# Isolation Distributional Kernel: A new tool for kernel based anomaly detection

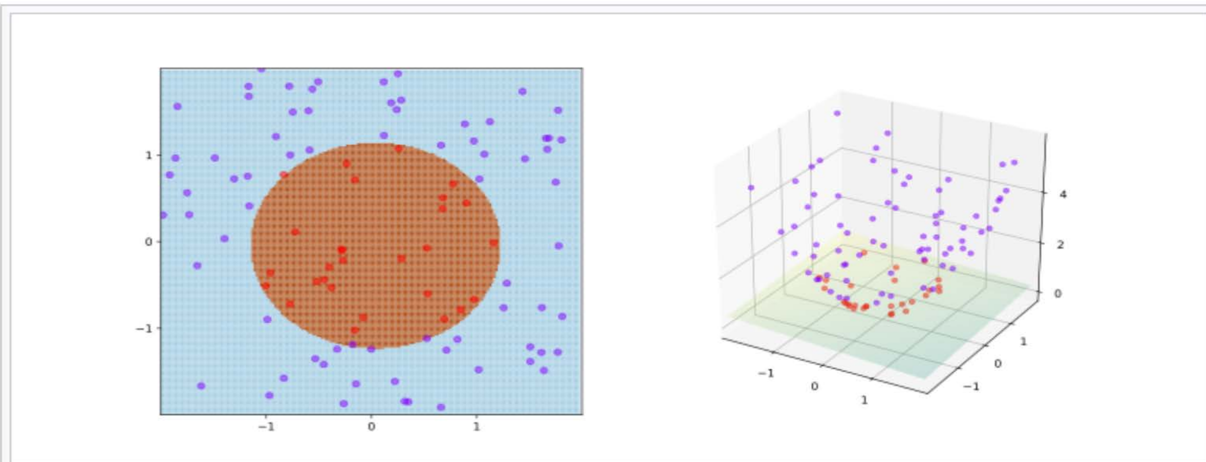
Kai Ming Ting & Bi-Cun Xu, Takashi Washio, Zhi-Hua Zhou
{xubc, zhouzh}@lamda.nju.edu.cn kaiming.ting@federation.edu.au

南京大学 NANJING UNIVERSITY 1902

LAMDA Learning And Mining from DatA
http://lamda.nju.edu.cn

## Kernel Based Learning

In machine learning, kernel methods are a class of algorithms for pattern analysis, whose best known member is the support vector machine (SVM).



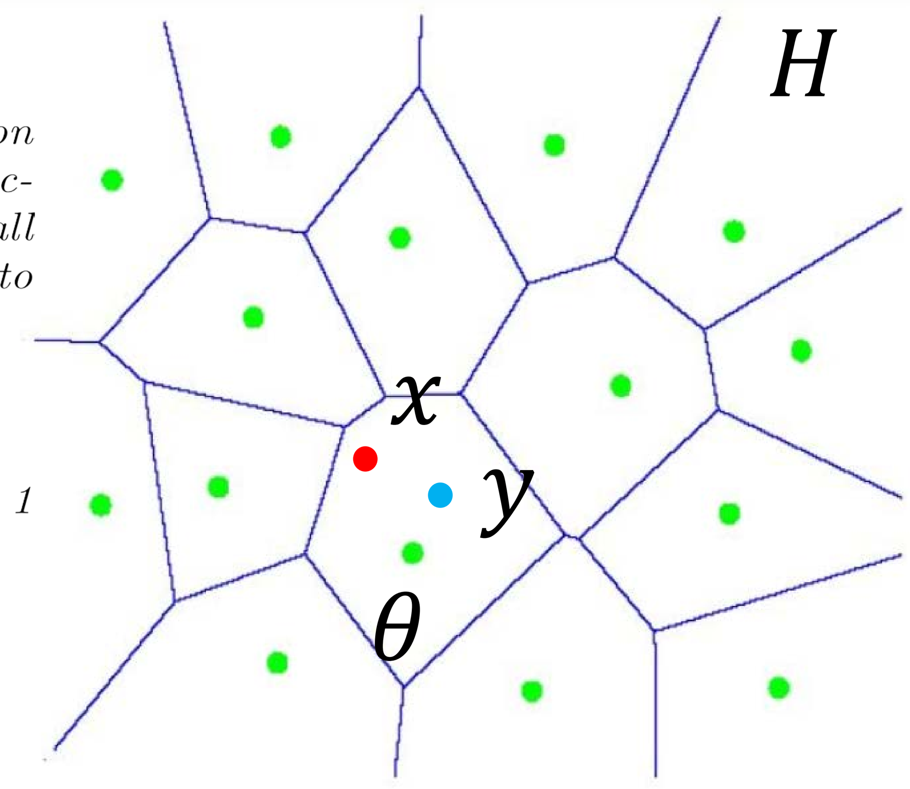
SVM with kernel given by $\varphi((a, b)) = (a, b, a^2 + b^2)$ and thus $K(\mathbf{x}, \mathbf{y}) = \mathbf{x}\cdot\mathbf{y} + \mathbf{x}^2\,\mathbf{y}^2$. The training points are mapped to a 3-dimensional space where a separating hyperplane can be easily found.

## Isolation Kernel

**Definition 1** *For two instances $x, y \in \mathbb{R}^d$, Isolation Kernel of $x$ and $y$ wrt $D$ is defined to be the expectation taken over the probability distribution on all partitioning $H \in \mathcal{H}_\psi(D)$ that both $x$ and $y$ fall into the same isolation partition $\theta \in H$:*

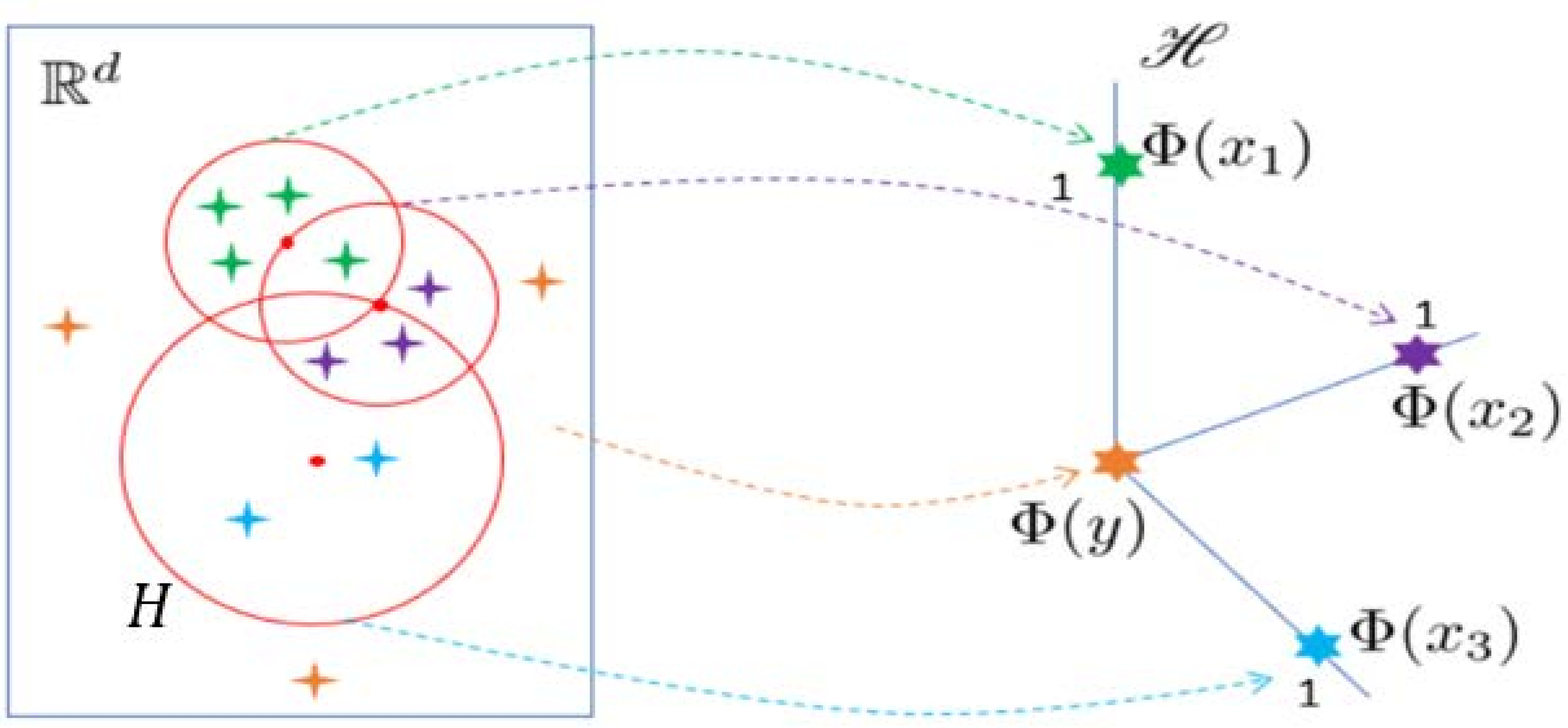
$$K(x, y \mid D) = \mathbb{E}_{\mathcal{H}_\psi(D)}[\mathbb{I}(x, y \in \theta | \theta \in H)]$$

*where $\mathbb{I}(B)$ is the indicator function which output 1 if $B$ is true; otherwise $\mathbb{I}(B) = 0$.*




## New Implementation of Isolation Kernel and Data-Dependent Property

Given a partitioning $H_i$, let $\Phi_i(x)$ be a $\psi$-dimensional binary column vector representing all hyperspheres $\theta_j \in H_i, j = 1, \ldots, \psi$; where $x$ falls into either only one of the $\psi$ hyperspheres or none. The $j$-component of the vector is: $\Phi_{ij}(x) = \mathbb{1}(x \in \theta_j \mid \theta_j \in H_i)$. Given $t$ partitionings, $\Phi(x)$ is the concatenation of $\Phi_1(x), \ldots, \Phi_t(x)$.




Theorem 1. *Given two probability distributions $\mathcal{P}_D, \mathcal{P}_{D'} \in \mathbb{P}$ from which points in datasets $D$ and $D'$ are drawn, respectively. Let $\mathcal{E} \subset \mathcal{X}$ be a region such that $\forall_{w\in\mathcal{E}}, \mathcal{P}_D(w) < \mathcal{P}_{D'}(w)$, i.e., $D$ is sparser than $D'$ in $\mathcal{E}$. Assume that $\psi = |\mathcal{D}|$ is large such that $\tilde{z}$ is the nearest neighbour of $z$, where $z, \tilde{z} \in \mathcal{D} \subset D$ in $\mathcal{E}$, under a given metric distance $\ell$ (the same applies to $z', \tilde{z}' \in \mathcal{D}' \subset D'$ in $\mathcal{E}$.)*

*Isolation Kernel $\kappa_I$ based on hyperspheres $\theta(z) \in H$ has the property that $\kappa_I(x, y \mid D) > \kappa_I(x, y \mid D')$ for any point-pair $x, y \in \mathcal{E}$.*

**For exactly the same two points, the similarity as measured by Isolation Kernel derived from a sparse dataset is larger than that derived from a dense dataset.**

## Isolation Distributional Kernel and Applying to Anomaly Detection

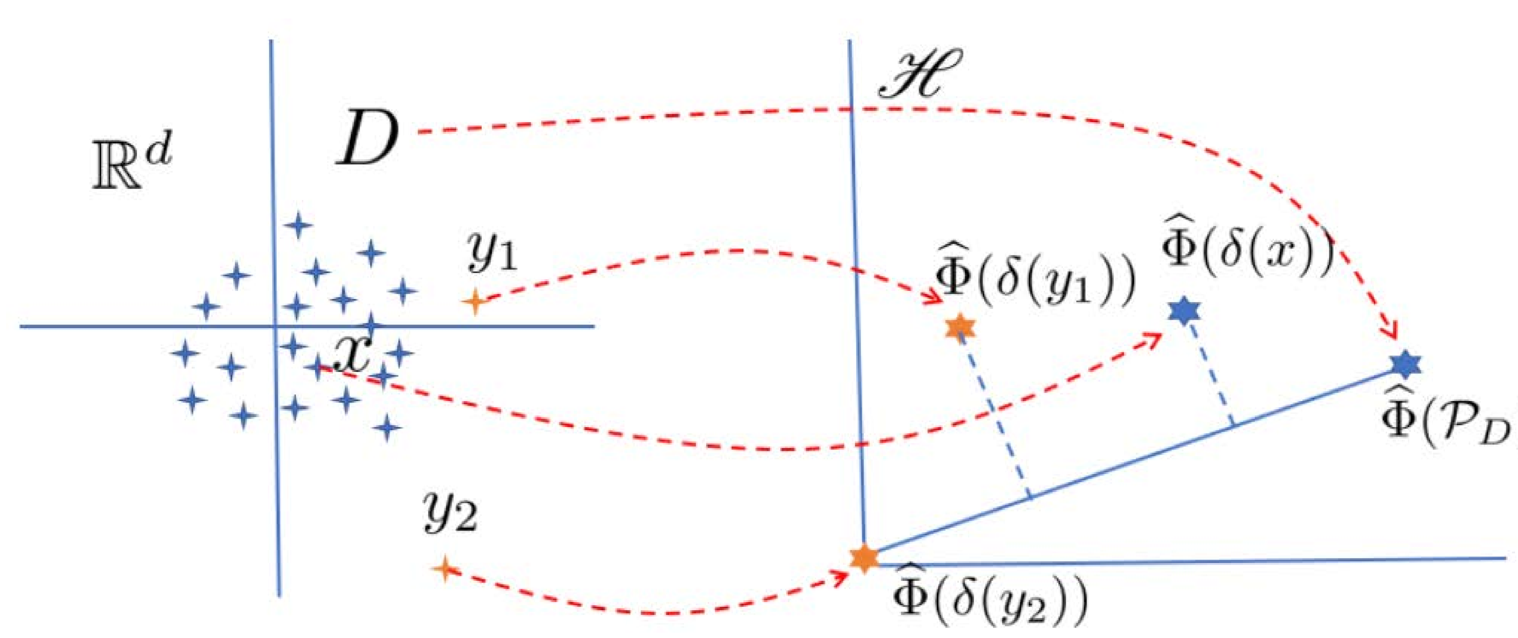
**Definition 1.** *Isolation Distributional Kernel of two distributions $\mathcal{P}_S$ and $\mathcal{P}_T$ is given as:*

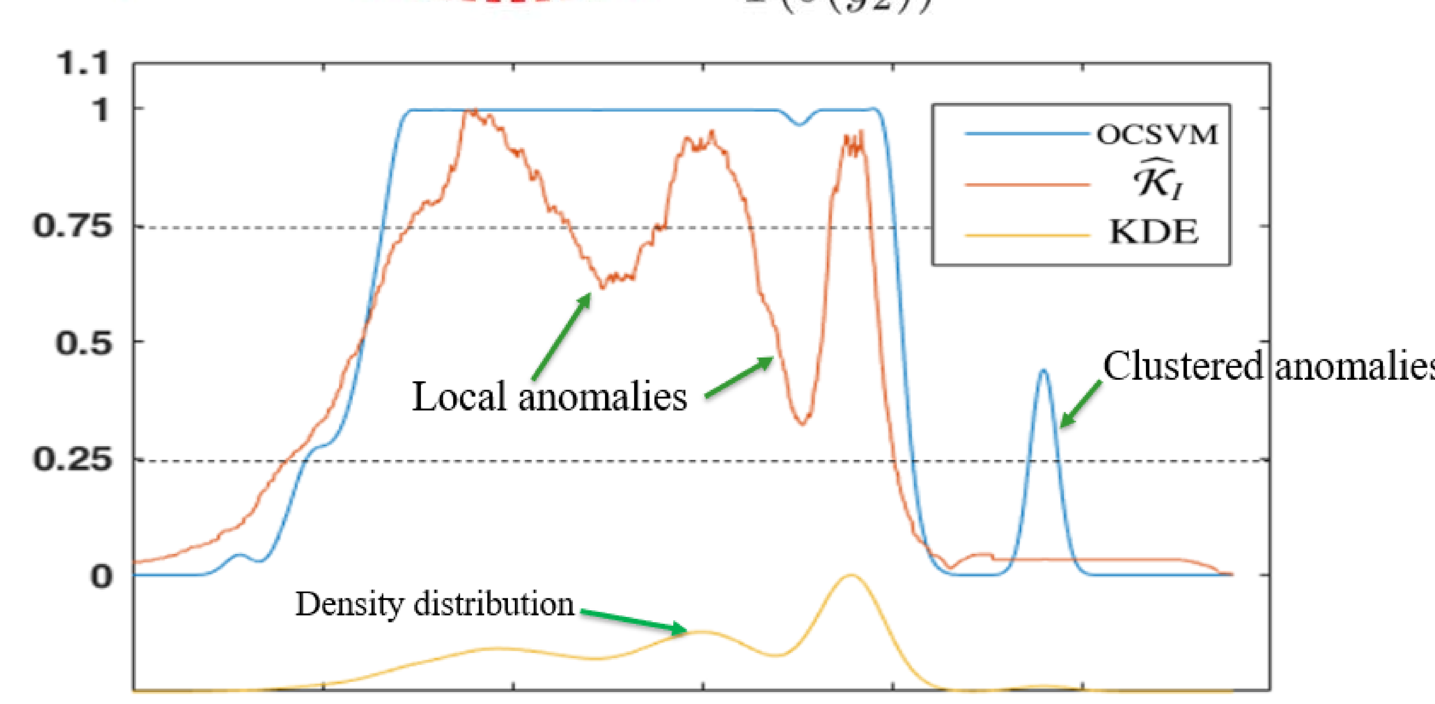
$$\begin{aligned}
\widehat{\mathcal{K}}_I(\mathcal{P}_S, \mathcal{P}_T|D) &= \frac{1}{|S||T|}\sum_{x\in S}\sum_{y\in T}\kappa_I(x, y|D) \\
&= \frac{1}{t|S||T|}\sum_{x\in S}\sum_{y\in T}\langle\Phi(x|D), \Phi(y|D)\rangle \\
&= \frac{1}{t}\left\langle\widehat{\Phi}(\mathcal{P}_S|D), \widehat{\Phi}(\mathcal{P}_T|D)\right\rangle
\end{aligned}$$

*where $\widehat{\Phi}(\mathcal{P}_S|D) = \frac{1}{|S|}\sum_{x\in S}\Phi(x|D)$ is the empirical feature map of the kernel mean embedding; and Isolation Kernel: $\kappa_I(x, y|D) = \frac{1}{t}\langle\Phi(x|D), \Phi(y|D)\rangle$.*

Proposition 1. *Under the conditions on $\mathcal{P}_D, \mathcal{P}_{D'}, D, D'$ and $\mathcal{E}$ of Theorem 1, given two distribution-pairs $\mathcal{P}_S, \mathcal{P}_T \in \mathbb{P}$ where the supports of both $\mathcal{P}_S$ and $\mathcal{P}_T$ are in $\mathcal{E}$, the IDK $\widehat{\mathcal{K}}_I$ based on hyperspheres $\theta(z) \in H$ has the property that $\widehat{\mathcal{K}}_I(\mathcal{P}_S, \mathcal{P}_T|D) > \widehat{\mathcal{K}}_I(\mathcal{P}_S, \mathcal{P}_T|D')$.*

$$\widehat{\mathcal{K}}_I(\delta(x), \mathcal{P}_D) = \frac{1}{t}\left\langle\widehat{\Phi}(\delta(x)), \widehat{\Phi}(\mathcal{P}_D)\right\rangle$$




## Conclusion

- ♠ New Implementation of Isolation Kernel.
- ♠ Prove the data dependent property of the new Isolation Kernel.
- ♠ Validity of the new Isolation Distributional Kernel (IDK): Injective. Isolation kernel is positive definite and a characteristic kernel.
- ♠ IDK is an effective and efficient kernel-based anomaly detector.