Theoretical Foundations for Learning Kernels in Supervised Kernel PCA

Mehryar Mohri
Courant Institute and Google
251 Mercer Street
New York, NY 10012
mohri@cs.nyu.edu

Afshin Rostamizadeh
Google
76 Ninth Avenue
New York, NY 10011
rostami@google.com

Dmitry Storcheus
Google
76 Ninth Avenue
New York, NY 10011
ds3626@nyu.edu

Abstract

This paper presents a novel learning scenario which combines dimensionality reduction, supervised learning as well as kernel selection. We carefully define the hypothesis class that addresses this setting and provide an analysis of its Rademacher complexity and thereby provide generalization guarantees. The proposed algorithm uses KPCA to reduce the dimensionality of the feature space, i.e. by projecting data onto top eigenvectors of covariance operator in a kernel reproducing space. Moreover, it simultaneously learns a linear combination of base kernel functions, which defines a reproducing space, as well as the parameters of a supervised learning algorithm in order to minimize a regularized empirical loss. The bound on Rademacher complexity of our hypothesis is shown to be logarithmic in the number of base kernels, which encourages practitioners to combine as many base kernels as possible.

1 Introduction

In this paper we propose and analyze a hypothesis class for an algorithm that simultaneously learns a projection of data points onto a low-dimensional manifold as well as selects a discriminative function defined over the low-dimensional manifold. There are many well known techniques of non-linear manifold learning, such as Isometric Feature Mapping [11], Locally Linear Embedding [9] and Kernel Principal Component Analysis [10]. The setting suggested here is different from the standard dimensionality reduction setting in two ways: first we use KPCA in a supervised setting, i.e. coupled with a discriminative algorithm, and second we also learn the kernel function used by KPCA in a supervised manner.

Our hypothesis set is built around the KPCA algorithm, which learns a manifold by projecting onto top eigenvectors of a sample covariance operator in a reproducing kernel Hilbert space. We choose this particular dimensionality reduction technique, since existing literature has shown that several other manifold learning methods essentially reduce to KPCA, given the appropriate choice of kernel function.

It has been shown [6] that most of popular dimensionality reduction techniques are equivalent to Kernel PCA with a specific kernel matrix. Thus, choosing a dimensionality reduction method is equivalent to choosing a kernel for KPCA. However, one can come up with numerous different kernels to do dimensionality reduction, where every kernel could potentially be better than others for certain data sets and problems. The question of which kernel to use for a particular problem is not easy to answer and often involves trial and error. Our suggested algorithm improves upon this situation by considering a set of many base kernels, instead of only a single kernel, and learns a final kernel which is a non-negative linear combination of base kernels. Thus, one does not need to commit a priori to single kernel appropriate for the task, instead a kernel will be learned. We show that the Rademacher complexity of the resulting hypothesis set is logarithmic in the number of base kernels. Thus, one can potentially combine a very large number of kernels and still be guaranteed that our algorithm will no over-fit.

Our algorithm is supervised in the sense that it uses a training sample to learn an optimal weighted sum of kernels that gives "best" KPCA projection. Here, best means the smallest loss when the projected data is used as features for classification. Traditionally manifold learning was viewed as an unsupervised procedure, mostly used for exploratory data analysis or visual representation. However, it is shown in recent papers [4], [5], [8] that tuning manifold construction to directly benefit the classification algorithm used on with the reduced features gives considerably better performance. Treating the dimensionality reduction problem and the classification problem, which uses the reduced data, as a joint problem is usually called a "coupled" problem [5]. Thus, the novelty of this work is in analysing the learning kernel problem in the context of coupled dimensionality reduction and classification problem.

This short paper is focused on theoretical analysis of the algorithm: we provide a rigorous definition of the hypothesis set and its properties via operators in Hilbert spaces and derive a bound on the Rademacher complexity. We also present the form of the suggested algorithm, while investigating efficient implementations and empirical results are left for a longer version of this work.

2 Preliminaries

Since the algorithm we propose is learning a projection in reproducing Hilbert space, we will be explicitly using the algebra of operators on separable Hilbert spaces. Our notations are in line with [2], readers are encouraged to refer to that paper for more detailed description of relevant theorems.

Let $\{K_k\}_{k=1}^p$ be a set of base kernels and $\boldsymbol{\mu} \in \mathbb{R}^p$ be a vector with nonnegative coordinates. Let K be a weighted sum of base kernels $K = \sum\limits_{k=1}^{p} \mu_k K_k$. Denote $\mathbb{H}_{\mu_k K_K}$ the reproducing space of kernel function $\mu_k K_k$ and $\mathbb{H}$ the reproducing space of K. Assume that data is sampled from a manifold $\mathbb{X}$. Let $\Phi_{\mu_k K_k} : \mathbb{X} \to \mathbb{H}_{\mu_k K_K}$ be the feature map corresponding to kernel $\mu_k K_k$, in particular $\Phi_{\mu_k K_k}(x) = \mu_k K_k(\cdot, x)$ and Φ is the feature map corresponding to kernel K. By $\mathbb{H}^0_{\mu_k K_k}$ we denote the subspace of $\mathbb{H}_{\mu_k K_k}$ spanned by $\Phi_{K_k}(x'_1), ..., \Phi_{K_k}(x'_{m'})$. For the purpose of the numerical we only need information about $\mathbb{H}^0_{\mu_k K_k}$.

If data x is sampled from $\mathbb{X}$ according to some distribution $\mathcal{D}$, then $\Phi(x)$ is a random element in Hilbert space $\mathbb{H}$. Let C be the true covariance operator with respect to $\Phi(x)$ defined as $\langle f, Cg \rangle_{\mathbb{H}} = E\left\{\langle f, \Phi(x)\rangle_{\mathbb{H}} \langle g, \Phi(x)\rangle_{\mathbb{H}}\right\}$. Ideally we are interested in learning the projection onto the top r eigenfunctions of C. To estimate C we use an unlabeled sample $S' = \{x'_1, ..., x'_{m'}\}$ and define a sample covariance operator $C_{S'}$ by $\langle f, C_{S'} g \rangle_{\mathbb{H}} = \frac{1}{m'} \sum\limits_{n=1}^{m'} f(x_n) g(x_n)$. As shown by [12] both eigenvalues and eigenspaces of $C_{S'}$ converge to those of C. Thus our algorithm will learn a projection onto top r nonzero eigenfunctions of $C_{S'}$ within the space $\mathbb{H}$ generated by kernel K. We will denote this orthogonal rank r projection by $P^r_{S'}$.

Learning is done by fitting weights $\boldsymbol{\mu}$ of base kernels. We use a labeled sample $S = \{x_1, ..., x_m\}$ for learning. Since we are considering two samples S and S', we need to distinguish between kernel matrices, their eigenvalues and eigenvectors on different sample. Let $\mathbf{K}_k$ be the kernel matrix of K_k on sample S and $\mathbf{v}_{k,j}$ with $\gamma_{k,j}$ be its eigenvector(normalized)-eigenvalue pair, ordered by eigenvalues in decreasing order. The same objects built on unlabeled sample S' are $\mathbf{K}'_k$,$\mathbf{v}'_{k,j}$ and $\gamma'_{k,j}$ respectively.

We regularize $\boldsymbol{\mu}$ by bounding $\sup\limits_{|I|=r} \sum\limits_{(k,j)\in I} \mu_k \gamma'_{k,j} \leq \Lambda$, where $\mid I \mid = r$ means that the cardinality of index set I is r. To put it more rigorously, define a set $\Delta_{\gamma',\Lambda}$ as follows:

$$\Delta_{\gamma',\Lambda} = \left\{\boldsymbol{\mu} : \mu_k \geq 0 \wedge \sup_{|I|=r} \sum_{(k,j)\in I} \mu_k \gamma'_{k,j} \leq \Lambda \wedge \mu_k = 0 \textit{ if } \mu_k \notin \textit{ top r coordinates of } \left\{\mu_k \gamma'_{k,j}\right\}\right\}$$

Our algorithm learns projection by optimizing $\boldsymbol{\mu}$ subject to $\boldsymbol{\mu} \in \Delta_{\gamma',\Lambda}$. To ensure that the relation between $\boldsymbol{\mu}$ and eigenfunctions of $C_{S'}$ is explicit, we impose a mild assumption on base kernels, which we call orthogonality.

Definition 1. ***Orthogonal Kernels.*** *Let $\{K_k\}_{k=1}^p$ be a finite set of PDS kernels, then this set is called orthogonal with respect to sample $S = \{x_1, ...x_m\}$ if and only if $\mathbb{H}_i^0 \cap \mathbb{H}_j^0 = 0$ for any $i \neq j$. Where $\mathbb{H}_i^0$ and $\mathbb{H}_j^0$ are the subspaces of $\mathbb{H}_i$ and $\mathbb{H}_j$ spanned by $\Phi_{K_i}(x_1), ..., \Phi_{K_i}(x_m)$ and $\Phi_{K_j}(x_1), ..., \Phi_{K_j}(x_m)$ respectively.*

Orthogonality essentially means that reproducing spaces $\mathbb{H}_{\mu_k K_k}$ of base kernels restricted to span of sample points are disjoint, thus by [1], section 6, they are orthogonal components of $\mathbb{H}$. Orthogonality typically holds in practice, e.g. for polynomial and Gaussian kernels on $\mathbb{R}^n$. In case orthogonality is not satisfied, we can easily modify the support of base kernels to make this condition hold.

Let $C_{k,S'}$ be the restriction of $C_{S'}$ to $\mathbb{H}_{\mu_k K_k}$ and $u_{k,j}$ be the $j-th$ eigenfunction of $C_{k,S'}$ with eigenvalue $\lambda_{k,j}$. Orthogonality of kernels ensures that $u_{k,j}$ is also an eigenfunction of $C_{S'}$. Throughout the paper we will assume that base kernel functions $\{K_k\}$ are bounded and orthogonal as in Definition 1 with respect to samples S and S'

3 Learning scenario

Our formulation of the hypothesis, H_Λ, set is derived from [3], where they learn an optimal sum of kernels for SVM-style classification: $H_\Lambda = \{x \to \langle w, P_{S'}^r \Phi(x)\rangle_{\mathbb{H}} : K = \sum_{k=1}^p \mu_k K_k\}$ s.t. $\boldsymbol{\mu} \in \Delta_{\gamma',\Lambda}$, $\|w\|_{\mathbb{H}} \leqslant 1$ and $\{K_k\}_{k=1}^p$ are orthogonal. The notation H_Λ stresses that Λ is an important parameter of the hypothesis set, however one should keep in mind that H_Λ is also parametrized by the number of base kernels p, the rank of projection r and the unlaballed sample S'. In order to regularize $\boldsymbol{\mu}$ we control $\sup\limits_{|I|=r} \sum\limits_{(k,j)\in I} \mu_k \gamma'_{k,j}$, which is in fact a seminorm on $\mathbb{R}^{p,+}$, we denote it by $\|\boldsymbol{\mu}\|_{\gamma'}$. Thus, our algorithm learns $\boldsymbol{\mu}$ subject to $\|\boldsymbol{\mu}\|_{\gamma'} \leq \Lambda$. This norm has a direct connection to the spectrum of sample covariance operator $C_{S'}$, namely $\|\boldsymbol{\mu}\|_{\gamma'} = \frac{1}{m'} \sum\limits_{i=1}^{r} \lambda_i(C_{S'})$. Therefore, bounding $\|\boldsymbol{\mu}\|_{\gamma'}$ means bounding the spectrum of covariance operator.

The hypothesis set is described in terms of projection in $\mathbb{H}$, however for numerical computations only eigenvectors $\mathbf{v}'_{k,j}$ and eigenvalues $\gamma'_{k,j}$ of sample kernel matrices $\mathbf{K'}_k$ on sample S' are available. The following lemma shows how to compute the value of our hypothesis using the information about $\mathbf{K'}_k$.

Lemma 2. ***Computation of hypothesis.*** *For each $x \in \mathbb{X}$, every $h \in H_\Lambda$ is described as follows*

$$h(x) = \sum_{n=1}^{m}\sum_{i=1}^{m}\sum_{k=1}^{p} \frac{\sqrt{\mu_k}}{\sqrt{\gamma'_{k,i}}} \alpha_{k,i} \left[\mathbf{v}'_{k,i}\right]_n K_k(x'_n, x) s_{k,i} \tag{1}$$

s.t. $\sum\limits_{k=1}^{p}\sum\limits_{i=1}^{m} \alpha_{k,i}^2 s_{k,i} \leq 1$, $\boldsymbol{\mu} \in \Delta_{\gamma',\Lambda}$ as well as $s_{k,i} = 1$ if $\mu_k \gamma'_{k,i}$ belongs to top r from $\left\{\mu_k \gamma'_{k,i}\right\}$ and $s_{k,i} = 0$ otherwise. Here $\boldsymbol{\mu} \in \mathbb{R}^p$, $\alpha_{k,i} \in \mathbb{R}$ and $s_{k,i} \in \{1, 0\}$ are variables.

A heuristic algorithm naturally follows from the expression for $h(x)$ in the theorem above: minimize a convex loss function subject to $\|\boldsymbol{\mu}\|_{\gamma'} \leq \Lambda$

4 Generalization Bound

We derived a bound on sample Rademacher complexity of H_Λ, which is used together with the results of [7] to provide a generalization guarantee.

Theorem 3. ***Generalization bound*** *Let $\gamma'_{\min}$ be the smallest nonzero eigenvalue of $\{\mathbf{K'}_k\}$ and $\gamma_{\max}$ be the largest nonzero eigenvalue of $\{\mathbf{K}_k\}$. Let $\delta_r = \frac{1}{2}\left(\lambda_r(C) - \lambda_{r+1}(C)\right)$. Assume $\sup\limits_{x\in\mathbb{X}} K(x, x) =$*

M and denote the rank of $\mathbf{K}_k$ by r_k. Let $\rho > 0$ and $\hat{R}_\rho(h)$ be the margin loss of h. Then for any $\delta > 0$ with probability at least $1-\delta$ for any $h \in H_\Lambda$ the error $R(h)$ is bounded by

$$\hat{R}_\rho(h) + \frac{2}{\rho m}\sqrt{\frac{\gamma_{\max}}{\gamma'_{\min}} 2\Lambda \left(\log 2p + \log 2 \sum_{k=1}^{p} r_k\right)} \left(1 + \frac{4M}{\delta_r} + \frac{4M}{\delta_r}\sqrt{\frac{\log\frac{3}{\delta}}{2}}\right) + 3\sqrt{\frac{\log\frac{6}{\delta}}{2m}} \qquad (2)$$

The term $\frac{\gamma_{\max}}{\gamma'_{\min}}$ is a quantity that is similar to a condition number. We conjecture the dependency of generalization bound on $\frac{\gamma_{\max}}{\gamma'_{\min}}$ can be improved significantly in subsequent research. For the purposes of this discussion we treat the ratio as a constant. In the worst case, when kernel matrices have full rank, the sample Rademacher complexity is order of $O\left(\frac{\sqrt{\Lambda \log pm}}{m}\right)$. Since there is only a logarithmic dependency on number of base kernels p, our generalization bound encourages the use of large number of base kernels. Moreover, it suggests an algorithm for supervised Kernel PCA that consists of minimizing the empirical while controlling the upper bound Λ on semi-norm $\|\boldsymbol{\mu}\|_{\gamma'}$, which is equivalent to controlling the spectral radius of covariance operator.

5 Conclusion

In this paper we have defined a novel learning algorithm that combines nonlinear dimensionality reduction and classification. A rigorous learning scenario has been provided together with generalization bound based on Rademacher complexity. That bound tells us that if we use as many base kernels as we want, we still do not overfit due to logarithmic dependency on the number of kernels. The next step in this research is to analyze the empirical performance of the algorithm.

References

[1] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, pages 337–404, 1950.

[2] Gilles Blanchard, Olivier Bousquet, and Laurent Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2-3):259–294, 2007.

[3] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 247–254, 2010.

[4] Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *The Journal of Machine Learning Research*, 5:73–99, 2004.

[5] Mehmet Gönen. Coupled dimensionality reduction and classification for supervised and semi-supervised multilabel learning. *Pattern recognition letters*, 38:132–141, 2014.

[6] Jihun Ham, Daniel D Lee, Sebastian Mika, and Bernhard Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the twenty-first international conference on Machine learning*, page 47. ACM, 2004.

[7] Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II*, pages 443–457. Springer, 2000.

[8] Yen-Yu Lin, Tyng-Luh Liu, and Chiou-Shann Fuh. Multiple kernel learning for dimensionality reduction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(6):1147–1160, 2011.

[9] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[10] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *Artificial Neural NetworksICANN'97*, pages 583–588. Springer, 1997.

[11] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[12] Laurent Zwald and Gilles Blanchard. On the convergence of eigenspaces in kernel principal component analysis. 2006.