Learning from Distributions via Support Measure Machines

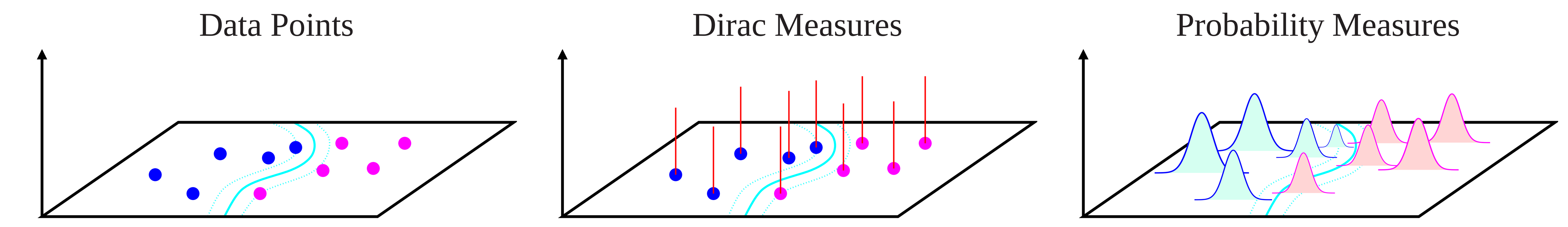
Krikamol Muandet[†], Kenji Fukumizu[‡], Francesco Dinuzzo[†], Bernhard Schölkopf[†]

[†]Department of Empirical Inference
MPI for Intelligent Systems, Tübingen, Germany
{krikamol,fdinuzzo,bs}@tuebingen.mpg.de

[‡]Department of Mathematical Analysis and Statistical Inference
The Institute of Statistical Mathematics, Tokyo, Japan
fukumizu@ism.ac.jp

MAX-PLANCK-GESELLSCHAFT

From Data Points to Probability Measures



Potential applications: Learning with noisy/uncertain examples (astronomical/biological data). Learning from groups of samples (population genetics, group anomaly detection, and preference learning). Learning under changing environments (domain adaptation/generalization). Large-scale machine learning (data squashing).

Hilbert Space Embedding

The kernel mean map from a space of distributions $\mathscr{P}$ into a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$:

$$\mu : \mathscr{P} \to \mathcal{H}, \quad \mathbb{P} \longmapsto \int_{\mathcal{X}} k(x,\cdot)\,\mathrm{d}\mathbb{P}(x) \ .$$

The kernel k is said to be *characteristic* if and only if the map μ is injective, i.e., there is no loss of information.

Representer Theorem

Given training examples $(\mathbb{P}_i, y_i) \in \mathscr{P} \times \mathbb{R}, i = 1,\dots,m$, a strictly monotonically increasing function $\Omega : [0,+\infty) \to \mathbb{R}$, and a loss function $\ell : (\mathscr{P} \times \mathbb{R}^2)^m \to \mathbb{R} \cup \{+\infty\}$, any $f \in \mathcal{H}$ minimizing the regularized risk functional

$$\ell\left(\mathbb{P}_1, y_1, \mathbb{E}_{\mathbb{P}_1}[f], \dots, \mathbb{P}_m, y_m, \mathbb{E}_{\mathbb{P}_m}[f]\right) + \Omega\left(\|f\|_{\mathcal{H}}\right)$$

admits a representation of the form $f = \sum_{i=1}^m \alpha_i \mu_{\mathbb{P}_i} = \sum_{i=1}^m \alpha_i \mathbb{E}_{x\sim\mathbb{P}_i}[k(x,\cdot)]$ for some $\alpha_i \in \mathbb{R}$, $i = 1,\dots,m$.

Key Observations

The standard representer theorem is recovered as a special case when $\mathbb{P}_i = \delta_{x_i}$. Thus, our framework generalizes the machine learning framework on data points. Moreover, our framework is different from minimizing the functional

$$\mathbb{E}_{\mathbb{P}_1} \dots \mathbb{E}_{\mathbb{P}_m} \ell(\{x_i, y_i, f(x_i)\}_{i=1}^m) + \Omega(\|f\|_{\mathcal{H}}) \qquad (1)$$

for the special case of the additive loss ℓ (**intractable**). It is also different from minimizing the functional

$$\ell(\{M_i, y_i, f(M_i)\}_{i=1}^m) + \Omega(\|f\|_{\mathcal{H}}) \qquad (2)$$

where $M_i = \mathbb{E}_{x\sim\mathbb{P}_i}[x]$ (**loss of information**).

The proposed framework does not lose information, but optimizes a less expensive problem than (1).

Kernels on Distributions

For distributions $\mathbb{P}, \mathbb{Q} \in \mathscr{P}$, a ***linear kernel*** on $\mathscr{P}$ is

$$K(\mathbb{P},\mathbb{Q}) = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} = \iint k(x,z)\,\mathrm{d}\mathbb{P}(x)\,\mathrm{d}\mathbb{Q}(z),$$

which can be approximated as

$$K(\widehat{\mathbb{P}}, \widehat{\mathbb{Q}}) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i, z_j), \quad x_i \sim \mathbb{P}, z_j \sim \mathbb{Q} \ .$$

For some distributions and kernel k, the kernel $K(\mathbb{P},\mathbb{Q})$ has an analytic form. Assume that $\mathbb{P}_i = \mathcal{N}(m_i, \Sigma_i)$:

Linear $k(x,y) = \langle x, y\rangle$:
$K(\mathbb{P}_i,\mathbb{P}_j) = m_i^\top m_j + \delta_{ij}\mathrm{tr}\,\Sigma_i$.

Gaussian RBF $k(x,y) = \exp(-\frac{\gamma}{2}\|x-y\|^2)$:
$K(\mathbb{P}_i,\mathbb{P}_j) = \exp(-\frac{1}{2}(m_i - m_j)^\top(\Sigma_i + \Sigma_j + \gamma^{-1}\mathbf{I})^{-1}(m_i - m_j))/|\gamma\Sigma_i + \gamma\Sigma_j + \mathbf{I}|^{\frac{1}{2}}$

Polynomial degree 2 $k(x,y) = (\langle x,y\rangle + 1)^2$:
$K(\mathbb{P}_i,\mathbb{P}_j) = (\langle m_i, m_j\rangle + 1)^2 + \mathrm{tr}\,\Sigma_i\Sigma_j + m_i^\top\Sigma_j m_i + m_j^\top\Sigma_i m_j$

Polynomial degree 3 $k(x,y) = (\langle x,y\rangle + 1)^3$:
$K(\mathbb{P}_i,\mathbb{P}_j) = (\langle m_i, m_j\rangle + 1)^3 + 6m_i^\top\Sigma_i\Sigma_j m_j + 3(\langle m_i, m_j\rangle + 1)(\mathrm{tr}\,\Sigma_i\Sigma_j + m_i^\top\Sigma_j m_i + m_j^\top\Sigma_i m_j)$

A ***nonlinear kernel*** can be defined as

$$K(\mathbb{P},\mathbb{Q}) = \kappa(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}}) = \langle \Phi(\mu_{\mathbb{P}}), \Phi(\mu_{\mathbb{Q}})\rangle_{\mathcal{F}}$$

where κ is a positive definite kernel function on $\mathcal{H}$. For example, $K(\mathbb{P},\mathbb{Q}) = \exp\left(-\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}^2/2\sigma^2\right)$ and $K(\mathbb{P},\mathbb{Q}) = (\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}}\rangle_{\mathcal{H}} + c)^d$.

The embedding kernel k defines the vectorial representation of the distributions, whereas the level-2 kernel κ allows for non-linear learning algorithms on probability distributions.

Risk Deviation Bound & Flexible SVMs

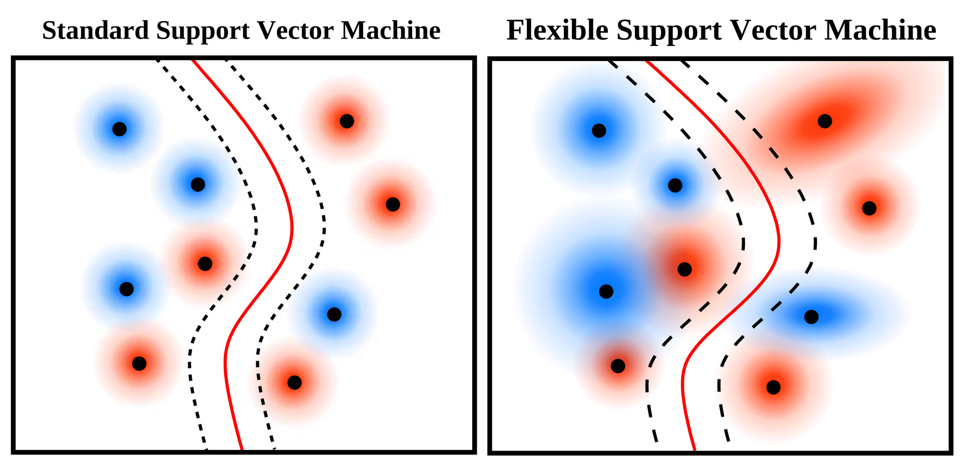
Risk Deviation Bound: Given an arbitrary distribution $\mathbb{P}$ with finite variance σ^2, a Lipschitz continuous function $f : \mathbb{R} \to \mathbb{R}$ with constant C_f, an arbitrary loss function $\ell : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ that is Lipschitz continuous in the second argument with constant C_ℓ, it follows, for any $y \in \mathbb{R}$, that

$$|\mathbb{E}_{x\sim\mathbb{P}}[\ell(y, f(x))] - \ell(y, \mathbb{E}_{x\sim\mathbb{P}}[f(x)])| \leq 2C_\ell C_f \sigma$$

Flexible SVMs: Assume that the densities of distributions $\mathbb{P}$ and $\mathbb{Q}$ are $g_x(\cdot)$ and $g_z(\cdot)$, where x and z are the parameters in the density family. Hence, we have

$$K(\mathbb{P},\mathbb{Q}) = \left\langle \int k(\tilde{x},\cdot)g_x(\tilde{x})\,\mathrm{d}\tilde{x}, \int k(\tilde{z},\cdot)g_z(\tilde{z})\,\mathrm{d}\tilde{z}\right\rangle_{\mathcal{H}} = k_g(x,z),$$

where k_g is a data-dependent p.d. kernel. That is, k_g depends not only on $x, z \in \mathcal{X}$, but also on other parameters of $\mathbb{P}, \mathbb{Q} \in \mathscr{P}$. For example, if $\mathbb{P}$ and $\mathbb{Q}$ are Gaussian distributions and the kernel k is a Gaussian RBF kernel, then we have different Gaussian RBF kernels at each data point, i.e., the means of the distributions (see figure).



Flexible SVM allows for data-dependent kernel functions, for example, pointwise uncertainties.

Experimental Results

In the experiments, we primarily consider three different learning algorithms: i) **SVM** trained on the means of the distributions is considered as a baseline algorithm (cf. (2)). ii) **Augmented SVM (ASVM)** is an SVM trained on augmented samples drawn according to the distributions $\{\mathbb{P}_i\}_{i=1}^m$ (cf. (1)). iii) **SMM** is our distribution-based method that is applied directly on the distributions.

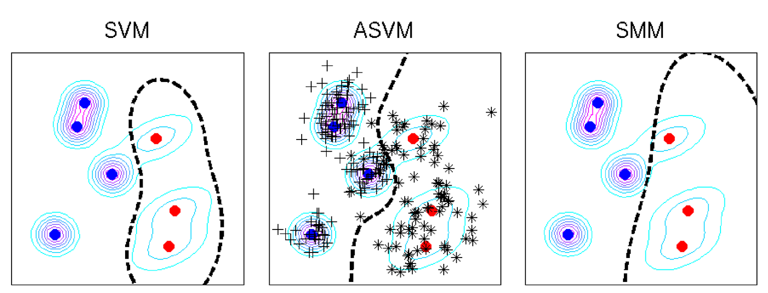


Figure 1: The decision boundaries of SVM, ASVM, and SMM on the synthetic dataset of distributions.

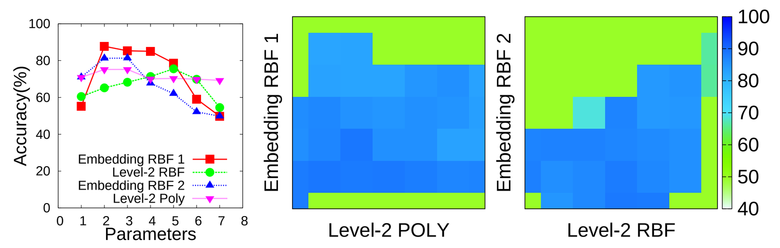


Figure 2: The parameter sensitivity of embedding kernels and level-2 kernels. The heatmaps depict the accuracy at different parameter values.

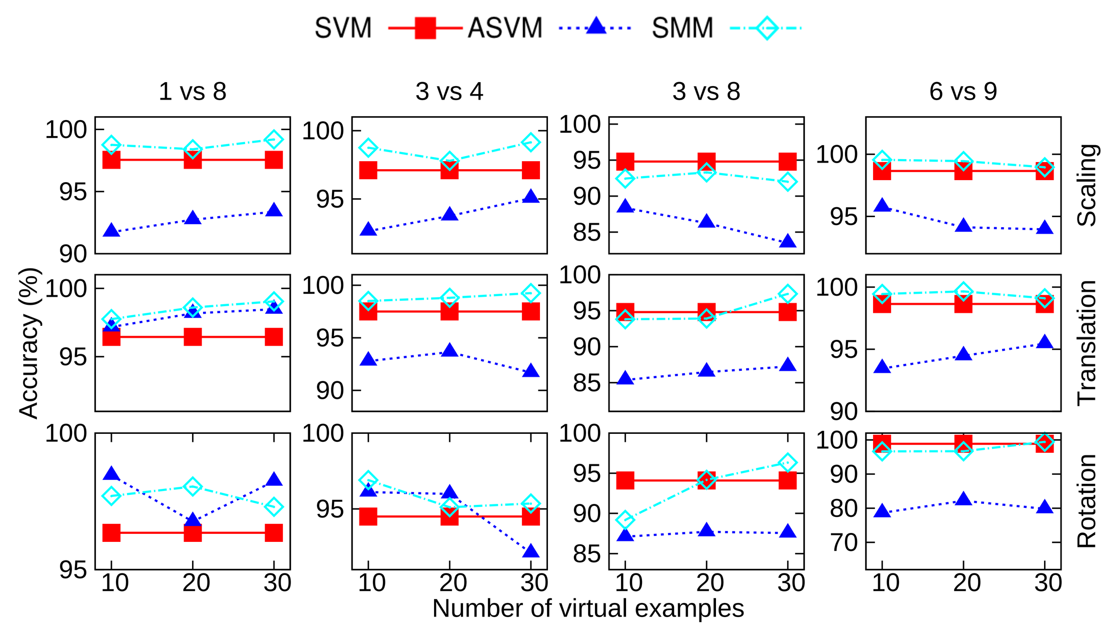


Figure 3: The comparison of SVM, ASVM, and SMM on the USPS handwritten digits dataset. The virtual examples are generated according to three basic operations, namely *scaling*, *translation*, and *rotation*. The pseudo-samples are drawn from the distributions associated with the parameters of each operation. We consider the linear SMM with Gaussian RBF kernel.

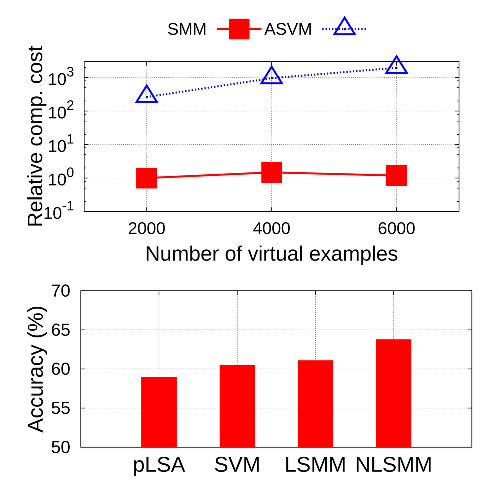


Figure 4: The computational cost on the USPS dataset (top). The results of different approaches on natural scene categorization using the bag-of-word representation (bottom).

The results demonstrate the benefits of distribution-based approach over sample-based approach.

References

[1] B. Schölkopf, R. Herbrich, and A. J. Smola. *A generalized representer theorem.* In COLT'01/EuroCOLT'01, pages 416–426. Springer-Verlag, 2001.

[2] A. Smola, A. Gretton, L. Song, and B. Schölkopf. *A Hilbert space embedding for distributions.* In ALT, pages 13–31. Springer-Verlag, 2007.

[3] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and Gert R. G. Lanckriet. *Hilbert space embeddings and metrics on probability measures.* JMLR, 99:1517–1561, 2010.