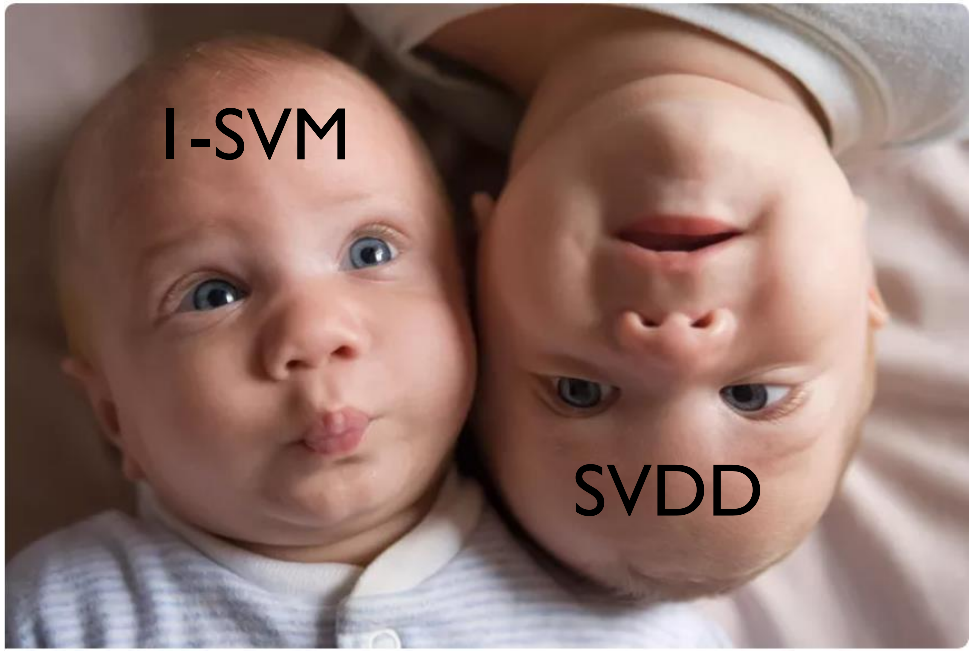
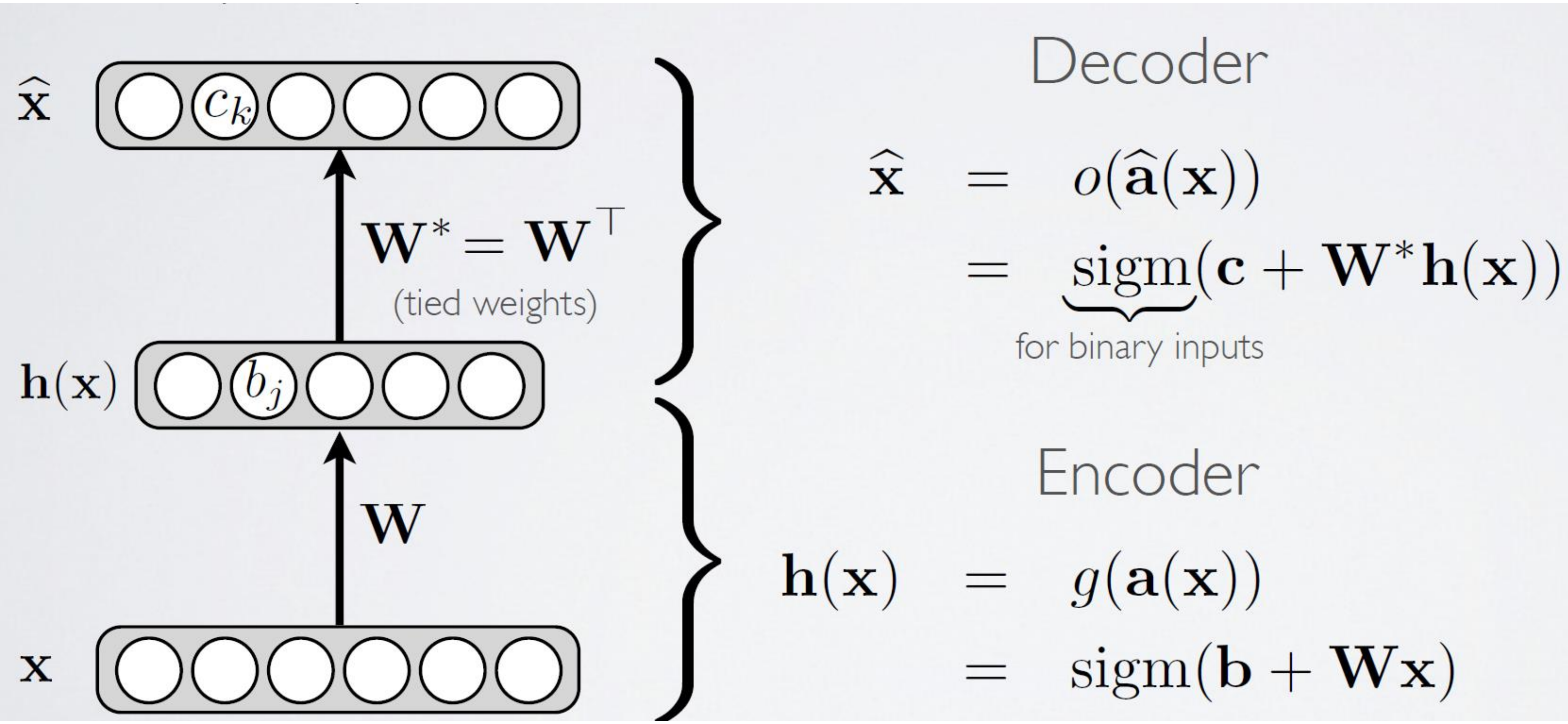



# Anomaly Detection:
# Auto-Encoder, 1-SVM, SVDD

Pilsung Kang
School of Industrial Management Engineering
Korea University

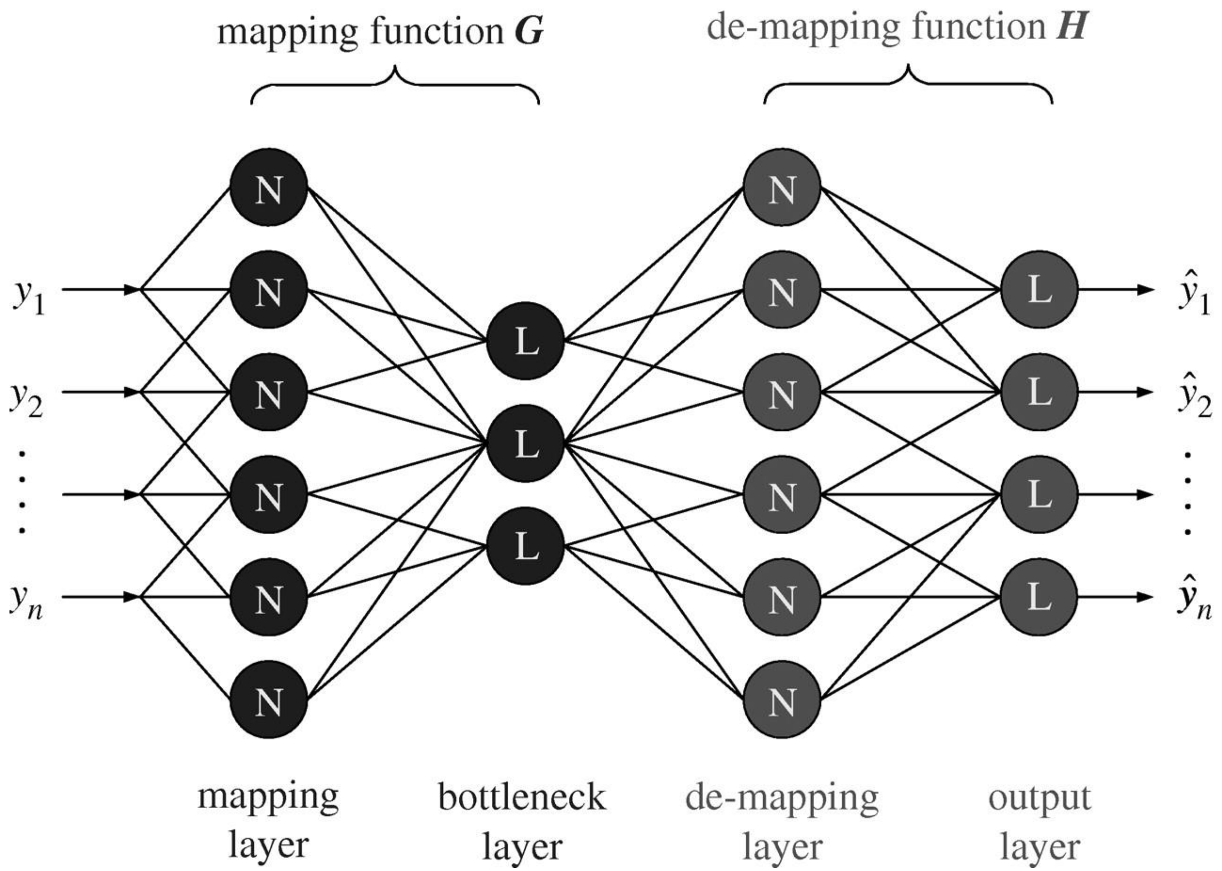
# Auto-Encoder for Anomaly Detection

- Auto-Encoder (Auto-Associative Neural Network)
  - ✓ Feed-forward neural network trained to reproduce its input at the output layer
    - Loss function: $l(f(\mathbf{x})) = \frac{1}{2}\sum_k (\widehat{x}_k - x_k)^2$



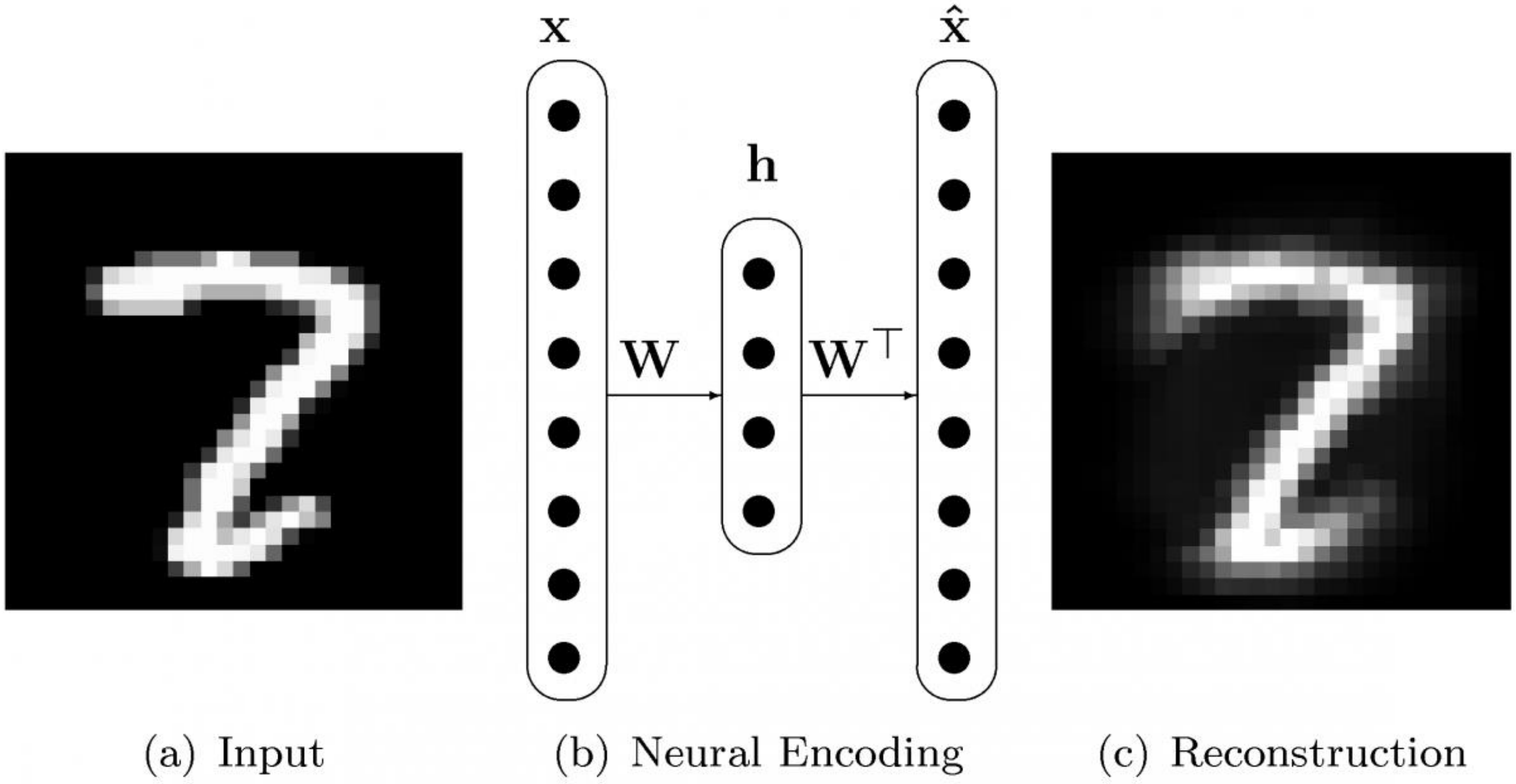
# Auto-Encoder for Anomaly Detection

- Auto Encoder (Auto-Associative Neural Network)
  - ✓ Feed-forward neural network trained to reproduce its input at the output layer
  - ✓ Overcomplete and Undercomplete hidden layers for AE



# Auto-Encoder for Anomaly Detection

- Auto Encoder (Auto-Associative Neural Network)
    - ✓ Example

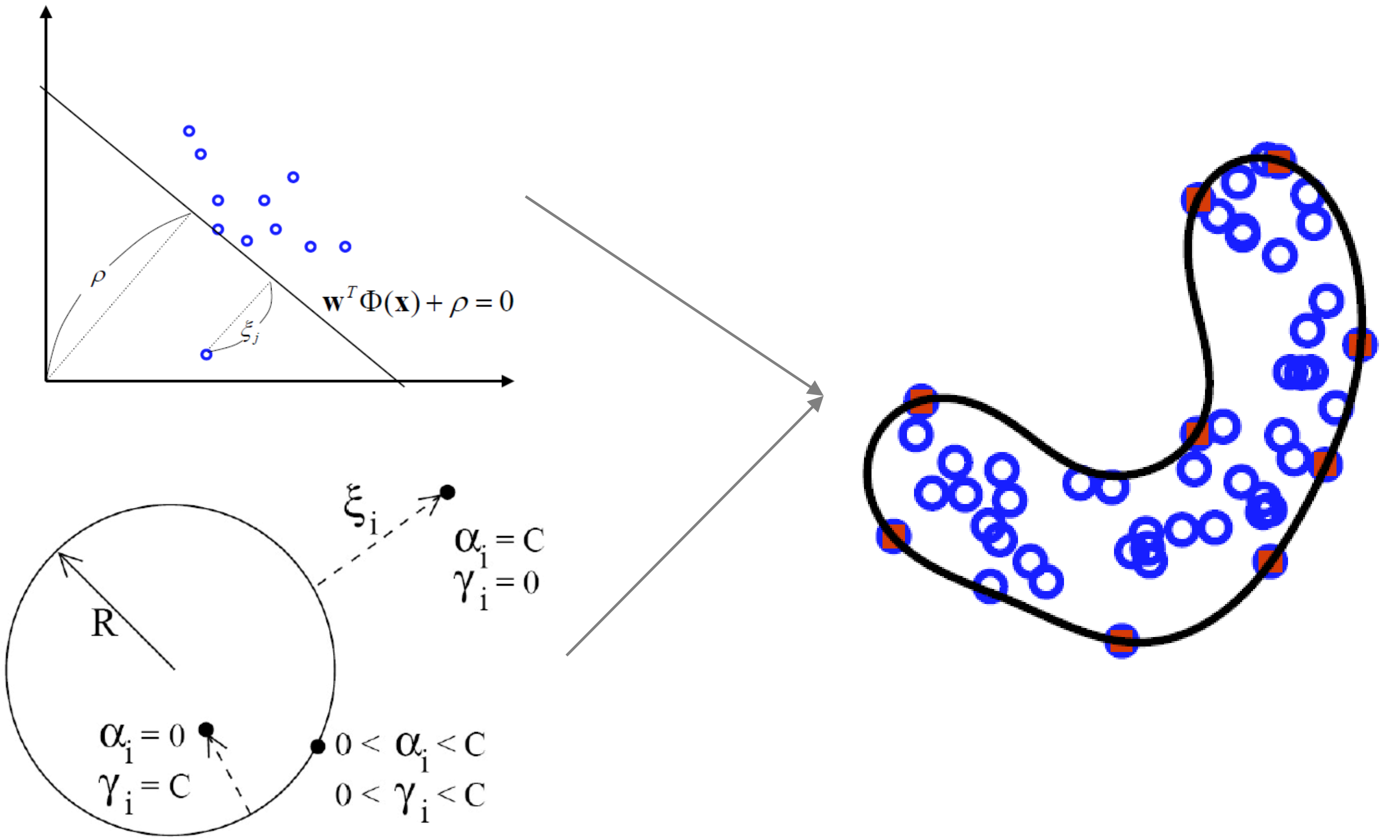


# Support Vector-based Novelty Detection

- Support vector-based novelty detection
  - ✓ Define boundaries of normal regions directly by finding function that separates the normal and abnormal observations

1-SVM

SVDD



# One-Class Support Vector Machine

Scholkopf et al. (2001)

- One-class support vector machine (1-SVM)
  - ✓ Map the data into the feature space corresponding to the kernel and to separate them from the origin with maximum margin
    - Optimization problem

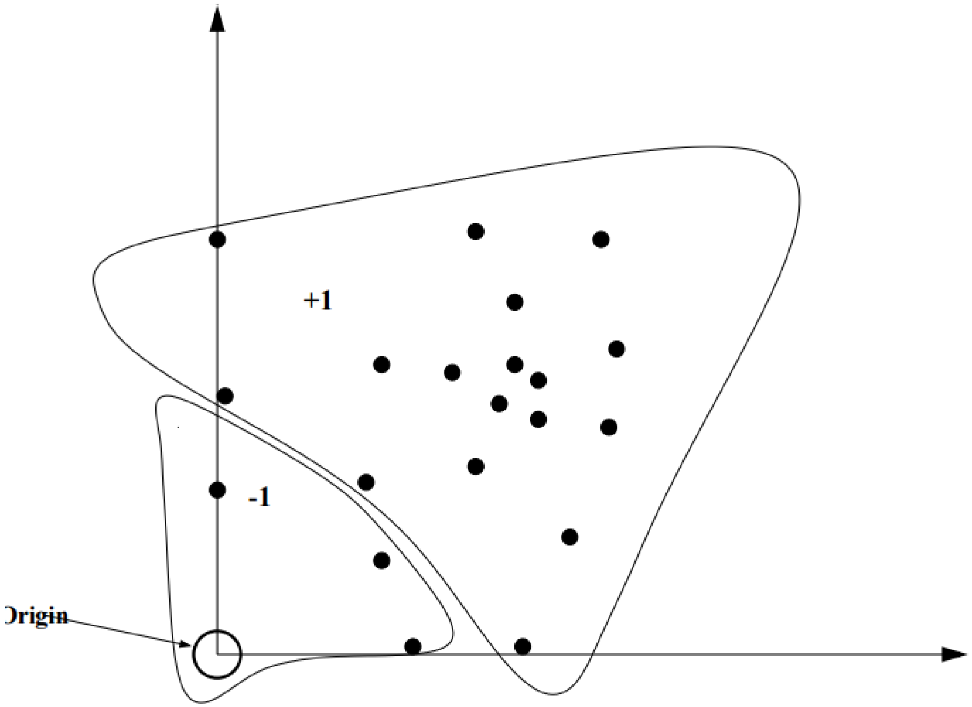
$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 + \frac{1}{\nu l}\sum_{i=1}^{l}\xi_i - \rho$$

$$s.t. \quad \mathbf{w}\cdot\Phi(\mathbf{x}_i) \geq \rho - \xi_i$$

$$i = 1, 2, \cdots, l, \quad \xi_i \geq 0$$

- Decision function

$$f(\mathbf{x}_i) = sign\big(\mathbf{w}\cdot\Phi(\mathbf{x}_i) - \rho\big)$$



# One-Class Support Vector Machine

- One-class support vector machine (1-SVM)
  - ✓ Primal Lagrangian problem (Minimize)

$$L = \frac{1}{2}||\mathbf{w}||^2 + \frac{1}{\nu l}\sum_{i=1}^{l}\xi_i - \rho - \sum_{i=1}^{l}\alpha_i\big(\mathbf{w}\cdot\Phi(\mathbf{x}_i) - \rho + \xi_i\big) - \sum_{i=1}^{l}\beta_i\xi_i$$

  - ✓ KKT condition

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{l}\alpha_i\Phi(\mathbf{x}_i) = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{l}\alpha_i\Phi(\mathbf{x}_i)$$

$$\frac{\partial L}{\partial \xi_i} = \frac{1}{\nu l} - \alpha_i - \beta_i = 0 \quad \Rightarrow \quad \alpha_i = \frac{1}{\nu l} - \beta_i$$

$$\frac{\partial L}{\partial \rho} = -1 + \sum_{i=1}^{l}\alpha_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{l}\alpha_i = 1$$

# One-Class Support Vector Machine

- One-class support vector machine (1-SVM)
  - ✓ Dual Lagrangian problem (Maximize)

$$L = \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j) + \frac{1}{\nu l}\sum_{i=1}^{l}\xi_i - \rho$$

$$-\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j) + \rho\sum_{i=1}^{l}\alpha_i - \sum_{i=1}^{l}\alpha_i\xi_i - \sum_{i=1}^{l}\beta_i\xi_i$$

  - ✓ We should solve

$$\min \;\; L = \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j)$$

$$s.t. \;\; \sum_{i=1}^{l}\alpha_i = 1, \quad 0 \le \alpha_i \le \frac{1}{\nu l}$$

# One-Class Support Vector Machine

- One-class support vector machine (1-SVM)
  - ✓ Employ Kernel Trick for a non-linear mapping

$$\min \; L = \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j)$$

$$s.t. \; \sum_{i=1}^{l}\alpha_i = 1, \quad 0 \leq \alpha_i \leq \frac{1}{\nu l}$$

$$\min \; L = \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

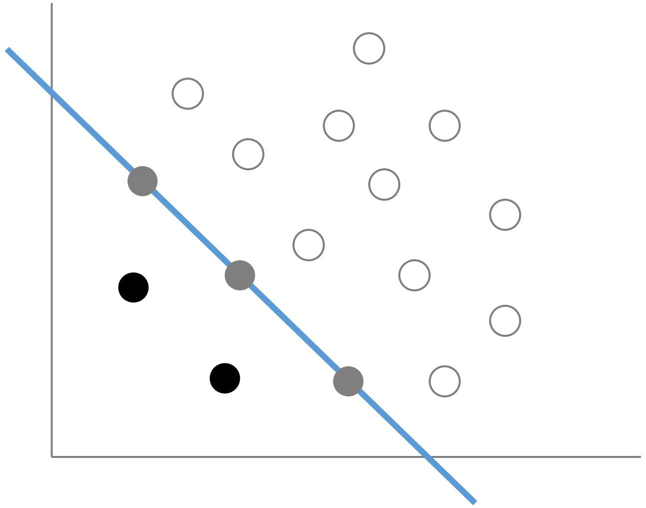
$$s.t. \; \sum_{i=1}^{l}\alpha_i = 1, \quad 0 \leq \alpha_i \leq \frac{1}{\nu l}$$

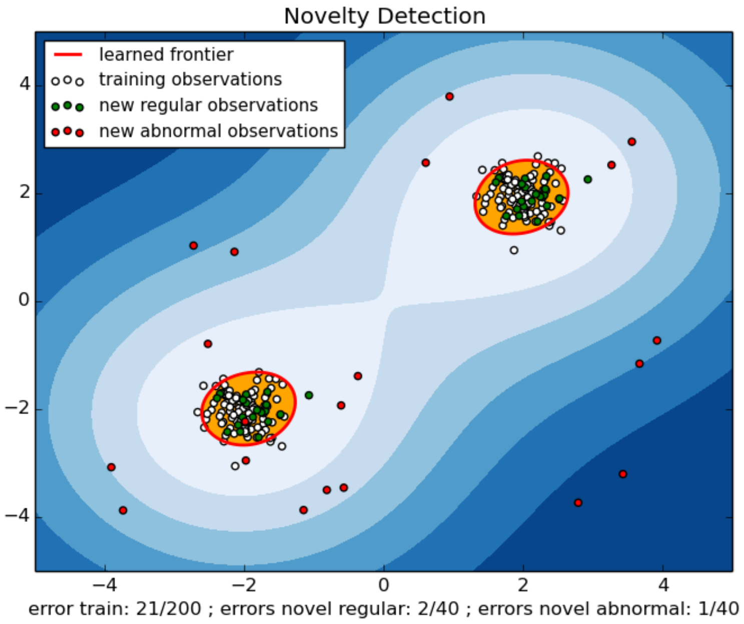
Some possible kernels $K(\cdot,\cdot)$:

$K(x, x_i) = x_i^T x$ (linear SVM)
$K(x, x_i) = (x_i^T x + \tau)^d$ (polynomial SVM of degree $d$)
$K(x, x_i) = \exp(-\|x - x_i\|_2^2/\sigma^2)$ (RBF kernel)
$K(x, x_i) = \tanh(\kappa x_i^T x + \theta)$ (MLP kernel)

# One-Class Support Vector Machine

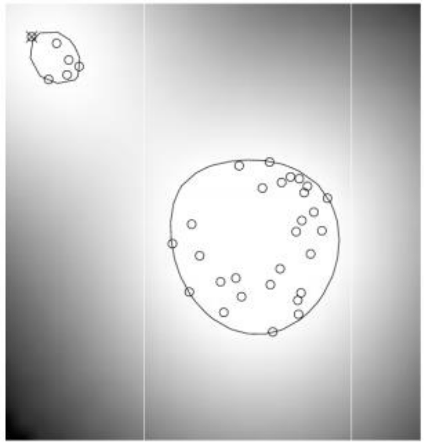
- One-class support vector machine (1-SVM)
  - ✓ Location of a point w.r.t. $\alpha_i$
    - Case 1: $\alpha_i = 0$ $\Rightarrow$ a non-support vector
    - Case 2: $\alpha_i = \frac{1}{\nu l}$ $\Rightarrow$ $\beta_i = 0$ $\Rightarrow$ $\xi_i > 0 \Rightarrow$ Support vector (outsider the hyperplane)
    - Case 3: $0 < \alpha_i < \frac{1}{\nu l}$ $\Rightarrow$ $\beta_i > 0$ $\Rightarrow$ $\xi_i = 0 \Rightarrow$ Support vector (on the hyperplane)

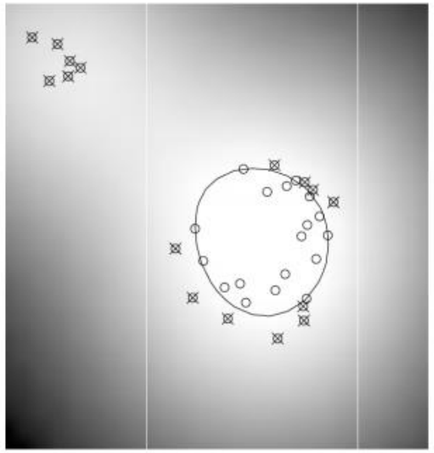




# One-Class Support Vector Machine

- One-class support vector machine (1-SVM)
  - ✓ The role of $\nu$
    - The maximum possible value of $\alpha_i = \frac{1}{\nu l}$
    - At least $\nu l$ support vectors exist
    - At most $\nu l$ support vectors can be located outside the hyperplane
    - Thus, $\nu$ is the lower bound for the fraction of support vectors and the upper bound for the fraction of errors
  - ✓ The higher the $\nu$, the more complex decision boundary is generated



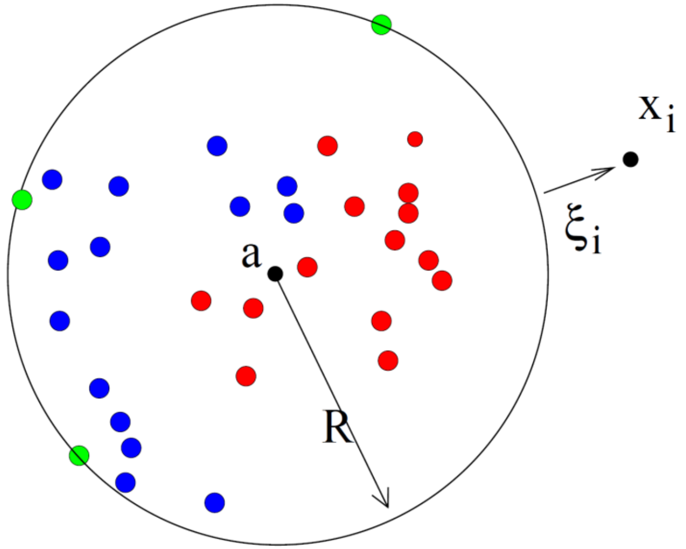


# Support Vector Data Description

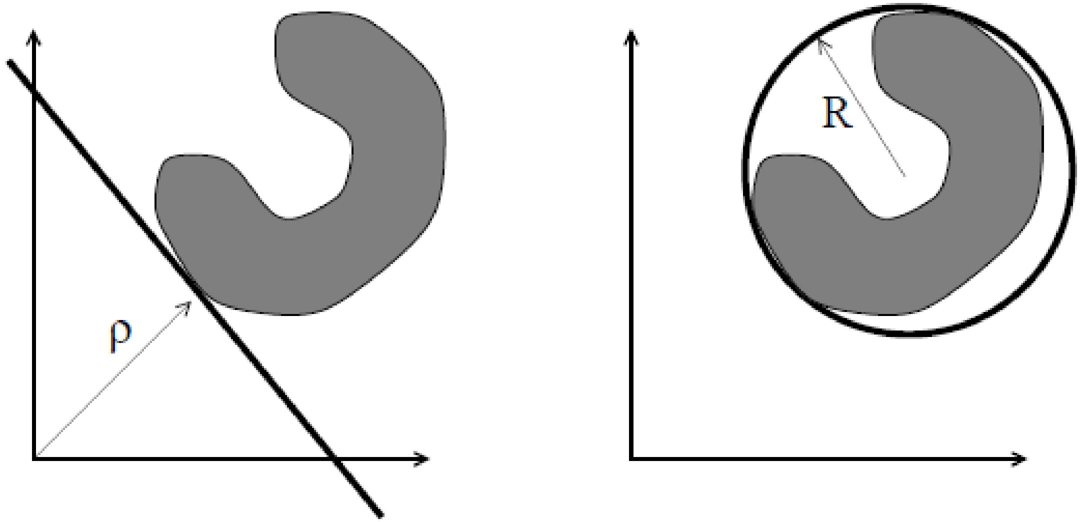
Tax and Duin (2004)

- Support Vector Data Description (SVDD)
  - ✓ Find a hypersphere enclosing all the normal instances in a feature space

SVDD




1-SVM vs. SVDD



# Support Vector Data Description

- Support Vector Data Description (SVDD)
  - ✓ Find a hypersphere enclosing all the normal instances in a feature space
    - Optimization function

$$\min_{R,\mathbf{a},\xi_i} R^2 + C\sum_{i=1}^{l}\xi_i$$

$$s.t. \quad ||\Phi(\mathbf{x}_i) - \mathbf{a}||^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0, \quad \forall i.$$

    - Decision function

$$f(\mathbf{x}) = sign\big(R^2 - ||\Phi(\mathbf{x}_i) - \mathbf{a}||^2\big)$$

# Support Vector Data Description

- Support Vector Data Description (SVDD)
  - ✓ Find a hypersphere enclosing all the normal instances in a feature space
    - Primal Lagrangian problem (Minimization)

$$L = R^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i\Big(R^2 + \xi_i - (\Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_i) - 2\cdot\mathbf{a}\cdot\Phi(\mathbf{x}_i) + \mathbf{a}\cdot\mathbf{a})\Big) - \sum_{i=1}^{l}\beta_i\xi_i$$

$$\alpha_i \geq 0, \quad \beta_i \geq 0$$

    - KKT condition

$$\frac{\partial L}{\partial R} = 2R - 2R\sum_{i=1}^{l}\alpha_i = 0 \quad \Rightarrow \sum_{i=1}^{l}\alpha_i = 1$$

$$\frac{\partial L}{\partial \mathbf{a}} = 2\sum_{i=1}^{l}\alpha_i\cdot\Phi(\mathbf{x}_i) - 2\mathbf{a}\sum_{i=1}^{l}\alpha_i = 0 \quad \Rightarrow \quad \mathbf{a} = \sum_{i=1}^{l}\alpha_i\cdot\Phi(\mathbf{x}_i)$$

$$\frac{\partial L}{\xi_i} = C - \alpha_i - \beta_i = 0 \quad \forall i$$

고려대학교 KOREA UNIVERSITY



DSBA Data Science & Business Analytics

# Support Vector Data Description

- Support Vector Data Description (SVDD)
  - ✓ Find a hypersphere enclosing all the normal instances in a feature space
    - Dual Lagrangian problem (Maximization)

$$L = R^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i\left(R^2 + \xi_i - (\Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_i) - 2\cdot\mathbf{a}\cdot\Phi(\mathbf{x}_i) + \mathbf{a}\cdot\mathbf{a})\right) - \sum_{i=1}^{l}\beta_i\xi_i$$

$$L = R^2 - R^2\sum_{i=1}^{l}\alpha_i + \sum_{i=1}^{l}\xi_i(C - \alpha_i - \beta_i)$$

$$+\sum_{i=1}^{l}\alpha_i\Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_i) - 2\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j) + \sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j)$$

$$L = \sum_{i=1}^{l}\alpha_i\Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_i) - \sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j) \qquad (0 \le \alpha_i \le C)$$

# Support Vector Data Description

- Support Vector Data Description (SVDD)
  - ✓ Find a hypersphere enclosing all the normal instances in a feature space
    - Dual Lagrangian problem (Maximization)

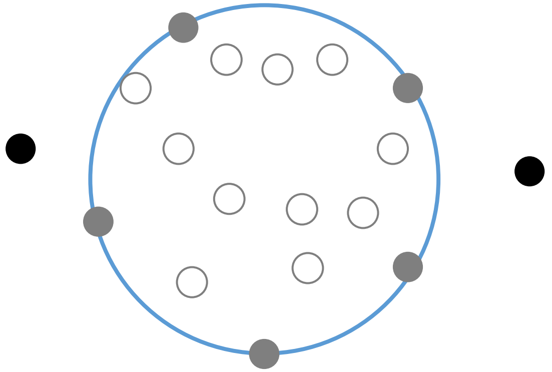
$$L = \sum_{i=1}^{l} \alpha_i \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_i) - \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j \Phi(\mathbf{x}_i) \Phi(\mathbf{x}_j) \qquad (0 \leq \alpha_i \leq C)$$

    - Dual Lagrangian problem (Minimization)

$$L = \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j \Phi(\mathbf{x}_i) \Phi(\mathbf{x}_j) - \sum_{i=1}^{l} \alpha_i \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_i) \qquad (0 \leq \alpha_i \leq C)$$

# Support Vector Data Description

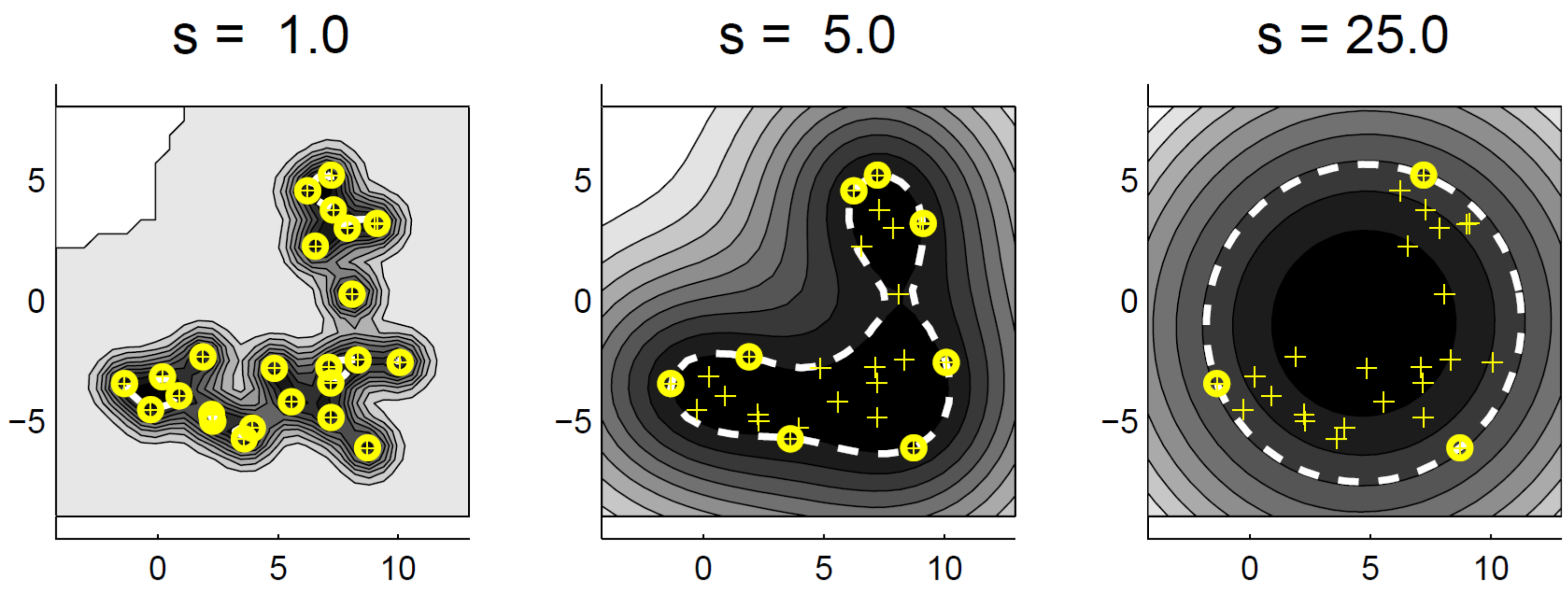
- Support Vector Data Description (SVDD)
  - ✓ Location of a point w.r.t. $\alpha_i$
    - Case 1: $\alpha_i = 0 \quad \Rightarrow$ a non-support vector
    - Case 2: $\alpha_i = C \quad \Rightarrow \quad \beta_i = 0 \quad \Rightarrow \quad \xi_i > 0 \Rightarrow$ Support vector (outsider the hypersphere)
    - Case 3: $0 < \alpha_i < C \quad \Rightarrow \quad \beta_i > 0 \quad \Rightarrow \quad \xi_i = 0 \Rightarrow$ Support vector (on the hypersphere)



# Support Vector Data Description

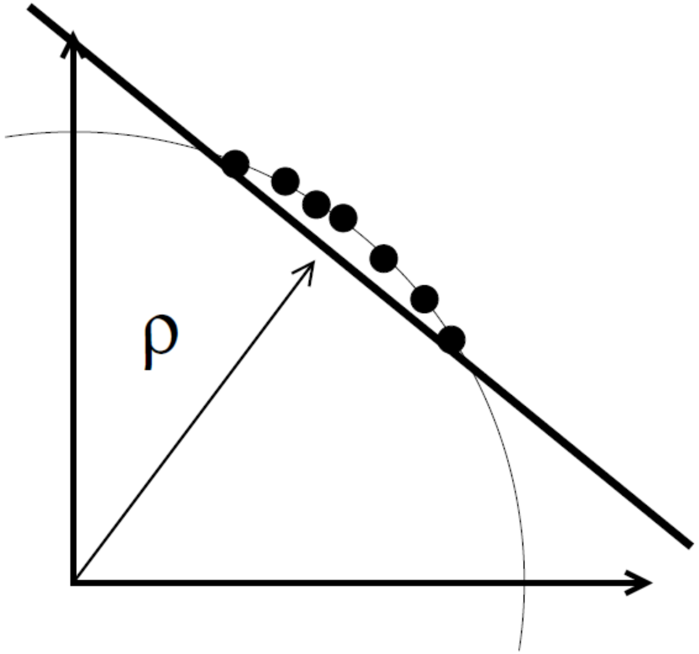
- Support Vector Data Description (SVDD)
  - ✓ SVDD with Gaussian (RBF) kernels

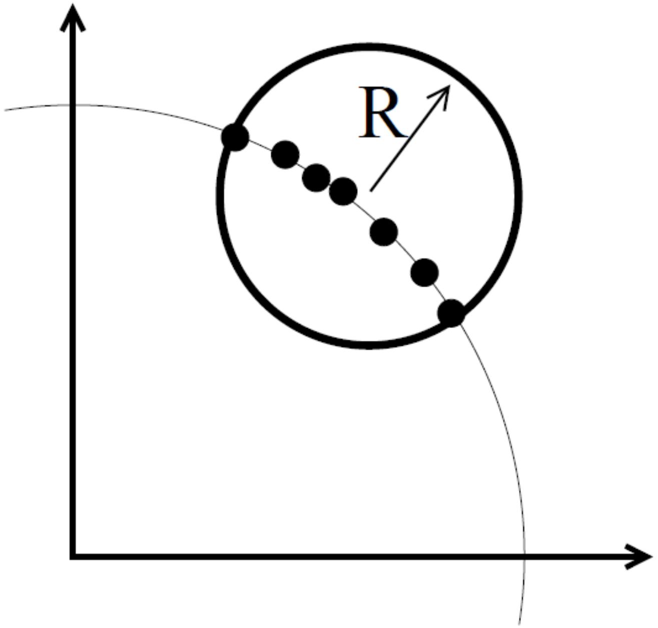
$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{s^2}\right)$$



# Support Vector Data Description

- Support Vector Data Description (SVDD)
  - ✓ When all data is normalized to unit norm vector, SVDD and 1-SVM are equivalent

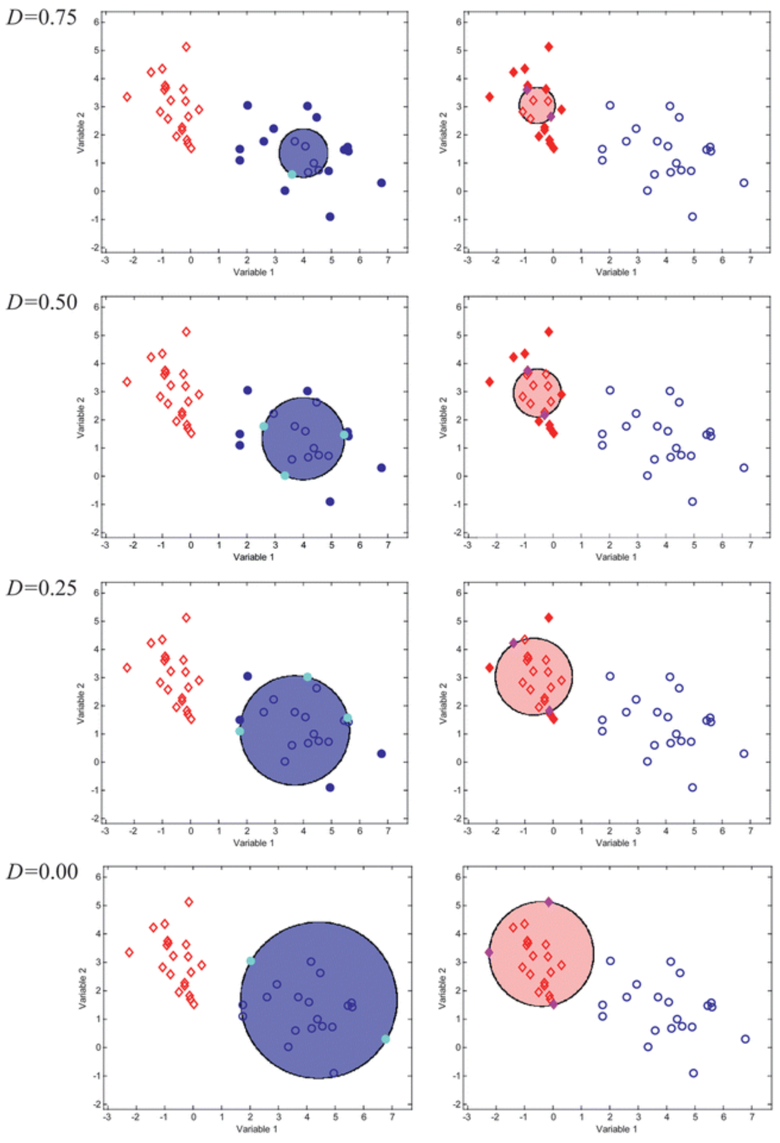



  - For the detailed proof, please refer to Tax (2001) pp. 39-41.

# Support Vector Data Description

- Support Vector Data Description (SVDD)
  - ✓ As in 1-SVM, $\upsilon$-SVDD can also be formulated
  - ✓ D in the right figure is $\upsilon$




http://pubs.rsc.org/-/content/articlehtml/2010/an/b918972f


Q & A

# References

Research Papers


- Breunig, M.M., Kriegel, H.-P., Ng, R.T., and Sander, J. (2000). LOF: Identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data: 93-104.
- Chalapathy, R., Menon, A. K., & Chawla, S. (2018). Anomaly detection using one-class neural networks. arXiv preprint arXiv:1802.06360.
- Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys 41(3): 15.
- Harmeling, S. Dornhege, G., Tax, D. Meinecke, F., and Muller, K-R. (2006). From outliers to prototype: Ordering data. Neurocomputing 69(13-15): 1608-1618.
- Hariri, S., Kind, M. C., & Brunner, R. J. (2018). Extended Isolation Forest. arXiv preprint arXiv:1811.02141.
- Kang, P. and Cho, S. (2009). A hybrid novelty score and its use in keystroke dynamics-based user authentication. Pattern Recognition 42(11): 3115-3127.
- Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on (pp. 413-422). IEEE.
- Liu, F. T., Ting, K. M., & Zhou, Z. H. (2012). Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(1), 3.
- Oza, P., & Patel, V. M. (2018). One-class convolutional neural network. IEEE Signal Processing Letters, 26(2), 277-281.
- Perera, P., Nallapati, R., & Xiang, B. (2019). Ocgan: One-class novelty detection using gans with constrained latent representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2898-2906).

# References

Research Papers

- Pimentel, M.A.F., Clifton, D.A., Clifton, L., and Tarassenko, L. (2014). A review of novelty detection, Signal Processing 99: 215-249.
- Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017, June). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In International Conference on Information Processing in Medical Imaging (pp. 146-157). Springer, Cham.
- Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., and Williamson, R. C.(2001). Estimating the support of a high-dimensional distribution. Neural computation 13(7): 1443-1471.
- Tax, D.M. (2001). One-class classification, Ph.D. Thesis, Delft University of Technology, Netherlands.
- Tax, D.M. and Duin, R.P. (2004). Support vector data description. Machine learning 54(1): 45-66.

Other materials

- Pages 28-33 & 36: http://research.cs.tamu.edu/prism/lectures/pr/pr_l7.pdf
- Figures in Auto-encoder section: https://dl.dropboxusercontent.com/u/19557502/6_01_definition.pdf
- Gramfort, A. (2016). Anomaly/Novelty detection with scikit-learn: https://www.slideshare.net/agramfort/anomalynovelty-detection-with-scikitlearn