

Distribution Embeddings in Reproducing Kernel Hilbert Spaces

Arthur Gretton, Karsten Borgwardt, Kenji Fukumizu¹, Bernhard Schölkopf, Alexander Smola²

The “kernel trick” is well established as a means of constructing nonlinear algorithms from linear ones, by transferring the linear algorithms to a high dimensional *feature space*: specifically, a reproducing kernel Hilbert space (RKHS). Recently, it has become clear that a potentially more far reaching use of kernels is as a linear way of dealing with higher order statistics, by embedding *probability distributions* in a suitable RKHS. These representations allow us to painlessly represent high order properties of distributions, and to compare distributions in a nonparametric setting.

As a first application, we consider the problem of testing whether two samples are from two different distributions, or from the same distribution (this is called the two sample or homogeneity problem). For instance, we might wish to find whether DNA microarray measurements obtained on the same tissue type by different labs are distributed identically, or whether differences in lab procedure are such that the data have dissimilar distributions (and cannot be aggregated). We solve the two sample problem by comparing the sample feature space means, using as our statistic the distance between the means, or *Maximum Mean Discrepancy* (MMD). When the population feature means are identical (and given a sufficiently rich RKHS), the distributions are guaranteed to be the same [1]. Our test outperforms competing approaches on data of high dimensionality and low sample size (such as microarray data); moreover, being based on kernels, it can be used to compare distributions on graphs (in our work, these represent proteins of different families), for which no alternative test currently exists.

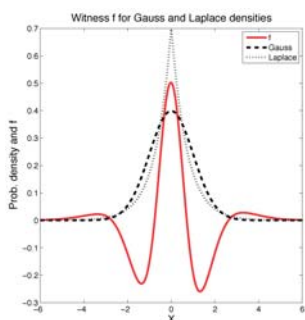
A related problem that can be solved using kernel mean representations is to determine whether two variables are dependent or independent. For instance, does a particular electrode or fMRI measurement of brain activity fluctuate together with a given stimulus, or do they bear no relation? We provide a measure of the strength of this dependence, and a test of whether the dependence is significant, in [2]. As with the two-sample case, this kernel test applies to structured data: for instance, we reveal the dependence between text and its translation, as compared to extracts on the same topic but otherwise unrelated. The strength of the dependence between random variables can be expressed as a maximum mean discrepancy between the joint distribution and the product of the marginals [3]. Along analogous lines, we also formulate measures of conditional dependence [4] and associated tests, which are a key element in graphical modeling and causal

inference. Further applications of kernel mean matching we have addressed include sample selection bias correction [5], feature selection [6], clustering, density estimation, maximum variance unfolding with side information, and independent component analysis [7].

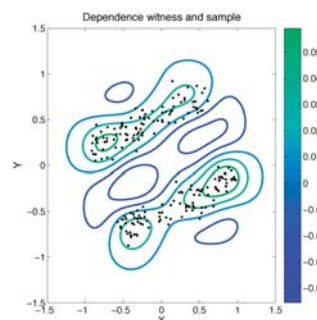


References

1. **Gretton, A.**, K. M. Borgwardt, M. Rasch, **B. Schölkopf**, A. Smola: A Kernel Method for the Two-Sample-Problem. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, 513–520. (Eds.) Schölkopf, B., J. Platt, T. Hofmann, MIT Press, Cambridge, MA, USA (2007).
2. **Gretton, A.**, K. Fukumizu, C. H. Teo, **L. Song**, **B. Schölkopf**, A. J. Smola: A Kernel Statistical Test of Independence. In *Advances in Neural Information Processing Systems 20: Proceedings of the 2007 Conference*, 585–592. (Eds.) Platt, J. C., D. Koller, Y. Singer, S. Roweis, MIT Press, Cambridge, MA, USA (2008).
3. Smola, A., **A. Gretton**, L. Song, **B. Schölkopf**: A Hilbert Space Embedding for Distributions. In *Algorithmic Learning Theory: 18th International Conference (ALT 2007)*, 13–31. (Eds.) Hutter, M., R. A. Servedio, E. Takimoto, Springer, Berlin, Germany (2007).
4. Fukumizu, K., **A. Gretton**, **X. Sun**, **B. Schölkopf**: Kernel Measures of Conditional Dependence. In *Advances in Neural Information Processing Systems 20: Proceedings of the 2007 Conference*, 489–496. (Eds.) Platt, J. C., D. Koller, Y. Singer, S. Roweis, MIT Press, Cambridge, MA, USA (2008).
5. Huang, J., A. Smola, **A. Gretton**, K. M. Borgwardt, **B. Schölkopf**: Correcting Sample Selection Bias by Unlabeled Data. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, 601–608. (Eds.) Schölkopf, B., J. Platt, T. Hofmann, MIT Press, Cambridge, MA, USA (2007).
6. Song, L., A. J. Smola, **A. Gretton**, K. M. Borgwardt, J. Bedo: Supervised Feature Selection via Dependence Estimation. In *Proceedings of the 24th Annual International Conference on Machine Learning (ICML 2007)*, 823–830. (Eds.) Ghahramani, Z. ACM Press, New York, NY, USA (2007).
7. **Shen, H.**, **S. Jegelka**, **A. Gretton**: Fast Kernel ICA using an Approximate Newton Method. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS 2007)*, 476–483. (Eds.) Meila, M., X. Shen, Microtome, Brookline, MA, USA (2007).



a. The function f maximizing the mean discrepancy in the case where a Gaussian is being compared with a Laplace distribution. This function encodes the difference in the distributions being compared.



b. The function f maximizing the mean discrepancy when testing independence. A sample from dependent random variables x and y is shown in black, and the function encoding the departure from independence is plotted as a contour.

Maximum mean discrepancy for two-sample and independence problems.

¹ Institute for Statistical Mathematics, Tokyo, Japan; ² Yahoo Research, Santa Clara, USA

RKHS Metrics on Probability Measures

Bharath Sriperumbudur^{1,2}, Arthur Gretton, Kenji Fukumizu³, Gert Lanckriet¹, Bernhard Schölkopf



The concept of distance between probability measures is a fundamental one and has many applications in probability theory and statistics. Popular applications include homogeneity tests (the two-sample problem), independence tests, goodness-of-fit tests, establishing central limit theorems, density estimation, etc. Distances between probability measures have been widely studied, where examples include the Kullback-Liebler distance, Hellinger distance, total variation distance, Wasserstein distance, and Dudley metric.

In this work, we study a novel measure of distance between probability measures obtained by embedding them into a reproducing kernel Hilbert space (RKHS), and computing the RKHS distance between these embeddings [1,4]. Formally, if P and Q are probability measures on a measure space M , then the proposed distance between them is given by

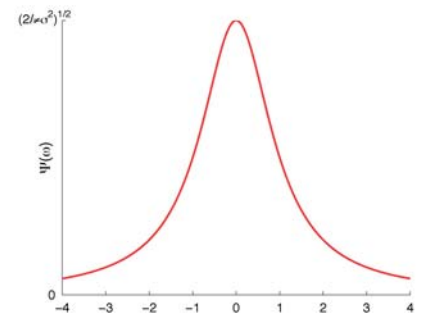
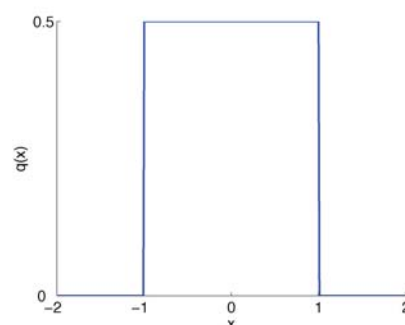
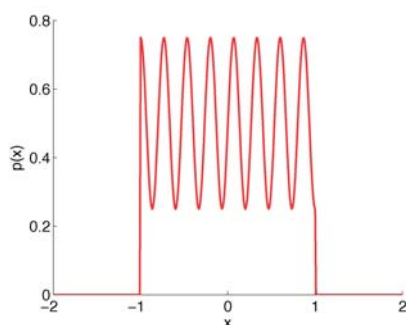
$$D(P, Q) = \left\| \int_M k(\cdot, x) dP(x) - \int_M k(\cdot, x) dQ(x) \right\|_{\mathcal{H}}, \quad (1)$$

where $\|\cdot\|_{\mathcal{H}}$ is the norm in RKHS \mathcal{H} and k is the reproducing kernel associated with \mathcal{H} (the kernel is assumed to be measurable and bounded). While D in Eq. (1) is a semi-metric on the space of all probability measures P on the domain M for any k , this does not guarantee the important property $D(P, Q) = 0 \Leftrightarrow P = Q$. In other words, we require D to be a *metric* so as to distinguish arbitrary probability distributions (and to be relevant to the applications cited earlier). Our main research question is therefore: Under what conditions on the kernel and on the class of probability measures considered is D a metric? We call such kernels *characteristic kernels*. As an example, suppose we are given two probability distributions P and Q shown in Figures 1(a) and 1(b). Can we distinguish P and Q by using the kernel function shown in Figure 1(c)? Instead of answering this question for a particular kernel, we are concerned with the broader question of determining the properties necessary and sufficient to *all* characteristic kernels on a given domain. Fukumizu *et al.* [2] have shown that k is characteristic if and

only if the direct sum of \mathcal{H} and \mathbb{R} is dense in the Banach space of q -integrable functions. This condition can be difficult to verify, however. More recently, we proposed in [5] a much more easily checked condition for the case $M = \mathbb{R}^n$ and k continuous and translation invariant ($k(x, y) = \psi(x - y)$): k is characteristic if and only if the support of the Fourier transform of ψ is \mathbb{R}^n . We also showed that all compactly supported, bounded, continuous, translation-invariant positive definite kernels on \mathbb{R}^n are characteristic. Recently, we extended these results to locally compact Abelian groups, compact groups, and semigroups [3].

References

- Berlinet, A., C. Thomas-Agnan: Reproducing Kernel Hilbert Spaces in Probability and Statistics. *Kluwer Academic Publishers*, London, UK, (2004).
- Fukumizu, K., A. Gretton, X. Sun, B. Schölkopf: Kernel Measures of Conditional Dependence. In *Advances in Neural Information Processing Systems 20: Proceedings of the 2007 Conference*, 489-496. (Eds.) Platt, J. C., D. Koller, Y. Singer, S. Roweis, MIT Press, Cambridge, MA, USA (2008).
- Fukumizu, K., B. K. Sriperumbudur, A. Gretton, B. Schölkopf: Characteristic Kernels on Groups and Semigroups. In *Advances in Neural Information Processing Systems 21: Proceedings of the 2008 Conference*, MIT Press, Cambridge, MA, USA (2009).
- Smola, A., A. Gretton, L. Song, B. Schölkopf: A Hilbert Space Embedding for Distributions. *Algorithmic Learning Theory: 18th International Conference (ALT 2007)*, 13-31. (Eds.) Hutter, M., R. A. Servedio, E. Takimoto, Springer, Berlin, Germany (2007).
- Sriperumbudur, B. K., A. Gretton, K. Fukumizu, G. Lanckriet, B. Schölkopf: Injective Hilbert Space Embeddings of Probability Measures. *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, 111-122. (Eds.) Servedio, R., T. Zhang, Springer, Berlin, Germany (2008).



1 (a) and 1 (b) represent the probability distributions P and Q respectively. 1 (c) represents the kernel, k .

¹ University of California, San Diego, USA.; ² Max-Planck-Institut für biologische Kybernetik, Tübingen, Germany

³ Institute for Statistical Mathematics, Tokyo, Japan.

Clustering and Taxonomy Discovery

Matthew B. Blaschko, Christoph H. Lampert, Arthur Gretton

We have developed clustering algorithms that extend beyond the usual notion of predicting a partition of a data sample. Specifically, we have developed algorithms for clustering with weak forms of supervision in the form of additional modalities, e.g. images with text captions. Additionally, we have developed algorithms for simultaneously learning a data partition as well as a taxonomy that relates the clusters. Unlike previous methods that have relied on greedy approaches, or sampling intensive optimization, we directly optimize a measure of dependence between the original data and the predicted partition and structure. Spectral clustering algorithms are an important subset of unsupervised methods that have been shown to have strong theoretical guarantees as well as good empirical performance. While traditional spectral clustering algorithms are completely unsupervised, we have developed a generalization, correlational spectral clustering, that can incorporate weak forms of supervision in the form of additional modalities [1]. Because data may be limited for which correspondences between modalities are known, we have developed a semi-supervised extension to the algorithm that allows for the incorporation of additional data in either modality [2].

Experimental results show that correlational spectral clustering consistently results in a data partition that is closer to that which has been defined by human labeling as compared with spectral clustering or linear dimensionality reduction (Table 1). Furthermore, the semi-supervised extension of KCCA results in improved class separation over KCCA without additional data. We have also introduced a family of unsupervised algorithms, numerical taxonomy clustering, to simultaneously cluster data, and to learn a taxonomy that encodes the relationship between the clusters. This problem is widely encountered in biology, when grouping different species; and in computer science, when summarizing and searching over documents and images. The algorithms work by maximizing the dependence between the taxonomy and the original data. The resulting taxonomy is a more informative visualization of complex data than simple clustering; in addition, taking into account the relations between

different clusters is shown to substantially improve the quality of the clustering, when compared with state-of-the-art algorithms in the literature (both spectral clustering and a previous dependence maximization approach) [3,4].

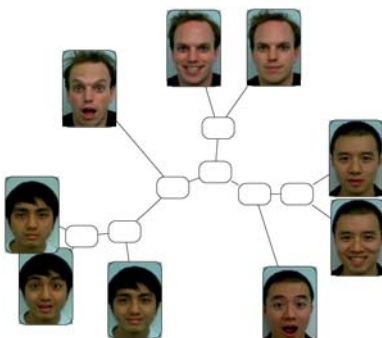
We have demonstrated in experiments that that numerical taxonomy clustering improves accuracy over existing algorithms and over clustering with a fixed structure. Experiments on a dataset of multiple images of different people and facial expressions (Figure 1), as well as on a collection of 12 years of NIPS papers (Figure 2) suggest that the algorithm leads to taxonomies that make intuitive sense. Clusters that are similar are closer together on the resulting tree, and different facial expressions and topics within the NIPS community were successfully partitioned and related. This indicates that numerical taxonomy clustering is a useful tool, both for improving the accuracy of clusterings and for the visualization of complex data.



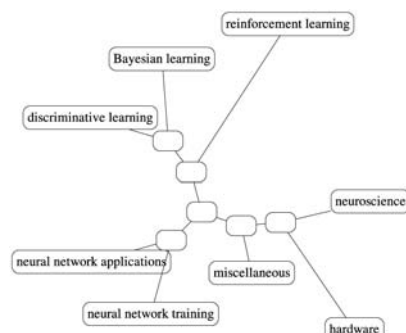
References

1. **Blaschko, M. B., C. H. Lampert:** Correlational Spectral Clustering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1–8 (2008).
2. **Blaschko, M. B., C. H. Lampert, A. Gretton:** Semi-Supervised Laplacian Regularization of Kernel Canonical Correlation Analysis. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008*, 133–145. (Eds.) Daelemans, W. Springer, Berlin, Germany (2008).
3. **Blaschko, M. B., A. Gretton:** A Hilbert-Schmidt Dependence Maximization Approach to Unsupervised Structure Discovery. *6th International Workshop on Mining and Learning with Graphs (MLG 2008)* 6, 1–3 (2008).
4. **Blaschko, M. B., A. Gretton:** Learning Taxonomies by Dependence Maximization. In *Advances in Neural Information Processing Systems 21: Proceedings of the 2008 Conference*, (2009).

	PCA	CCA	KPCA	KCCA
Israeli	3.1318	3.0638	2.9722	2.8046*
S.A.P.	0.9224	1.4699	0.8957	0.8588
Bass	2.2372	2.1880	2.1825	2.1053*
Crane	2.6416	2.6297	2.5642	2.5075*
Squash	2.3485	2.3452	2.2697	2.2517



1. Face dataset and the resulting taxonomy that was discovered by the algorithm.

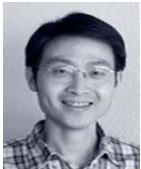


2. The taxonomy discovered for the NIPS dataset.

Table 1: Mean conditional entropy scores. Lower values indicate better clusters, and * indicates statistical significance. The proposed method, labeled KCCA, outperforms the other methods.

A Local Learning Approach for Clustering, Transductive Learning and Data Projection

Mingrui Wu¹, Bernhard Schölkopf



In supervised learning, the goal is to use a set of labeled data to train a model in order to predict the labels of unobserved data points. However, a single function may not be a good predictor for the whole input space. In such cases, it might be more appropriate to learn different functions that can make good predictions in different local regions. Given this motivation, in *local learning*, we build a model for each test data point by using only its neighboring training data, and use this locally trained model to predict the label of this particular test point.

The success of local learning in the supervised setting motivated us to adapt local learning to clustering problems, where we aim to achieve a clustering with small local learning error. Namely, the cluster label of each data point can be predicted based on its neighboring data and their cluster labels using current supervised learning methods. The key observation is that the mathematical relationship between the cluster label of a data point and the cluster labels of its neighbors can be obtained analytically via supervised learning even if the exact value of these labels are unknown. This allows us to formulate an optimization problem such that the solution can fit this relationship as well as possible [1]. Experimental results over several high dimensional image and news datasets show that our approach outperforms the widely used k-means and spectral clustering algorithms.

In [2], we apply local learning to transductive learning, where the test data set is available during training. In our approach, the goal is to find a model that has small local learning error and makes accurate predictions on labeled data. Along this direction, we analyze the transductive learning algorithm based on the graph Laplacian regularizer. We show that it implicitly performs local learning with the weighted average learning algorithm, and thus can be regarded as a special case of our local learning approach.

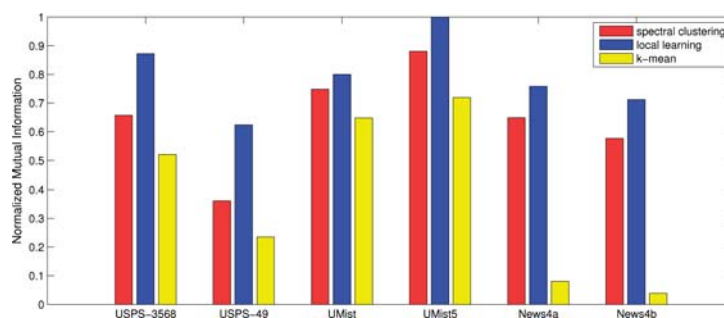
Another application of the local learning idea is feature extraction, where the goal is to project the data to a lower dimensional

subspace in order to suppress noisy and irrelevant information. In [3], we propose a Local Learning Projection (LLP) algorithm which searches for the minimal local estimation error. This means that the projection value of each data point can be accurately estimated based on its neighboring data points within the same class. Investigating the popular Principal Component Analysis (PCA) algorithm from this point of view, we find that PCA essentially gives the projection that minimizes the global estimation error, i.e. the projection of one data point can be accurately estimated based on the projection values of all the other points irrespective of their class.

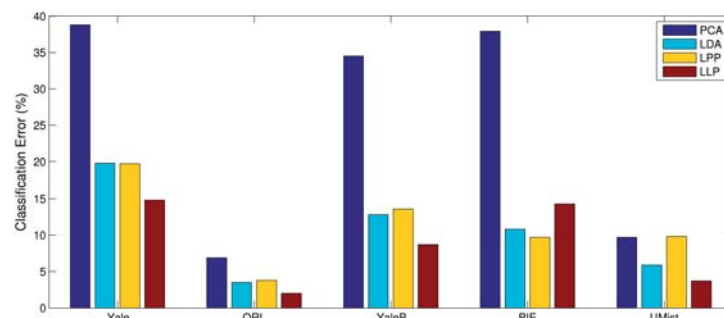
In summary, we extend local learning, originally developed only for labeled data and supervised learning, to explore relationships among labeled as well as unlabeled data points and propose new algorithms for clustering, transductive learning and data projection.

References

- Wu, M. B. Schölkopf:** A Local Learning Approach for Clustering. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, 1529–1536. (Eds.) Schölkopf, B., J. Platt, T. Hofmann, MIT Press, Cambridge, MA, USA (2007).
- Wu, M., B. Schölkopf:** Transductive Classification via Local Learning Regularization. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS 2007)*, 628–635. (Eds.) Meila, M., X. Shen, Microtome, Brookline, MA, USA (2007).
- Wu, M., K. Yu, S. Yu, B. Schölkopf:** Local Learning Projections. In *Proceedings of the 24th Annual International Conference on Machine Learning (ICML 2007)*, 1039–1046. (Eds.) Ghahramani, Z., ACM Press, New York, NY, USA (2007).



Clustering results of spectral clustering, local learning approach and the k-means algorithm.



Test error rate of different data projection methods on five face image datasets.

¹ Now at: Yahoo Research, Santa Clara, USA

Analyzing the Universum Inference Principle

Fabian H. Sinz, Olivier Chapelle¹, Alekh Agarwal², Bernhard Schölkopf

Probably the most important insight of philosophy of science is that no empirical fact can be inferred with absolute certainty. The machine learning counterpart of this result is the no free lunch theorem. In essence it states that the only way learning algorithms can outperform random guessing is to make correct assumptions about the underlying regularity in the data. While traditional learning algorithms make very general assumptions in form of a prior or a regularizer, a new form of regularizers which depend on another dataset has recently been proposed under the general term of data-dependent regularization. One specific example thereof is the Universum principle proposed by Vapnik [4], which was algorithmically implemented into a Support Vector Machine by Weston et al. [5].

The idea of learning with the Universum is to train a model that classifies the training data accurately and remains agnostic about the class labels of another set – called the Universum set – at the same time. Our goal in this project was to analyze which features of the Universum set mainly influence the solution of the Universum algorithm in order to have principled criteria to choose Universum sets and to relate this regularization principle to existing algorithms.

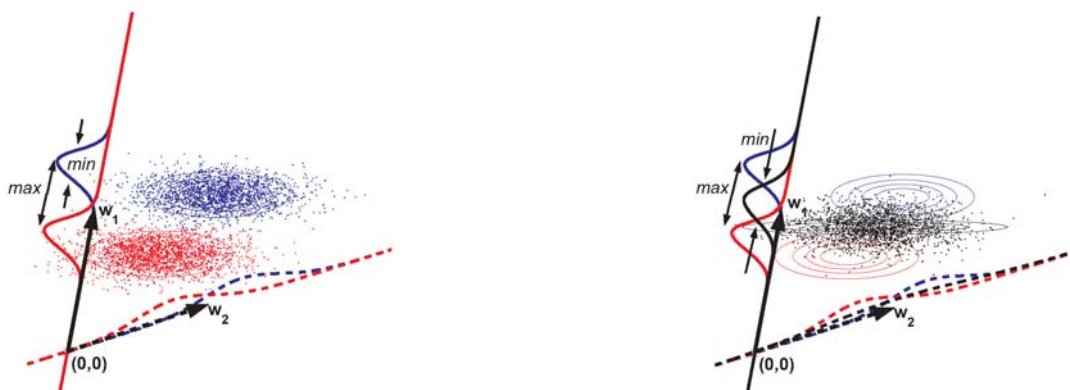
The Universum algorithm uses a standard Support Vector Machine with an additional loss term that forces the elements of the Universum set to be close to the hyperplane separating the two classes of the training set. We analyzed the algorithm for two different loss functions on the training and the Universum set: the hinge loss and the quadratic loss. We showed that for both loss functions the Universum algorithm aims to make the solution as orthogonal as possible to the subspace spanned by the Universum points [3, 2]. One would expect that the principal components of the Universum set have the strongest influence on the solution in this approach. For the quadratic loss this intuition is exact. We showed that the Universum algorithm with quadratic loss is in fact equivalent to a hybrid of oriented Principal Component Analysis and Fisher Discriminant Analysis (see Figure for illustration). We verified our theoretical results on toy data and real world datasets [3].

Our analysis shows that the implementation of data-dependent regularization exhibits a strong link to known statistical learning methods such as oriented Principal Component Analysis and Fisher Discriminant Analysis. The Universum set serves as a means to implicitly specify variations the resulting classifier should be invariant against. For instance, in digit recognition this could be small rotations, translations or different stroke widths. There are other algorithms that are able to incorporate invariances (e.g. [1]). However, the transformations have to be known explicitly for these algorithms. While variations are usually not known in an analytic form, the Universum algorithms offers a generic way to specify them implicitly via the Universum set that contains data subject to those variations. This renders it a general form of data-dependent regularization.



References

1. **Chapelle, O., B. Schölkopf:** Incorporating Invariances in Nonlinear SVMs. In *Advances in Neural Information Processing Systems 14: Proceedings of the 2001 Conference*. (Eds.) Dietterich, T. G., S. Becker, Z. Ghahramani, MIT Press, Cambridge, MA, (2002).
2. **Sinz F. H.:** A priori Knowledge from Non-examples. Master's thesis. URL: <http://www.kyb.tuebingen.mpg.de/bs/people/fabee/universvm.html>.
3. **Sinz, F. H., O. Chapelle, A. Agarwal, B. Schölkopf:** An Analysis of Inference with the Universum. In *Advances in Neural Information Processing Systems 20: Proceedings of the 2007 Conference*, 1369–1376. (Eds.) Platt, J. C., D. Koller, Y. Singer, S. Roweis, MIT Press, Cambridge, MA, USA (2008).
4. Vapnik, V. N.: *Statistical Learning Theory*. John Wiley & Sons Inc (1998).
5. Weston, J., R. Collobert, **F. Sinz**, L. Bottou, V. Vapnik: Inference with the Universum. In *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*, 1009–1016. (Eds.) Cohen, W. W., A. Moore, ACM Press, New York, NY, USA (2006) URL: <http://www.kyb.tuebingen.mpg.de/bs/people/fabee/universvm.html>.



Visualization of the objectives of Fisher Discriminant Analysis (*left*) and the hybrid of Fisher Discriminant Analysis and Oriented Principal Component Analysis (*right*). FDA aims to minimize the variance of the projections onto the normal (black arrows) of the separating hyperplane (blue and red Gaussians) while pushing them as far apart as possible. The hybrid does exactly the same with the only difference that it additionally aims to minimize the projections of the Universum set (black Gaussian).

¹ Now at: Yahoo Research, Santa Clara, USA; ² Now at: University of California, Berkeley, USA

New Directions in Structured Output Prediction

Yasemin Altun



Discriminative learning methods, such as Support Vector Machines (SVMs) and Gaussian Processes (GPs), traditionally do not exploit dependencies between class labels where more than one label is predicted. Many real-world classification problems, on the other hand, involve sequential, temporal or structural dependencies between multiple labels. The goal of structured output (SO) learning is to generalize discriminative learning methods for scenarios where the prediction consists of multiple inter-dependent variables. In recent years, we proposed various structured output learning problems in supervised learning setting. In [1], [2] and [3], we generalized SVMs, GPs and Boosting to structured prediction respectively. Furthermore, we studied SO learning for semi-supervised learning [4] where we used graph Laplacian techniques in order to ensure the structured prediction classifier is smooth over the structured input space.

In our recent work, we build upon our previous research and investigate making inference on *complex* tasks, in the sense that correct prediction of the task depends on reliable inference of multiple simpler tasks. In particular, we consider problems that consist of multiple interdependent structured prediction problems. Two simple examples, Named-Entity Recognition and parsing, are seen in Figure. Instead of the predominant cascaded approach, where subtasks are solved in a cascaded manner and their outputs are used as features, we propose learning these tasks jointly using techniques from structured prediction learning and multi task learning. This approach, although computationally more intensive, can mitigate the error-propagation problem of the cascaded approach. Moreover, it allows information flow from tasks higher in the cascaded architecture to the lower ones, which can lead to better performance of the lower tasks. More importantly, it provides a very informative regularizer that imposes smoothness of multiple predictors simultaneously. Our goal is to find a discriminant function over all tasks by simultaneously discovering a hidden representation θ of some shared characteristics across multiple structured prediction tasks and learning the corresponding parameters of each task w_i . In order to extract the common structures over the input space, we assume a low dimensional representation shared across all tasks. We optimize a loss function which consists of a standard

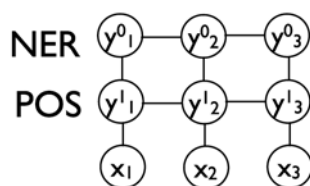
structured prediction loss (as in [1, 2]) for each task and a regularization term for each component, θ and w_i . This leads to an optimization problem over a loopy graph. We devise an approximation method that captures the specific dependencies of multiple structured prediction tasks, where it is assumed that the dependencies of variables within tasks are stronger than the dependencies of variables across tasks.

In our initial experiments, the target task is Named-Entity recognition and the subtask is part-of-speech tagging. We model these as two inter-dependent label sequence prediction problem and optimize conditional loglikelihood of both tasks (CRFs) regularized as above. We observe performance improvements in the target task as well as subtask tasks over the described cascaded approach. We are currently investigating the empirical results with other application domains and the use of structured prediction approaches other than CRFs.

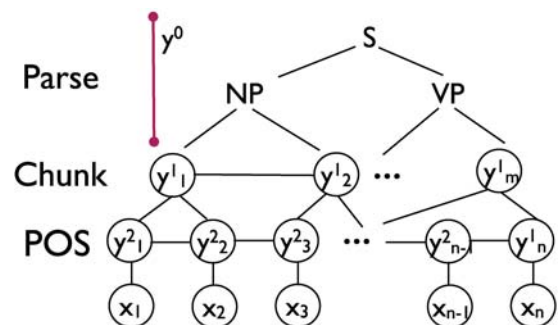
References

1. Altun, Y., T. Hofmann, I. Tsochantaridis: Large Margin Methods for Structured and Interdependent Output Variables. *Predicting Structured Data*, 85–104. (Eds.) Bakır, G., T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, S.V.N. Vishwanathan, MIT Press, Cambridge, MA, USA (2006).
2. Altun, Y., T. Hofmann, A. Smola: Gaussian Process Classification for Segmenting and Annotating Sequences. In *Proceedings of the 21st International Conference on Machine Learning*. (Ed.) Carla E. Brodley, ACM (2004).
3. Altun, Y., T. Hofmann, M. Johnson: Discriminative Learning for Label Sequences via Boosting. In *Advances in Neural Information Processing Systems 15*, 977–984. (Eds.) Becker, S., S. Thrun, K. Obermayer, MIT Press, Cambridge, MA, USA (2003).
4. Altun, Y., D. McAllester, M. Belkin: Maximum Margin Semi-Supervised Learning for Structured Variables. In *Advances in Neural Information Processing Systems 18: Proceedings of the 2005 Conference*, 33–40. (Eds.) Weiss, Y., B. Schölkopf, J. Platt, MIT Press, Cambridge, MA, USA (05/2006).

a.



b.



Examples of complex prediction problems (a) Target task:Name Entity Recognition (NER) Subtask:Part-of-Speech (POS) tagging (modeled as two label chains) (b) Target task:Parsing Subtasks:POS tagging and Chunking (modeled as a chain, semi-chain and tree).

Semi-supervised Structured Output Learning via Convex Duality

Ayse Naz Erkan, Yasemin Altun

In structured output (SO) learning, the goal is to learn a mapping from arbitrary input spaces to output spaces whose elements are *structured objects* such as sequences, trees, strings or graphs. Various SO algorithms have been proposed for supervised learning in the last few years. In our recent work, we provided a unifying analysis of the existing supervised learning methods for SO prediction by employing convex duality techniques. Moreover, using the insights that we gained from this analysis, we developed a class of semi-supervised learning methods for SO prediction problems.

In the last decade, a number of studies showed that logistic regression and boosting can be cast as the convex duals of KL divergence minimization with respect to some data constraints. These constraints enforce that the expected values of some features with respect to the target distribution match their empirical counterparts. In [1], we provided a unified framework for relating a large set of statistical inference methods to various divergence minimization problems. These inference methods include kernel methods, such as Gaussian Processes and kernel Least Squares for classification and regression. We recently extended this analysis to algorithms for SO learning via *structured data constraints*, where the features are defined over the structured objects (such as cliques of a graph). Combining these constraints with various divergence functions and taking the convex duals of these minimization problems result in various well-known structured prediction methods, e.g. references in [on p. 26].

Motivated by this unified analysis, we proposed extending the divergence minimization framework to semi-supervised learning. In particular, we modified the data constraints such that the

expectations are taken with respect to both labeled and unlabeled instances. If the labeled and unlabeled data come from the same marginal distribution, $p(x)$, unlabeled data can significantly improve the accuracy of the expected values, which in turn leads to tighter generalization bounds of the trained model. In order to satisfy this implicit assumption, we explicitly added constraints that force the expected values of features on unlabeled data to match the expected values of features on labeled data. Combining these data constraints with various divergence functions results in different semi-supervised learning methods. Moreover, when the constraints are structured (as mentioned above), our framework yields a family of structured semi-supervised learning methods. In our preliminary experiments, we combined the new set of constraints with the minimization of Kullback-Leibler divergence, which yields semi-supervised logistic regression (for standard classification) and semi-supervised Conditional Random Fields (for structured prediction). We observed around 30% error reduction with semi-supervised logistic regression over supervised logistic regression on various commonly used datasets such as g50c, Text. We are currently experimenting with semi-supervised Conditional Random Fields.



References

1. Altun, Y., T. Hofmann, I. Tsochantaridis: Large Margin Methods for Structured and Interdependent Output Variables. *Predicting Structured Data*, 85–104. (Eds.) Bakır, G., T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, S.V.N. Vishwanathan, MIT Press, Cambridge, MA, USA (2006).

Learning the Optimal Kernel Parameters for Support Vector Machines

Peter V. Gehler, Sebastian Nowozin

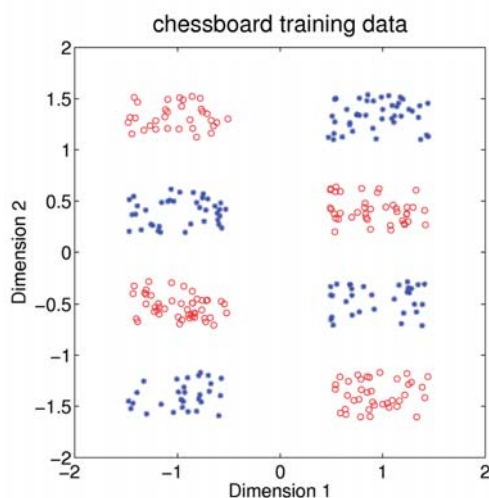


For every algorithm making use of positive definite kernels there is the same question to answer: What is the best kernel for a given task? Ideally it should be set such that the resulting function minimizes the expected risk of the task. Since this measure is not accessible, the most widely used approach is to approximate it using a Cross Validation scheme. When training Support Vector Machines (SVMs), we need to tune two components: one is the set of kernel parameters whilst the other is the regularization constant controlling the trade-off between the complexity of the function class and its ability to explain the data correctly. In Cross Validation, the parameters are picked from a range of values that perform the best on some held-out data. This becomes infeasible if the number of kernel parameters to be searched over is large, e.g. more than 10.

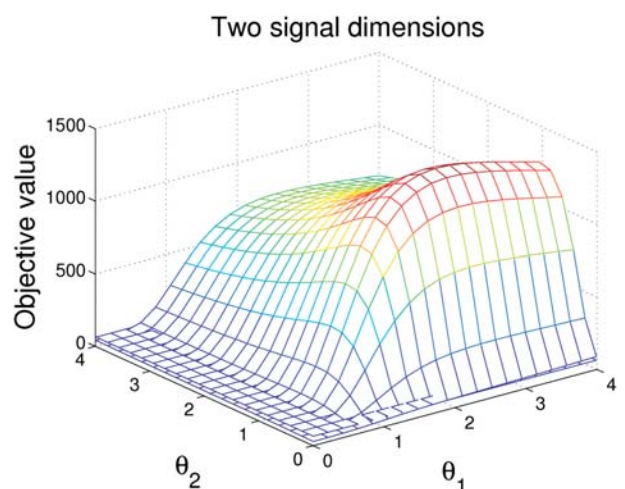
In our work we follow a different approach to select the correct kernel. We build upon the Multiple Kernel Learning (MKL) framework (see also [p. 59]) in which the kernel is sought as the convex combination of a finite number of base kernels. The mixing coefficients for this combination are jointly optimized with the remaining SVM parameters thus performing an automatic kernel selection. In our work we show that the restriction of MKL to consider only finite combinations of kernels is unnecessary. We reformulate the problem as finding the convex

combination in a possibly infinite set of kernels and construct an Semi Infinite Linear Program to solve for the parameters. Theoretical results show that the solution is always well defined, i.e., admits a finite representation, and furthermore give convergence statements for the algorithm.

The proposed algorithm along with the standard SVM and MKL are benchmarked on a large set of standard benchmark datasets commonly used in the literature. There are two interesting findings. The first observation is that although our algorithm has access to a large class of kernels (the Gram matrix ranges between the identity and an all-ones matrix) there is no sign of overfitting. Optimizing the parameters during SVM training performs as well as SVM paired with CV parameter estimation. The second observation is that a learned kernel mostly does not improve the classification performance. However there are some cases for which we record impressive performance gains. Selecting different kernels can also be viewed as selecting different features of the data. With our approach it is possible to embed the feature selection process into the SVM training. Furthermore we expect a successful application of this technique to settings where the sheer amount of possible kernels is too large for standard SVM and MKL training.



1. Projection of a 20 dimensional toy problem onto two signal dimensions. The remaining 18 dimensions are Gaussian noise. The two different classes are color and marker coded. The best Gaussian kernel to use in an SVM models the signal dimensions with two different bandwidths and ignores the remaining noise dimensions. Cross Validation is prohibitive in this case but our approach correctly identifies the best kernel.



2. Objective function guiding the selection of the best kernel which is parameterized over the two parameters θ_1 , θ_2 . Values with a high objective function correspond to kernels whose inclusion will yield to a better SVM objective. The algorithm searches for the maximum of this possibly non-convex function and includes the kernel in the problem. Note that the maximum of the objective function corresponds to the best kernel for the training data shown in Figure 1.

Learning Extremely Sparse Kernel Machines

Christian Walder¹, Kwang In Kim², Olivier Chapelle³, Bernhard Schölkopf

In the last decade or so, kernel methods such as support vector machines, Gaussian processes and splines have emerged among the most powerful tools available for a wide range of machine learning and more general function estimation problems. This can be attributed to the fact that, in addition to yielding fine practical results, such methods are theoretically elegant, and as a result may be readily adapted to solve a multitude of important problems. However, kernel methods are notoriously computationally heavy, precluding their application to a wide range of practical settings in which their good performance would otherwise make them a natural choice.

A large amount of work has been done to alleviate this computational problem, either by approximating the solution to the optimization problem solved by any particular kernel based algorithm, or constructing degenerate kernels functions for which the exact solution is less expensive to compute. Of these two categories, the latter is perhaps the most general, and is the one into which the present project falls. The majority of approaches from this category seek computational advantages by basing their computations on a reduced set of basis functions, with the constraint that these basis functions are the kernel function with one of its two inputs fixed. The present project strikes out at this constraint by considering more general classes of basis functions.

The project has led to good results on a number of problems. In particular, in Figure 1 we depict a high quality spline based surface fitting solution involving many millions of data points [3]. We have also derived a sparse Gaussian process regression algorithm which obtains state of the art results for a given number of basis functions – see Figure 2 as well as [2]. Although our new methods are necessarily more complex to train than their predecessors, it is possible forward to apply the underlying ideas to a range of new problems, given the results of this project.

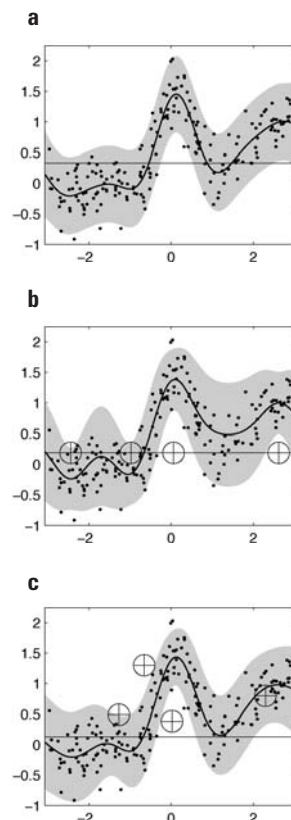


References

1. Snelson, E., Z. Ghahramani: Sparse Gaussian Processes using Pseudo-Inputs. In *Advances in Neural Information Processing Systems 18: Proceedings of the 2005 Conference*, 1257–1264. (Eds.) Weiss, Y., B. Schölkopf, J. Platt, MIT Press, Cambridge, MA, USA (05/2006).
2. Walder, C., K. I. Kim, B. Schölkopf: Sparse Multiscale Gaussian Process Regression. In *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, 1112–1119. (Eds.) Cohen, W. W., A. McCallum, S. T. Roweis, ACM Press, New York, NY, USA (2008).
3. Walder, C., B. Schölkopf, O. Chapelle: Implicit Surface Modelling with a Globally Regularised Basis of Compact Support. *Computer Graphics Forum (Proc. Eurographics 2006)*, 25(3), 635–644 (2006).



1(a). Rendered implicit surface model of “Lucy”, constructed from 56 million data points. (b) The center locations of the 364,982 compactly supported basis functions which are used to define the surface.



2. Predictive distributions (mean curve with \pm two standard deviations shaded) of *a*: the exact Gaussian process (which employs 200 basis functions), *b*: an approximation with 4 fixed-width basis functions as per [1], and *c*: our approximation, which also employs 4 basis functions, but with variable widths. For the approximations, we render the basis as crossed circles at vertical positions which are proportional to the width of the basis function, see [2] for more details.

¹ Now at: Technical University of Denmark, Lyngby, Denmark; ² Now at: Universität des Saarlandes, Saarbrücken, Germany;

³ Now at: Yahoo Research, Santa Clara, USA