ADVERSARIAL ANOMALY DETECTION

Prameesha S. Weerasinghe
Sarah M. Erfani
Tansu Alpcan
Christopher Leckie
Margreta Kuijper

University of Melbourne
Academic Centre for Cyber Security Excellence

caleckie@unimelb.edu.au

What is Machine learning?

- It is a method of data analysis including making decisions such as classification

How does it work?

- Automatically builds an analytical model by using algorithms that iteratively learn from data
- Machine learning allows computers to find hidden features without being explicitly programmed to extract these features.

Why is it popular now?

- Growing volume and variety of available data
- Increased computational capability
- Affordable data storage

SUPERVISED LEARNING

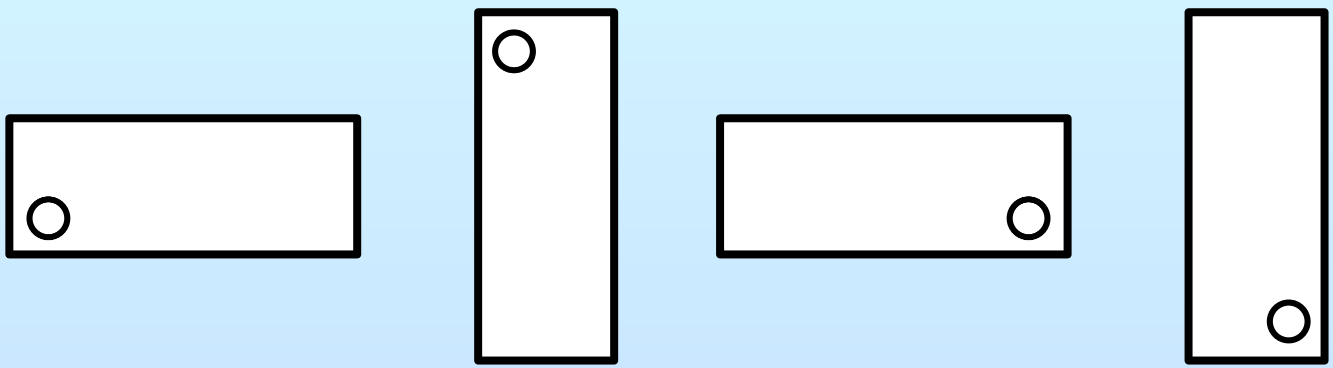
- We give data as well as labels
- The algorithm finds the relationship between the data and the labels - e.g., Classification

UNSUPERVISED LEARNING

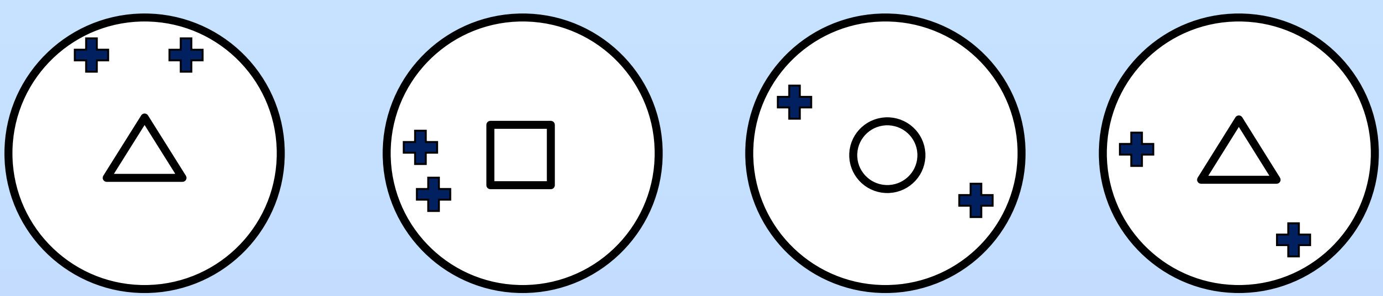
- Data is given without labels
- Algorithm finds patterns in data - e.g., Clustering or Anomaly Detection

Anomaly detection: a general challenge of intelligence?

Spot the odd one out:



a. b. c. d.



a. b. c. d.

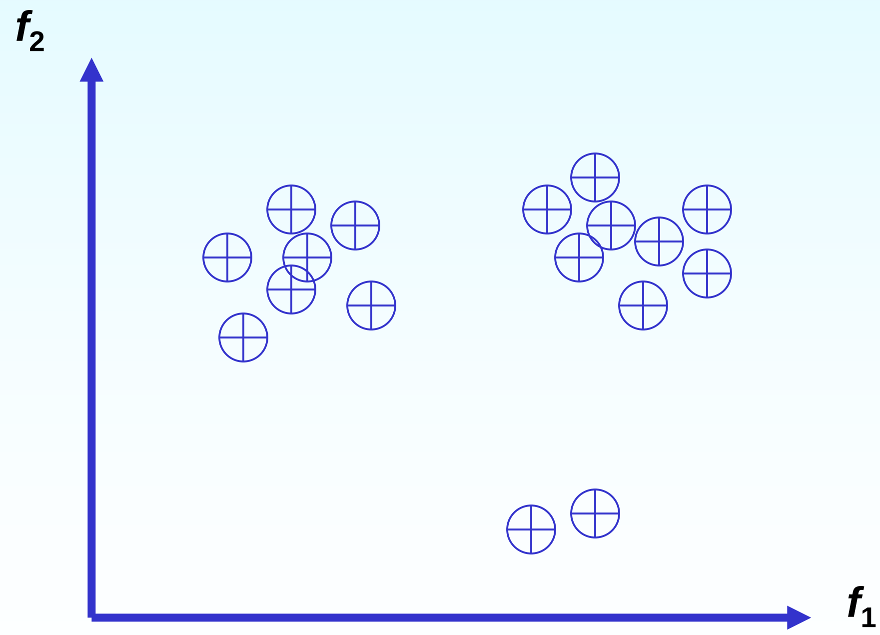
Learning Unusual Patterns (Anomaly Detection)

- Learn a model of “normal” database records
- Use this model to test new records for anomalies
- Any anomalies can be either interesting or errors

Unsupervised Anomaly Detection

[Eskin et al. 2002]

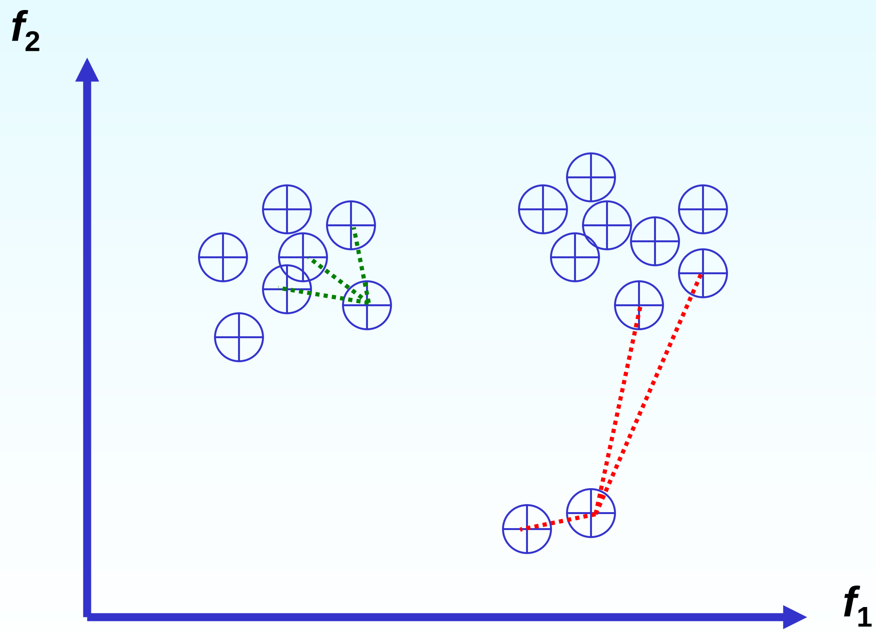
- **Map record fields into a feature space {$f_1 \ldots f_k$}**
- **Cluster similar records**
- **Use large clusters to represent normal records**



Unsupervised Anomaly Detection

K-nearest neighbours:

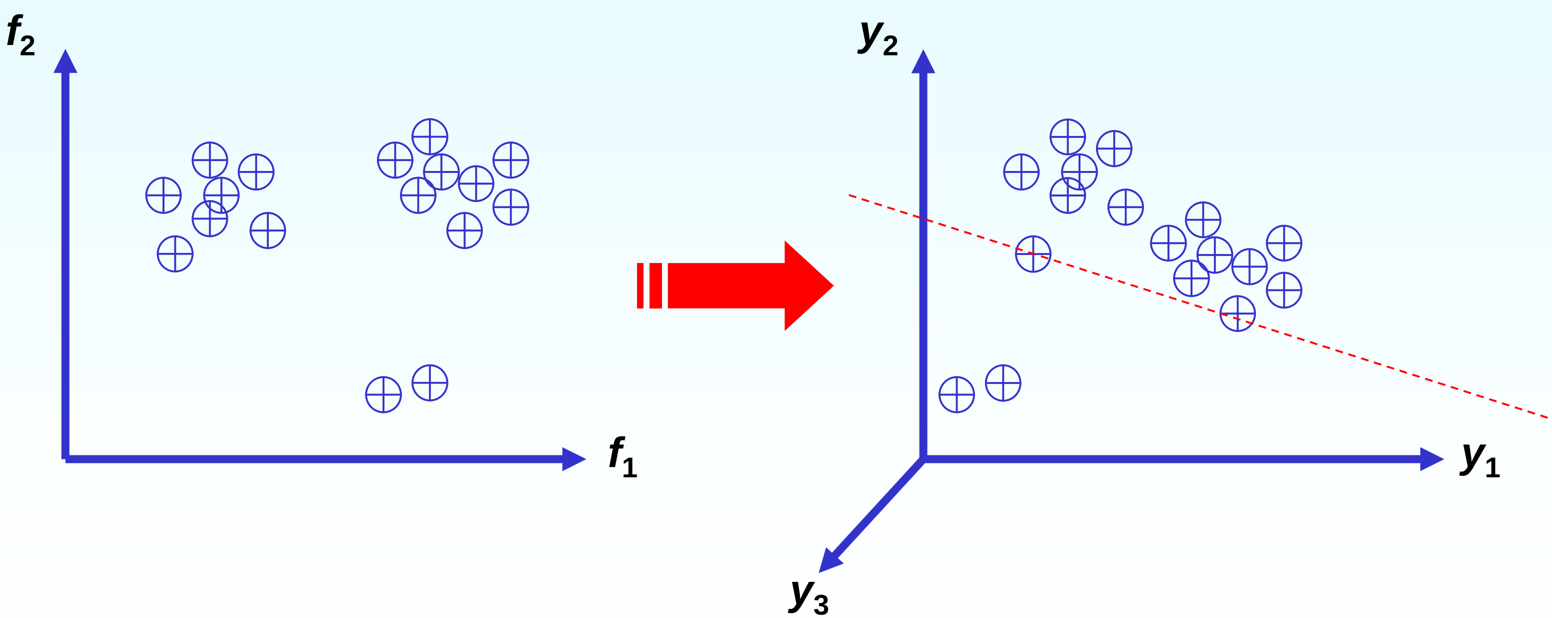
- **Find *k* nearest neighbours of each point**
- **Data points with high kNN distance are in sparse regions of space**



Unsupervised Anomaly Detection

One-class Support Vector Machine:

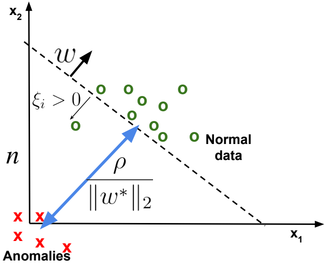
- **Map data points into a higher dimensional space**
- **Find a hyperplane that is *maximally distant* from origin while separating *most points* from origin**



ONE-CLASS SUPPORT VECTOR MACHINES

- An **unsupervised** learning algorithm to **detect anomalies**
- Linearly separates the training data w.r.t. the origin with the highest margin
- The primal optimization problem of OCSVMs is (Schölkopf et al. 2000)

$$\begin{aligned} \min_{w,\ \xi_i,\ \rho} \quad & \frac{1}{2}\|w\|^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & \langle w\,, x_i\rangle \geq \rho - \xi_i,\ \forall i = 1, \ldots, n \\ & \xi_i \geq 0,\ \forall i = 1, \ldots, n \end{aligned} \quad (1)$$



- where $\nu \in (0, 1)$ is the regularization parameter
- take larger value for ν if training set is suspected to be contaminated
- ρ is the offset from the origin
- ξ_i values are the slack variables

ONE-CLASS SUPPORT VECTOR MACHINES

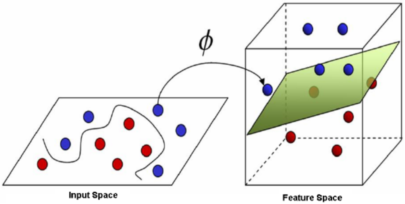
- The dual form of the OCSVM algorithm is (Schölkopf et al. 2000),

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j \langle x_i , x_j \rangle \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \frac{1}{\nu n},\ \forall i = 1, \ldots, n \\ & \sum_{i=1}^{n} \alpha_i = 1 \end{aligned} \tag{2}$$

where α_i are the dual variables

KERNEL TRICK

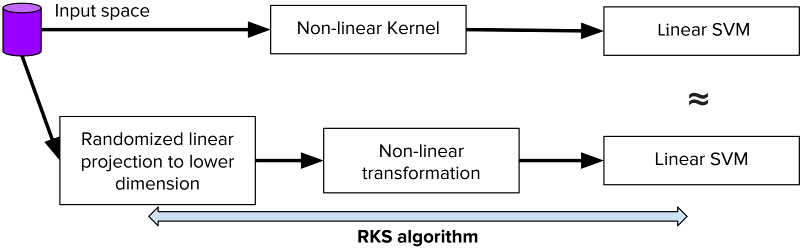
- Suppose input data is not linearly separable
- The original input space is mapped, via function ϕ, to a higher-dimensional feature space where the data is linearly separable
- Explicitly transforming each data point is computationally expensive (especially with high dimensional data)
- As optimization problem 2 uses the dot product between data points, the "kernel trick" can be used for positive definite kernel functions in order to reduce the computational load
  $\langle \phi(x_i)) , \phi(x_j) \rangle = k(x_i, x_j)$
- Time complexity: $\mathcal{O}(dn^2)$ where d is the dimension of input space and n is the number of training data samples



[1] https://www.researchgate.net/figure/260283043_fig13_Figure-A15-The-non-linear-SVM-classifier-with-the-kernel-trick

Alternative to the Kernel Trick

- Rahimi and Recht (2008) introduced *Random Features for Large Scale Machine Learning* in order to reduce the computational load (**RKS algorithm**)
- Map the input data to a randomized low-dimensional space, called feature space, and then apply existing fast linear methods
- Time complexity: $\mathcal{O}(dn)$ where d is the dimension of the feature space



S. Erfani, M. Baktashmotlagh, S. Rajasegarar, S. Karunasekera, C. Leckie, "R1SVM: A randomised nonlinear approach to large-scale anomaly detection" AAAI 2015

ACTIONS OF AN ADVERSARY

Source: Winnetka Animal Hospital

Can they "poison" our model of what is normal?

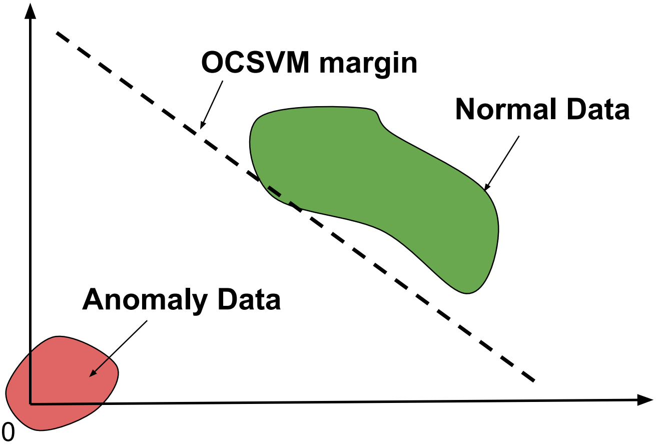
Attack on Integrity

- The ultimate objective of the attacker is to fool the user into labeling anomalies as normal during testing (increase **False Negatives**)
- The attacker would first compromise the classifier by injecting outliers into the training data
- After this, it would be easier for the attacker to craft harmful adversarial data points that are classified by the user as normal data points.
- Learners such as OCSVMs can withstand noise in data
- But are affected when adversaries deliberately distort data

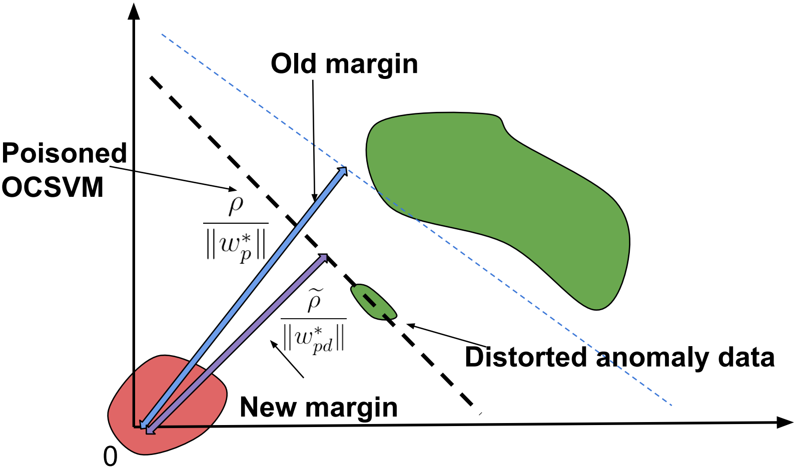
Increasing the Attack Resistance of OCSVMs

- It has been shown that transforming data using the RKS algorithm can create better separated data clouds
- There is a potential for adversarial distortions to have a less impact when data is projected to lower dimensions
- It becomes very difficult for the Adversary to predict the projection matrix because it is chosen randomly

OCSVM - BEFORE ATTACK



OCSVM - AFTER ATTACK

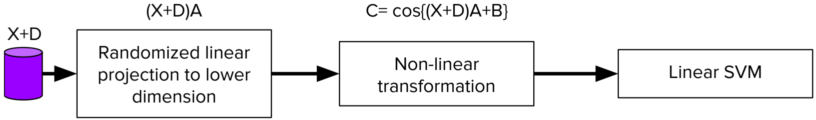


Impact on OCSVM margin

- Let w_p^* be the solution in the projected space **without** adversarial distortions
- Let w_{pd}^* be the solution in the projected space **with** adversarial distortions
- Margin of separation of a OCSVM is given by $\rho/\|w\|_2$
- Which implies that a small weight vector corresponds to a large margin of separation of the attack
- $\|w_p^*\|_2 - \|w_{pd}^*\|_2$ is an indicator of the attack's effectiveness
- As the learner cannot demarcate adversarial distortions from the normal data, it cannot empirically calculate $\|w_p^*\|_2$
- Therefore we derive an upper bound on $\|w_p^*\|_2 - \|w_{pd}^*\|_2$

Details of the RKS algorithm

- Training data - $X \in \mathbb{R}^{n \times d}$
- Adversarial distortions - $D \in \mathbb{R}^{n \times d}$
- Projection matrix - $A \in \mathbb{R}^{d \times r}$, where each element is an i.i.d. $\mathcal{N}(0, 1)$ random variable
- b is a $1 \times r$ row vector where each element is drawn uniformly from $[0, 2\pi]$
- Define B as a $n \times r$ matrix with b in each row
- Define $C \in \mathbb{R}^{n \times r}$ as $C := \cos\big((X + D)A + B\big)$



Assumption 1: Let $D = (d_{ij}) \in \mathbb{R}^{n \times d}$, THEN THE DISTORTIONS MADE BY THE ADVERSARY ARE SMALL S.T. $\cos(d_{ij}) = 1 - \frac{d_{ij}^2}{2}$ HOLDS (I.E., SMALL ANGLE APPROXIMATION)

Theorem 1: If Assumption 1 holds, then the difference between the lengths of the vectors w_p^* and w_{pd}^* are bounded above by

$$\|w_p^*\|_2 - \|w_{pd}^*\|_2 \leq \frac{3\sqrt{r}}{2}. \tag{3}$$

Key message: random projection of data to lower dimensional space limits ability of attacker to poison anomaly detector training!

Conclusions

- OCSVMs are designed to withstand **noise** in training data
- But are vulnerable to malicious **adversarial distortions**
- RKS algorithm was previously used to lower the computational requirements
- Projecting training data to lower dimensional spaces could mask the possible adversarial distortions
- Effectiveness of the adversarial distortions would be reflected on the difference between the margins of separation (after using RKS algorithm)
- We theoretically show that the difference can be reduced by projecting to lower dimensional spaces

P. Weerasinghe, S. Erfani, T Alpcan, C. Leckie, M. Kuijper, "Unsupervised Adversarial Anomaly Detection using One-Class Support Vector Machines," MTNS 2018.

Theorem - High-level proof

- Define $C^X := \cos(XA + B)$, $C^D := \cos(DA)$, $S^X := \sin(XA + B)$ and $S^D := \sin(DA)$
- Let $\tilde{\alpha}$ be the vector achieving the optimal solution in the projected space when adversarial distortions are present. The following is derived when obtaining the dual optimization problem of OCSVMs,

$$\left\|w^*_{pd}\right\|_2 = \left\|\tilde{\alpha}^T C\right\|_2. \tag{4}$$

- Using the cosine angle-sum identity on C (the symbol $\odot$ denotes the Hadamard product for matrices),

$$\left\|w^*_{pd}\right\|_2 = \left\|\tilde{\alpha}^T \left(C^X \odot C^D\right) - \tilde{\alpha}^T \left(S^X \odot S^D\right)\right\|_2. \tag{5}$$

Theorem - High-Level Proof

- From Assumption 1, the constraint conditions of the OCSVM problem and by using small angle approximation, we obtain

$$\left\|w_{pd}^*\right\|_2 \geq \left\|\tilde{\alpha}^T C^X\right\|_2 - \frac{3\sqrt{r}}{2} \tag{6}$$

- Since the optimization problem is a minimization problem the optimal solution for the OCSVM without any distortion (i.e., α^*) would give a value less than or equal to the value given by $\tilde{\alpha}$.

$$\left\|\alpha^{*,T} C^X\right\|_2 \leq \left\|w_{pd}^*\right\|_2 + \frac{3\sqrt{r}}{2}, \tag{7}$$

$$\left\|w_p^*\right\|_2 - \left\|w_{pd}^*\right\|_2 \leq \frac{3\sqrt{r}}{2}. \tag{8}$$

- The learner is able to make the upper bound tighter by reducing the dimensionality of the dataset (i.e., r).