MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Semi-Supervised Transfer Learning Using Marginal Predictors

Deshmukh, A.; Laftchiev, E.

TR2018-040 June 04, 2018

Abstract

This paper addresses the problem of using unlabeled data in transfer learning. Specifically, we focus on transfer learning for a new unlabeled dataset using partially labeled training datasets that consist of a small number of labeled data points and a large number of unlabeled data points. To enable transfer learning, we assume that the training and testing datasets are drawn from similar probability distributions and that the unlabeled data in each dataset can be described by similar underlying manifolds. The solution offered is a distribution free, kernel and graph Laplacian-based approach which optimizes empirical risk in the appropriate reproducing kernel Hilbert space. The approach is tested on a synthetic dataset for classification accuracy and on the Parkinson's Telemonitoring dataset from the UCI machine learning repository for prediction accuracy. Our results show a 27.3% improvement in miss-classification error and a 5.9% improvement in prediction error as compared to standard supervised learning algorithms. The results shown in this work can be widely applied in domains from medicine, to machine reliability, to prediction of human actions.

IEEE Data Science Workshop

SEMI-SUPERVISED TRANSFER LEARNING USING MARGINAL PREDICTORS

Aniket Anand Deshmukh[*]

University of Michigan
Electrical Engineering and Computer Science
Ann Arbor, MI 48105

Emil Laftchiev

Mitsubishi Electric Research Labs
Data Analytics
Cambridge, MA 02139

ABSTRACT

This paper addresses the problem of using unlabeled data in transfer learning. Specifically, we focus on transfer learning for a new unlabeled dataset using partially labeled training datasets that consist of a small number of labeled data points and a large number of unlabeled data points. To enable transfer learning, we assume that the training and testing datasets are drawn from similar probability distributions and that the unlabeled data in each dataset can be described by similar underlying manifolds. The solution offered is a distribution free, kernel and graph Laplacian-based approach which optimizes empirical risk in the appropriate reproducing kernel Hilbert space. The approach is tested on a synthetic dataset for classification accuracy and on the Parkinson's Telemonitoring dataset from the UCI machine learning repository for prediction accuracy. Our results show a 27.3% improvement in miss-classification error and a 5.9% improvement in prediction error as compared to standard supervised learning algorithms. The results shown in this work can be widely applied in domains from medicine, to machine reliability, to prediction of human actions.

***Index Terms*—** transfer learning, semi-supervised learning, unlabeled data, regression, classification

1. INTRODUCTION AND BACKGROUND

Recently, supervised learning methods which rely on the abundant availability of labeled training data have been very successful in solving challenges in computer vision and speech signal processing. Yet there exists many promising application areas where collecting labeled data is difficult and expensive.

As an example consider the problem of detecting and monitoring Parkinson's disease [1]. Detecting the presence and estimating the severity of the disease is difficult because symptoms are not readily observable until significant neurological damage has taken place. To detect early stages of the disease, scientists observe vocal measurements from the patient which can then be used to create a predictor to monitor the disease progression. However, learning such a predictor is difficult requiring subject cooperation, consistent data collection [2], and expensive expert labeling. Furthermore, each new patient suffers from a lack of labeled data that precludes a fast diagnosis of the progression of the disease. Thus the methodology required to achieve the results in [1] is slow and expensive necessitating new methods to reduce the cost of data collection and improve the speed of model training.

This paper proposes a novel algorithm that: uses transfer learning to reduce the number of required labeled training examples and removes the need for labels on the test dataset; uses unlabeled data to further reduce the need for labeled examples in the training dataset. This algorithm demonstrates that semi-supervised transfer learning, under the appropriate assumptions, can improve prediction as much as 5.9% on the Parkinson's Telemonitoring dataset and can improve classification as much as 27.3% in synthetic manifold data as compared to standard transfer learning methods.[1]

Prior semi-supervised learning and transfer learning algorithms such as the co-training method for inductive transfer learning by Yuan et. al. [3], transfer progressive transductive support vector machine method by Zhou et. al. [4] and the self-taught learning method by Raina et. al. [5] all used some unlabeled data during training, however these methods also rely on labeled samples in the test dataset. Critically, the work in this paper does not assume any labeled data points in the test dataset. A method that did not require labeled data points in the test dataset was shown in [6]; this method is limited by its assumption of a single source and test domain and was not easily extensible. The algorithm proposed in this paper is capable of leveraging multiple source domains.

2. MATHEMATICAL PRELIMINARIES

The following mathematical background is necessarily brief, but we refer the interested reader to [7] for more information. To begin let $\mathcal{X}$ be the feature space and $\mathcal{Y}$ be the labeled space. Let P_{XY} be the probability distribution on $\mathcal{X} \times \mathcal{Y}$. Training samples $\{x_i, y_i\}_{i=1}^m$ are i.i.d. drawn from P_{XY}. In case of supervised and semi-supervised learning, the goal is to find a

[*] Author performed the work while at Mitsubishi Electric Research Labs.

[1] Code implementing the approach herein can be found at www.merl.com.

function $f : \mathcal{X} \to \mathcal{Y}$. When $\mathcal{Y} \in \mathbb{R}$ then this is a regression problem and in the case when $\mathcal{Y} \in \mathbb{N}$, this is a classification problem. Kernel functions, k, are used to transfer the learning problem into a richer (non-linear) function space.

2.1. Kernel-based learning algorithms

The function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel on $\mathcal{X}$ if the matrix $[k(x_i, x_j)]_{i,j}$ is positive semi-definite (PSD). The existence of a kernel, k, on $\mathcal{X}$ implies the existence of a Hilbert Space H and a mapping $\phi : \mathcal{X} \to H$ such that $k(x, y) = \langle \phi(x), \phi(y) \rangle_H$, where H and ϕ are not uniquely determined by k. Kernel based algorithms are solved using the following result.

Theorem 1 (Representer Theorem) *[8] Let k be a kernel on $\mathcal{X}$ and let $\mathcal{H}$ be it's associated RKHS. Fix $x_1, \ldots, x_m \in \mathcal{X}$, and consider the optimization problem*

$$\min_{f \in \mathcal{H}} D(f(x_1), \ldots, f(x_m)) + P(\|f\|_H^2), \tag{1}$$

where P is non decreasing and D depends on f only through $f(x_1), \ldots, f(x_m)$. If (1) *has a minimizer, then it has a minimizer of the form $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$, where $\alpha_i \in R$. Furthermore if P is strictly increasing, then every solution of* (1) *has this form.*

2.2. Supervised Learning

For a sufficient large labeled dataset the problem is termed supervised and can be solved as follows. Let $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function. Then solve the following minimization problem to learn a function, f.

$$\min_{f \in H} J(f) = \min_{f \in H} \lambda \|f\|^2 + \frac{1}{m} \sum_{i=1}^{m} \ell(y_i, f(x_i)) \tag{2}$$

Here f can be computed directly via Thm 1. The function ℓ is chosen to be a hinge loss for classification problems and a square loss for regression problems.

2.3. Semi-Supervised Learning

When sufficient quantities of labeled data are not available, semi-supervised approaches are employed to use the usually ample available unlabeled data to augment the training dataset. Given some assumptions about the data, this unlabeled data can be used to improve the performance of the learned function, f. It is important to note that the performance of semi-supervised learning algorithms can approach but not exceed the performance of supervised algorithms (when labeled data is ample) which means that semi-supervised learning cannot replace supervised learning in all applications.

To incorporate unlabeled data, the training set contains both m labeled data points, $\{x_i, y_i\}_{i=1}^{m}$, and n unlabeled data points $\{x_i\}_{i=m+1}^{m+n}$. The unlabeled data points help to elucidate the data structure [9]. A critical assumption on the labeled data is that the labels are sufficiently similar for similar data points.

Solutions to semi-supervised learning can be found via expectation-maximization mixture models, self-training[10], transductive support vector machines [11], graph-based methods [12, 13] and manifold regularization [9]. In this paper we consider manifold regularization due to its close connection to kernel-based learning methods. Manifold regularization extends the problem shown in eq. (2) [9]. The idea is to create a graph using both labeled and unlabeled data and to penalize the supervised learning problem with the graph Laplacian. The graph Laplacian is a description of the data manifold and thereby indirectly the marginal distribution of the data. The learning objective is augmented as follows.

$$\begin{aligned}
\min_{f \in H} J(f) &= \min_{f \in H} \lambda \|f\|^2 + \frac{1}{m} \sum_{i=1}^{m} \ell(y_i - f(x_i)) \\
&+ \frac{\gamma}{(m+n)^2} \sum_{i,j=1}^{m+n} W_{ij} (f(x_i) - f(x_j))^2 \\
&= \min_{f \in H} \lambda \|f\|^2 + \frac{1}{m} \sum_{i=1}^{m} \ell(y_i - f(x_i)) \\
&+ \frac{\gamma}{(m+n)^2} \boldsymbol{f}^T L \boldsymbol{f}
\end{aligned}$$

where $W_{ij} = exp((x_i - x_j)^2 / 2\sigma^2)$ are edge weights in the adjacency graph of the data, $(x_1, \ldots, x_{m+n})$, for the K nearest neighbors of x_i, σ is set to 1 or is found in cross-validation, $\boldsymbol{f} = [f(x_1), \ldots, f(x_{m+n})]$, and L = D - W is the graph Laplacian. W is the graph weight matrix and D is the diagonal matrix whose elements are given by $D_{ii} = \sum_{j=1}^{m+n} W_{ij}$. The solution, obtained using a modified version of Theorem 1 [9], is expressed as $\hat{f(x)} = \sum_{i=1}^{m+n} \alpha_i K(x_i, x)$.

2.4. Transfer Learning

Extending supervised and semi-supervised learning results to new domains (datasets) is difficult and often requires new data collection and learning. One way to reduce this is to use transfer learning to learn in one setting and perform work in another. Ideally, learning can even be extended to different tasks and in different settings like batch learning, online learning, multi-armed bandits and reinforcement learning [14, 15, 16, 17, 18].

More concretely, recall that $\mathcal{X}$ is a feature space, $\mathcal{Y}$ is an output space and data samples $\{x, y\}$ are drawn from a distribution P_{XY}. Together (P_{XY},$\mathcal{X}$) form the domain pair. Thus the problem of transfer learning is to choose different domain pairs or choose different output spaces. The setting of this paper chooses different domain pairs. This means that there is a common output space for all datasets, but the marginal distribution of the underlying datasets is different. As a further

complexity, we assume that labels are only available for the training datasets but not for the testing set. This setting is the same as the case of learning marginal predictors (LMP) [15].

Thus choose N training datasets, with N similar but distinct distributions $P_{XY}^{(i)}$ on $\mathcal{X} \times \mathcal{Y}$, $i \in \{1, ..., N\}$. For each distribution, i, the training samples $S_i = (X_{ij}, Y_{ij})_{1 \leq j \leq m_i}$ are i.i.d. realizations from $P_{XY}^{(i)}$. The samples in the test dataset $S_T = (X_j^T, Y_j^T)$ are i.i.d. realizations from P_{XY}^T, but no labels Y_j are not observed. The goal is to predict the labels for the test dataset. For simplicity, assume that $m_i = m, \forall i$.

This problem is solved by leveraging the kernalized approaches in Section 2.1. Let $\bar{k}$ be a kernel on $\mathbb{P}_X \times \mathcal{X}$. Let $\hat{P}_X^{(i)}$ be the empirical marginal distribution corresponding to sample S_i. Let us denote $\mathbb{P}_X \times \mathcal{X}$ the extended input space and $\tilde{X}_{ij} = (\hat{P}_X^{(i)}, X_{ij})$ the extended data. Then solve the following minimization problem for transfer learning.

$$\min_{f \in H_{\bar{k}}} \lambda \|f\|^2 + \frac{1}{N} \sum_{i=1}^{N} \frac{1}{m} \sum_{j=1}^{m} \ell(Y_{ij}, f(\tilde{X}_{ij})) \tag{3}$$

3. SEMI-SUPERVISED TRANSFER LEARNING

Here we propose a new algorithm that leverages the unlabeled data in the transfer learning setting. Consider the same setting for the transfer learning problem as presented in Section 2.4. Then add to the m labeled points in each training dataset n unlabeled data points.

Adding the unlabeled data points means that in addition to using the marginal distributions for knowledge transfer between domains, we add knowledge of the data structure which helps learn a more accurate function f. This is particularly true in the case where relatively few labeled data points are available and thus the problem is under-determined with respect to the model.

The data structure information is added to the problem in the form of a graph Laplacian. This idea is similar to the approach used in semi-supervised learning algorithms. For transfer learning, the graph structure is built over each training dataset and added separately to the problem. More formally this can be represented by the augmented objective function as follows.

$$\begin{aligned}
& \min_{f \in H_{\bar{k}}} \lambda \|f\|^2 + \frac{1}{N} \sum_{i=1}^{N} \frac{1}{m} \sum_{j=1}^{m} \ell(Y_{ij}, f(\tilde{X}_{ij})) \\
+ \quad & \sum_{i=1}^{N} \frac{\gamma}{(m+n)^2} \sum_{p,q=1}^{m+n} W_{pq}^i (f(x_{ip}) - f(x_{iq}))^2 \\
\Longrightarrow \quad & \min_{f \in H_{\bar{k}}} \lambda \|f\|^2 + \frac{1}{N} \sum_{i=1}^{N} \frac{1}{m} \sum_{j=1}^{m} \ell(Y_{ij}, f(\tilde{X}_{ij})) \\
+ \quad & \sum_{i=1}^{N} \frac{\gamma}{(m+n)^2} f_i^T L_i f_i
\end{aligned}$$

Note here that in this notation, the i^{th} dataset has a graph representation L_i and all graph Laplacians are added in the penalty term. Note then that the penalty term is a manifold regularizer across all training datasets. The solution of the minimization above has the following form,

$$\hat{f}(\hat{P}_X, x) = \sum_{i=1}^{N} \sum_{j=1}^{m+n} \alpha_{ij} \bar{k}((\hat{P}_X^{(i)}, X_{ij}), (\hat{P}_X, x)), \tag{4}$$

which can be solved by defining the kernel $\bar{k}$ as a product kernel [15].

$$\bar{k}((P_1, x_1), (P_2, x_2)) = k_P(P_1, P_2) k_X(x_1, x_2) \tag{5}$$

In the equation above, k_X is a kernel on $\mathcal{X}$ and K_P is a kernel on $\mathbb{P}_X$. The kernel k_X on $\mathcal{X}$ is the standard kernel, but the kernel on the probability distributions, K_P, needs to be dataset dependent.

To find K_P define the mapping $\Psi : \mathbb{P}_X \to \mathcal{H}'_{k_X}$ where k'_X is a kernel on $\mathcal{X}$. Then define a new mapping such as the one used in the characteristic kernel framework [19]. (As an aside, there is an important connection between the injectivity of Ψ and universal kernels, which has been studied in [20].)

$$P_X \mapsto \Psi(P_X) := \int_{\mathcal{X}} k'_X(x, \cdot) dP_X(x). \tag{6}$$

Using these mappings define K_P on $\mathbb{P}_X$ as,

$$K_P(P_X, P'_X) = \kappa(\Psi(P_X), \Psi(P'_X)). \tag{7}$$

where κ is a kernel on $\mathcal{H}_{k'_X}$.

4. IMPLEMENTATION

To evaluate the proposed algorithm, the new method of semi-supervised transfer learning (SSTL) is implemented and compared with two existing methods: Pooling and learning marginal predictors (LMP). To create a predictor using the Pooling method, all datasets are combined into a single dataset and a function is learned on the aggregate set. The LMP predictors are learned as previously described in Section 2.4. Optimal parameters for LMP and Pooling algorithms are chosen using 5-fold cross-validation following the procedure described in [21]. The results showed that the selected hyperparameters are data set depenedent. Thus to increase robustness, hyperparameters are selected for each run of the experiment. The optimal LMP parameters are then used for the proposed algorithm, SSTL. Regularization parameters for SSTL are fixed. Note that SSTL results can further be improved if optimal parameters are chosen for SSTL.

All algorithms are implemented in Matlab using available packages. The Pooling and LMP algorithms are implemented via an existing package which is based on LibLinear [21, 22]. The proposed method, SSTL, is implemented via a modification of the LapSVM package [23]. Kernel approximation

techniques are used for the Pooling and LMP algorithms to speed up performance and the same approximation is adopted for SSTL to ensure a fair comparison [21]. All three algorithms are tested on two datasets: a synthetic dataset and the Parkinson's Telemonitoring Dataset [1]. In the synthetic dataset, the SSTL performance is demonstrated in the classification setting while in the Parkinson's dataset the performance is demonstrated in the regression setting.

Synthetic Data Generation: The synthetic dataset is composed of 25 training datasets and 5 testing datasets. Each dataset is a newly generated two dimensional manifold rotated by a randomly picked value from the set, $\{0, \frac{pi}{20}, \ldots, \frac{pi}{2}\}$. 10 of the dataset points are labeled and 400 of the points are unlabeled. Test datasets contain no labeled data points. An example manifold dataset can be seen in Fig. 1. The axes represent two-dimensional features and color represents the label. It is clear that marginal distribution for each dataset is different.

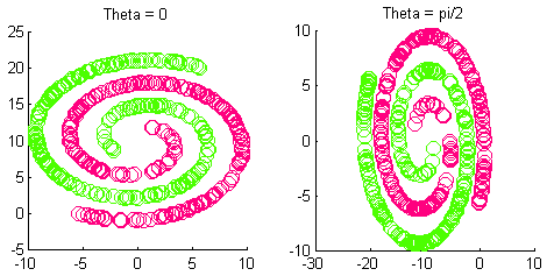

Fig. 1: Example of Synthetic Datasets

Parkinson's Telemonitoring Dataset: The Parkinson's Telemonitoring dataset is composed of voice measurements from 42 individuals with early-stage Parkinson's disease. Each individual was followed for 6 months and their voice samples were labeled on the UPDRS scale for Parkinson's. In addition, the dataset contains age, gender, and time since recruitment for each individual.

In the experiments for this paper, the goal is to predict the total UPDRS score of Parkinson's disease symptoms based on the age, gender, time since recruitment, and voice measurements. 35 of the individuals are in the training dataset, and 7 individuals are in the testing dataset. The data for all individuals contains 80 unlabeled data points, while training individuals also have 10 labeled data points. An example of the marginal distributions for one data feature for two patients is shown in Fig. 2. This type of difference holds between patients and features in the data.

For both datasets, 10 different train/test dataset splits are created and all three algorithms including parameter selection are run 20 times. The results are averaged over 200 experiments.

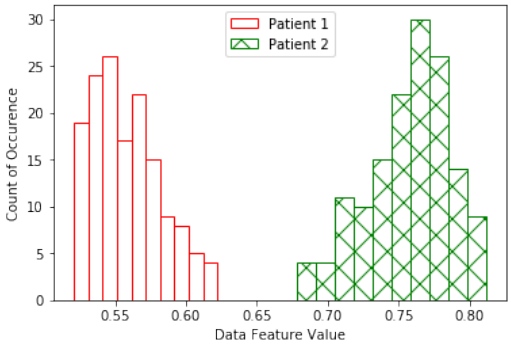

Fig. 2: Example of Parkinson's Telemonitoring Datasets

5. RESULTS AND DISCUSSION

Synthetic Data: Table 1 shows the performance of SSTL using synthetic data in the classification setting and with respect to miss-classification error. Here the worst performing method is the Pooling method which was expected because aggregating the datasets effectively destroys the manifold structure of the data and therefore increases the challenge when learning a classifier. In contrast, the LMP method improves classification results by 12.6% w.r.t. Pooling. However, LMP suffers from an extreme label deficit in the training dataset. In contrast, SSTL improves classification accuracy by 27.3% w.r.t. Pooling by explicitly learning the manifold structure of the data.

Dataset	Pooling	LMP	SSTL
Synthetic	29.28	25.57	21.28
Parkinson's	126.94	121.30	119.47

Table 1: Misclassification Error for Synthetic dataset and Mean Squared Error for Parkinson's dataset

Parkinson's Telemonitoring Dataset: The mean of the results of SSTL in the regression setting are shown in Table 1. Here again, the Pooling method has the worst performance resulting of a squared error of 126.94 on the UPDRS scale. LMP improves the results by 4.4%, while SSTL improves the results by 5.9%.

6. CONCLUSION

This paper presents an algorithm that combines labeled and unlabeled data during transfer learning. In particular, we leverage results from semi-supervised learning that show that penalizing the learning algorithm using the graph Laplacian reduces the need for labeled data samples while maintaining a constant error rate. The results are shown in both the classification and regression settings using synthetic manifold data and the real world Parkinson's Telemonitoring dataset.

7. REFERENCES

[1] Athanasios Tsanas, Max A Little, Patrick E McSharry, and Lorraine O Ramig, "Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests," *IEEE transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 884–893, 2010.

[2] John W. Graham, "Missing data analysis: Making it work in the real world," *Annual Review of Psychology*, vol. 60, no. 1, pp. 549–576, 2009.

[3] Yuan Shi, Zhenzhong Lan, Wei Liu, and Wei Bi, "Extending semi-supervised learning methods for inductive transfer learning," in *2009 Ninth IEEE International Conference on Data Mining*. IEEE, 2009, pp. 483–492.

[4] Huiwei Zhou, Yan Zhang, Degen Huang, and Lishuang Li, "Semi-supervised learning with transfer learning," in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pp. 109–119. Springer, 2013.

[5] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of the 24th International Conference on Machine learning*. ACM, 2007, pp. 759–766.

[6] Akinori Fujino, Naonori Ueda, and Masaaki Nagata, "Adaptive semi-supervised learning on labeled and unlabeled data with different distributions," *Knowledge and information systems*, vol. 37, no. 1, pp. 129–154, 2013.

[7] Arthur Gretton, "Introduction to rkhs, and some simple kernel algorithms," October 2015.

[8] Ingo Steinwart and Andreas Christmann, *Support Vector Machines*, Springer Science & Business Media, 2008.

[9] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of machine learning research*, vol. 7, no. Nov, pp. 2399–2434, 2006.

[10] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models," in *Application of Computer Vision*, Jan 2005, vol. 1, pp. 29–36.

[11] Ronan Collobert, Fabian Sinz, Jason Weston, and Léon Bottou, "Large scale transductive svms," *Journal of Machine Learning Research*, vol. 7, no. Aug, pp. 1687–1712, 2006.

[12] Avrim Blum, John Lafferty, Mugizi Robert Rwebangira, and Rajashekar Reddy, "Semi-supervised learning using randomized mincuts," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 13.

[13] Xiaojin Zhu, *Semi-supervised Learning with Graphs*, Ph.D. thesis, Carnegie Mellon University, 2005.

[14] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[15] Gilles Blanchard, Gyemin Lee, and Clayton Scott, "Generalizing from several related classification tasks to a new unlabeled sample," in *Advances in neural information processing systems*, 2011, pp. 2178–2186.

[16] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf, "Domain generalization via invariant feature representation," in *International Conference on Machine Learning*, 2013, pp. 10–18.

[17] Andrei A Rusu, Matej Večerík, Thomas Rothörl, Nicolas Heess, Razvan Pascanu, and Raia Hadsell, "Sim-to-real robot learning from pixels with progressive nets," in *Conference on Robot Learning*, 2017, pp. 262–270.

[18] Aniket Anand Deshmukh, Urun Dogan, and Clay Scott, "Multi-task learning for contextual bandits," in *Advances in Neural Information Processing Systems*, 2017, pp. 4851–4859.

[19] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander J Smola, "A kernel approach to comparing distributions," in *Association for the Advancement of Artificial*, July 2007, pp. 1637–1641.

[20] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet, "Hilbert space embeddings and metrics on probability measures," *Journal of Machine Learning Research*, vol. 11, no. Apr, pp. 1517–1561, 2010.

[21] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott, "Domain generalization by marginal transfer learning," *arXiv preprint arXiv:1711.07910*, 2017.

[22] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[23] Stefano Melacci and Mikhail Belkin, "Laplacian Support Vector Machines Trained in the Primal," *Journal of Machine Learning Research*, vol. 12, pp. 1149–1184, March 2011.