It is a long lasting dream of humanity to build a machine that can predict the future. In this work we aim at making a first step towards giving computers such abilities, at least on very short time scales.

We study the situation of a time-varying probability distribution from which sample sets at different time points are observed. Our main result is a method for learning an operator that captures the dynamics of the time-varying data distribution based on two recent machine learning techniques: the embedding of probability distributions into a reproducing kernel Hilbert space and vector-valued regression. By extrapolating the learned dynamics into the future we obtain an estimate of the future distribution in form of a (potentially weighted) set of samples.

Let $Z$ be a data domain and let $d_t(z)$ for $t \in \mathbb{N}$ be a time-varying data distribution over $z \in Z$. At a fixed point of time, $t$, we assume that we have access to sequences of sets, $S_t \{z_1^t, \ldots, z_n^t\}$, for $t = 1, \ldots, T$, that were sampled i.i.d. from the respective distributions, $d_1, \ldots, d_T$. Our goal is to construct a distribution, $d_{T+1}$, that is as close as possible to the so far unobserved $d_{T+1}$. Optionally, we might also want to construct a set, $S$, of samples that are distributed approximately according to $d_{T+1}$. In this work we propose a method for extrapolating the distribution dynamics (EDD) that consists of four steps, see Figure 1 for an illustration.

a) We represent each sample set, $S_t$, as a vector in a Hilbert space through its empirical (kernel mean) embedding [4]

$$S \rightarrow \tilde{\mu}(S), \quad \tilde{\mu}(S) = \frac{1}{|S|} \sum_{z \in S} \phi(z),$$

where $\phi: Z \rightarrow \mathcal{H}$ is the feature map induced by a positive definite kernel function $k: Z \times Z \rightarrow \mathbb{R}$ with associated reproducing kernel Hilbert space (RKHS) $\mathcal{H}$.

b) We learn an operator that reflects the dynamics between the vectors by solving the regularized vector-valued regression problem [3]

$$\min_{\tilde{\mu}_t \in \mathcal{H}} \sum_{t=1}^{T-1} \|\tilde{\mu}_{t+1} - A \tilde{\mu}_t\|_{\mathcal{H}}^2 + \lambda \|A\|_{F}^2$$

for $\lambda \geq 0$. The solution can be written in closed form (see [2])

$$\tilde{\mu} = \sum_{i=1}^{T-1} \tilde{\mu}_{t+1} \sum_{s=1}^{T-1} W_{ts}(\tilde{\mu}_{ts})_{\mathcal{H}},$$

with coefficient matrix $W = (K + \lambda I)^{-1}$, where $K \in \mathbb{R}^{(T-1) \times (T-1)}$ is the kernel matrix with entries $K_{ij} = \langle \tilde{\mu}_i, \tilde{\mu}_j \rangle_{\mathcal{H}}$, and $I$ is the identity matrix of the same size.

c) We extrapolating the distribution dynamics by one step by applying the learned operator, $\tilde{\mu}$, to the last vector in the sequence, $\tilde{\mu}_T$. The resulting prediction, $\tilde{\mu}_{T+1} = A \tilde{\mu}_T$, can be written as a weighted linear combination of the observed distributions,

$$\tilde{\mu}_{T+1} = \sum_{t=2}^{T} \tilde{\mu}_t \tilde{\mu}_t \text{ with } \tilde{\mu}_t = \sum_{s=1}^{T} W_{ts}(\tilde{\mu}_s, \tilde{\mu}_t)_{\mathcal{H}},$$

for $t = 1, \ldots, T - 1$. The coefficients, $\tilde{\mu}_t$, can be computed from the original sample sets by means of only kernel evaluations, because $\langle \tilde{\mu}_t, \tilde{\mu}_t \rangle_{\mathcal{H}} = \frac{1}{|S_t|} \sum_{z_i^t, z_j^t} k(z_i^t, z_j^t)$.

d) Optionally, we use kernel herding [1] to create a new sample set, $S_{T+1}$, for the extrapolated distribution, i.e. $\tilde{\mu}(S_{T+1}) \approx \tilde{\mu}_{T+1}$.

Experiments on synthetic and real data show that EDD is in fact able to extrapolate the distribution dynamics and that this can be used for practical tasks, such as domain adaptation in situations when no training examples from the target distribution are available, not even unlabeled ones.

We acknowledge funding from the ERC under grant agreement no 308036.


