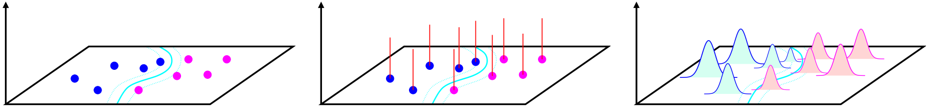
Learning from Distributions via Support Measure Machines

K. Muandet, K. Fukumizu, F. Dinuzzo, B. Schölkopf

MAX-PLANCK-GESELLSCHAFT



$$\mathcal{X} \longrightarrow \mathcal{H} \qquad x \longmapsto k(x,\cdot)$$

$$\Delta \longrightarrow \mathcal{H} \qquad \delta_x \longmapsto \int k(x,\cdot)\,\mathrm{d}\delta_x$$

$$\mathscr{P} \longrightarrow \mathcal{H} \qquad \mathbb{P} \longmapsto \int k(x,\cdot)\,\mathrm{d}\mathbb{P}(x)$$

Potential Applications:

- Uncertain/noisy data (astronomical/biological data)
- Groups of samples (group anomaly, preference learning)
- Changing environments (domain adaptation/generalization)
- Large-scale machine learning (data squashing)

Regularization on Probability Measures

Th54

Given a sample $(\mathbb{P}_1, y_1), (\mathbb{P}_2, y_2), \ldots, (\mathbb{P}_m, y_m)$, any solution f to

$$\ell\left(\mathbb{P}_1, y_1, \mathbb{E}_{\mathbb{P}_1}[f], \ldots, \mathbb{P}_m, y_m, \mathbb{E}_{\mathbb{P}_m}[f]\right) + \Omega\left(\|f\|_{\mathcal{H}}\right) \qquad (1)$$

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

Our framework (1) **is different from**

1. $\mathbb{E}_{\mathbb{P}_1}\mathbb{E}_{\mathbb{P}_2}\ldots\mathbb{E}_{\mathbb{P}_m}\ell(\{x_i, y_i, f(x_i)\}_{i=1}^m) + \Omega(\|f\|_{\mathcal{H}})$ **intracable**
2. $\ell(\{M_i, y_i, f(M_i)\}_{i=1}^m) + \Omega(\|f\|_{\mathcal{H}})$, $M_i = \mathbb{E}_{\mathbb{P}_i}[x]$ **information loss**

Risk Deviation Bound

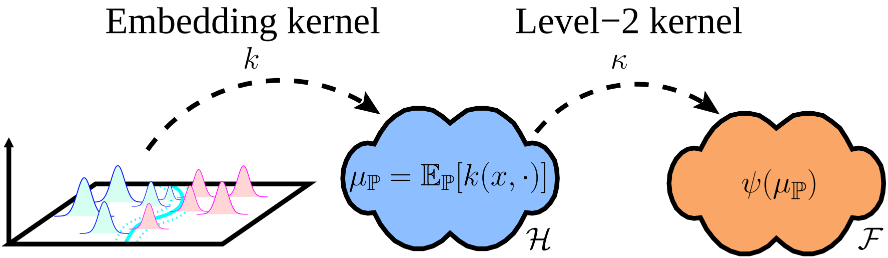
Given a distribution $\mathbb{P}$ with variance σ^2, a Lipschitz continuous function f with constant C_f, a loss function ℓ with constant C_ℓ, it follows for any $y \in \mathbb{R}$ that

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

Information preserving + computationally efficient.

Support Measure Machines (SMM)

Th54



Feature maps

$$K(\delta_x, \delta_y)$$
$$= \langle k(x,\cdot), k(y,\cdot)\rangle_{\mathcal{H}}$$
$$= k(x,y)$$

The SVM is recovered as a special case.

Linear kernels

$$K(\mathbb{P}, \mathbb{Q})$$
$$= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}}\rangle_{\mathcal{H}}$$
$$= \mathbb{E}_{x\sim\mathbb{P}, z\sim\mathbb{Q}}[k(x,z)]$$

It defines the feature for distributions.

Nonlinear kernels

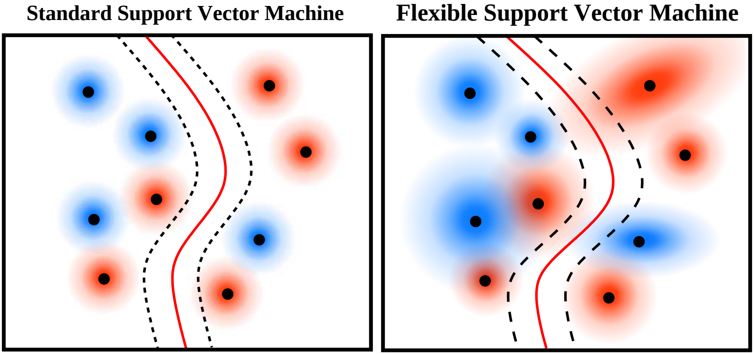
$$K(\mathbb{P}, \mathbb{Q})$$
$$= \kappa(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}})$$
$$= \langle \psi(\mu_{\mathbb{P}}), \psi(\mu_{\mathbb{Q}})\rangle_{\mathcal{F}}$$

It allows for nonlinear learning algorithms.

Flexible Support Vector Machines

Th54

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



$$f = \sum_{i=1}^{n} k(x_i,\cdot) \qquad f = \sum_{i=1}^{n} k_i(x_i,\cdot)$$

The flexible SVM places different kernels on training samples.