is needed for successful clustering. In [8, 7], the distance was derived from parametric density models, mostly the Gaussian mixtures, and excellent results were

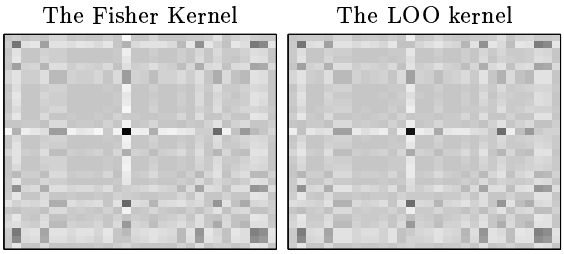

Fig. 1. The comparison between the two kernel matrices

obtained. Here, we will show a similar result by using a nonparametric density estimate, namely the k-nearest neighbor method. The two-dimensional artificial dataset is shown in Fig.2. We performed the k-NN density estimate (k=15), and the result is shown in the lower left of Fig.2. Then, the LOO kernel matrix is derived from this estimate. The lower right of Fig.2 shows the kernel matrix, where the clear separation of two clusters is already visible. When the k-means clustering is applied in the feature space, the two non-spherical clusters are nicely separated (Fig. 2 upper left), whereas the k-means in the input space cannot separate (Fig. 2 upper right). The advantage of this method over Gaussian mixture based methods[8, 7] is that the iterative EM learning is not necessary for computing kernels. Unlike the approaches based on the Gaussian mixture, our kernel does not have a local minima problem.

5 Conclusion

In this paper, we proposed the leave-one-out kernel, which is derived from parametric and nonparametric distribution estimates. Since the LOO kernel is approximated by the Fisher kernel in parametric cases, it can be considered as an extension of the Fisher kernel. The leave-one out process is commonly used for estimating generalization error (i.e. LOO cross validation). In future works, it would be interesting to explore the connections between the LOO kernel and the LOO cross validation. Notice that the LOO kernel is defined only among existing examples, not for unseen examples. So in supervised learning, we have to know the input vectors of test set in advance (i.e. the transductive setting). The generalization to unseen examples would be an interesting future topic as well.

Acknowledgements The authors would like to thank K.-R. Müller, G. Rätsch and S. Sonnenburg for fruitful discussions.

References

1. S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, 2001.

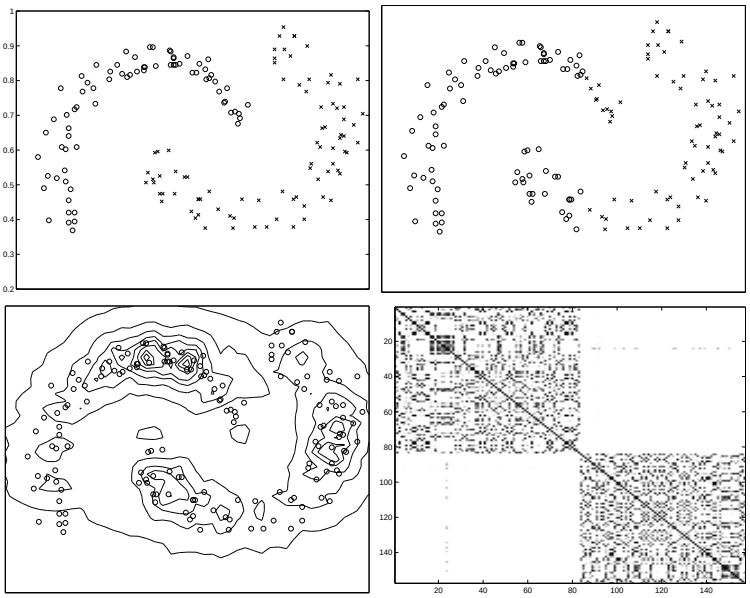

Fig. 2. (Upper Left) the result of k-means clustering in the LOO kernel's feature space derived from k-nn density estimate. (Upper Right) the result of k-means clustering in the input space. (Lower Left) Contour plot of density estimate by k-nn method ($k = 15$). (Lower Right) The plot of LOO kernel matrix. Clear separation of two clusters can be seen.

2. A.J. Izenman. Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86(413):205–224, 1991.
3. T.S. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. *J. Comp. Biol.*, 7:95–114, 2000.
4. T.S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS 11*, pages 487–493. MIT Press, 1999.
5. A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.
6. K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Networks*, 12(2):181–201, 2001.
7. M. Rattray. A model-based distance for clustering. In *Proc. IJCNN'00*, 2000.
8. M.E. Tipping. Deriving cluster analytic distance functions from gaussian mixture models. In D. Willshaw and A. Murray, editors, *Proceedings of ICANN'99*, pages 815–820. IEE Press, 1999.
9. V.N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.