Master Degree in Artificial Intelligence for Science and Technology

Anomaly Detection: Additional Algorithms

Fabio Stella

Department of Informatics, Systems and Communication

University of Milan-Bicocca

fabio.stella@unimib.it

OUTLOOK

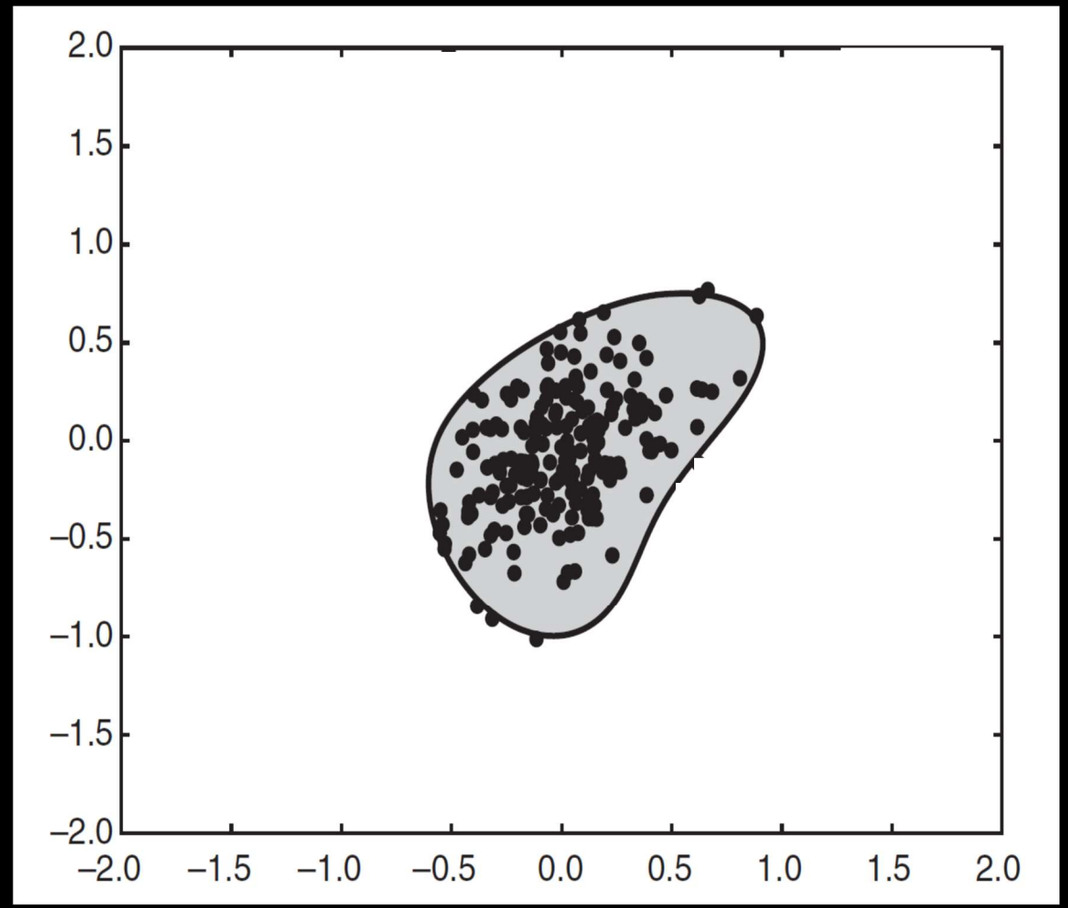
- One Class SVM
- Information Theoretic Approaches
- Evaluation of Anomaly Detection

POINT ANOMALY DETECTION: ONE CLASS SVM

- uses an SVM approach to classify normal objects
- uses the given data to construct such a model
- this data may contain outliers
- but the data does not contain class labels
- how to build a classifier given one class?

POINT ANOMALY DETECTION: ONE CLASS SVM

- only uses data objects from the normal class
- learns the boundary of normal data objects
- uses kernel functions
- leverages on the origin trick



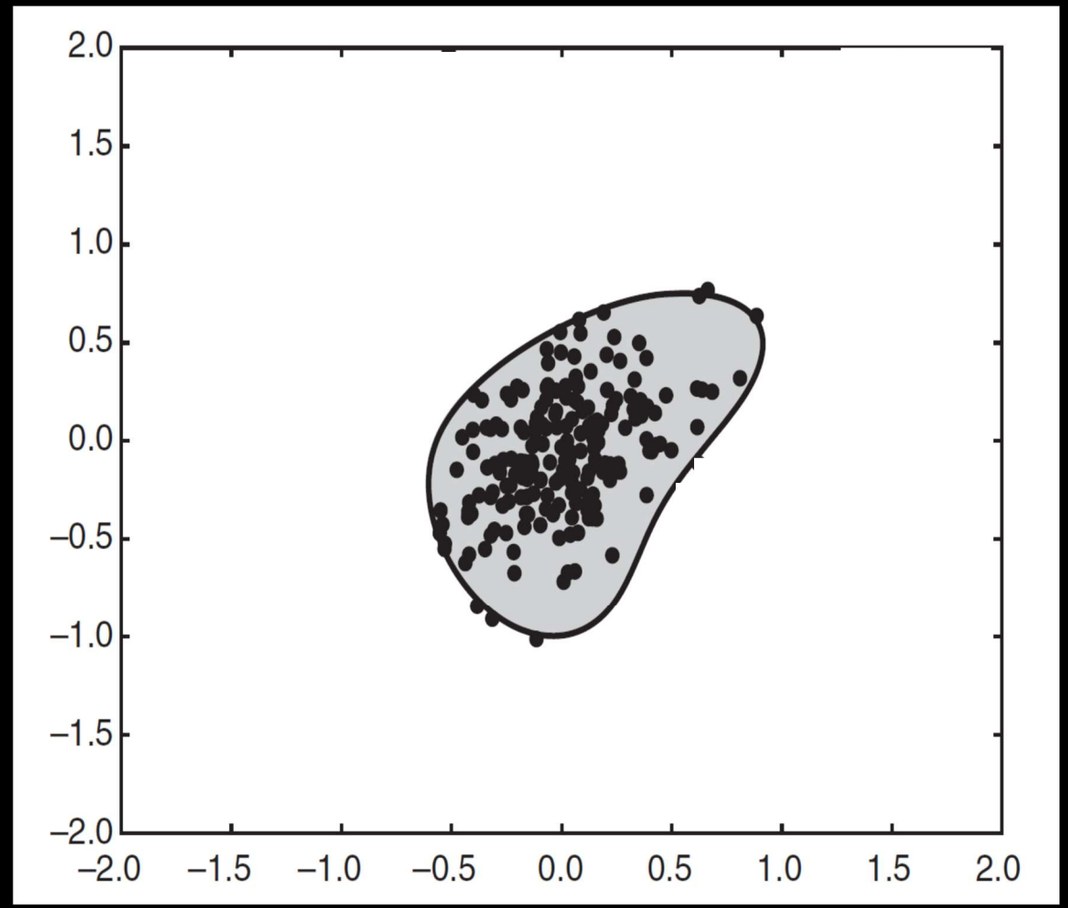
POINT ANOMALY DETECTION: ONE CLASS SVM

KERNELS

- to learn nonlinear boundary that encloses the normal class, we transform the data to a higher dimensional space where the normal class can be separated using a linear hyperplane
- this is done by using a function ϕ that maps every data object $\mathbf{x}$ from the original space of attributes to a point $\phi(\mathbf{x})$ in the transformed high dimensional space
- in the transformed space, the training instances can be separated using a linear hyperplane defined by parameters $(\mathbf{w}, b)$ as follows

$$\langle w, \phi(\mathbf{x}) \rangle = b$$

- thus we want a linear hyperplane that places all of the normal instances on one side, and thus we want $(\mathbf{w}, b)$ be such that…



$\langle w, \phi(\mathbf{x}) \rangle > b$ if $\mathbf{x}$ belongs to the normal class

$\langle w, \phi(\mathbf{x}) \rangle < b$ if $\mathbf{x}$ belongs to the anomaly class

POINT ANOMALY DETECTION: ONE CLASS SVM

ORIGIN TRICK

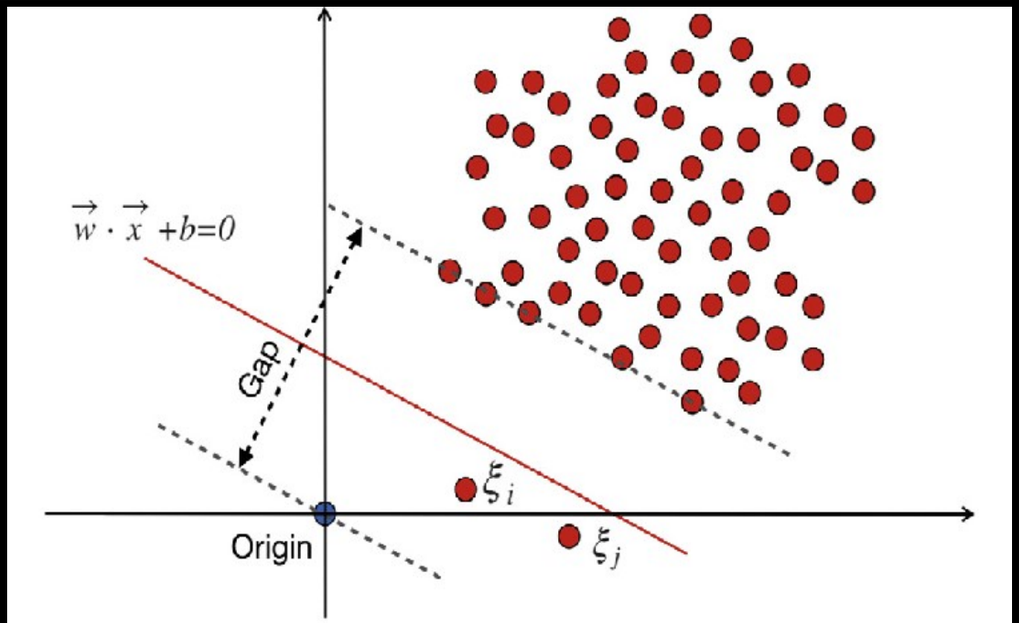
- use a Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)$
  - every point $\mathbf{x}$ is mapped to a unit hypersphere

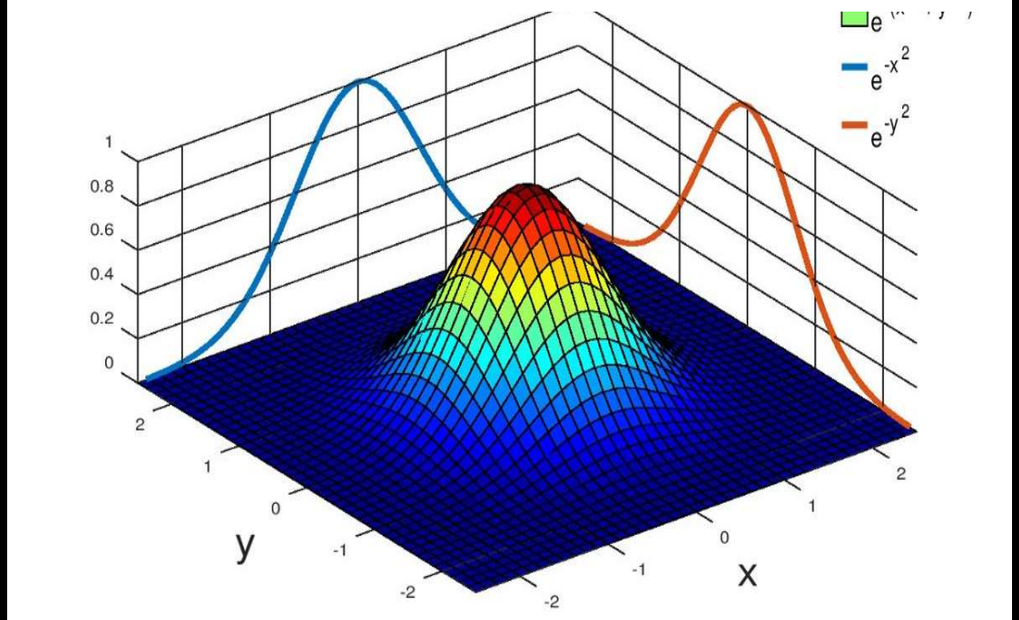
$$k(\mathbf{x}, \mathbf{x}) = \langle \phi(\mathbf{x}), \phi(\mathbf{x}) \rangle = \|\mathbf{x}\|^2 = 1$$

  - all points $\mathbf{x}$ and $\boldsymbol{y}$ in the same orthant (quadrant)

$$k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle \geq 0$$

- it aims to maximize the distance of the separating plane from the origin



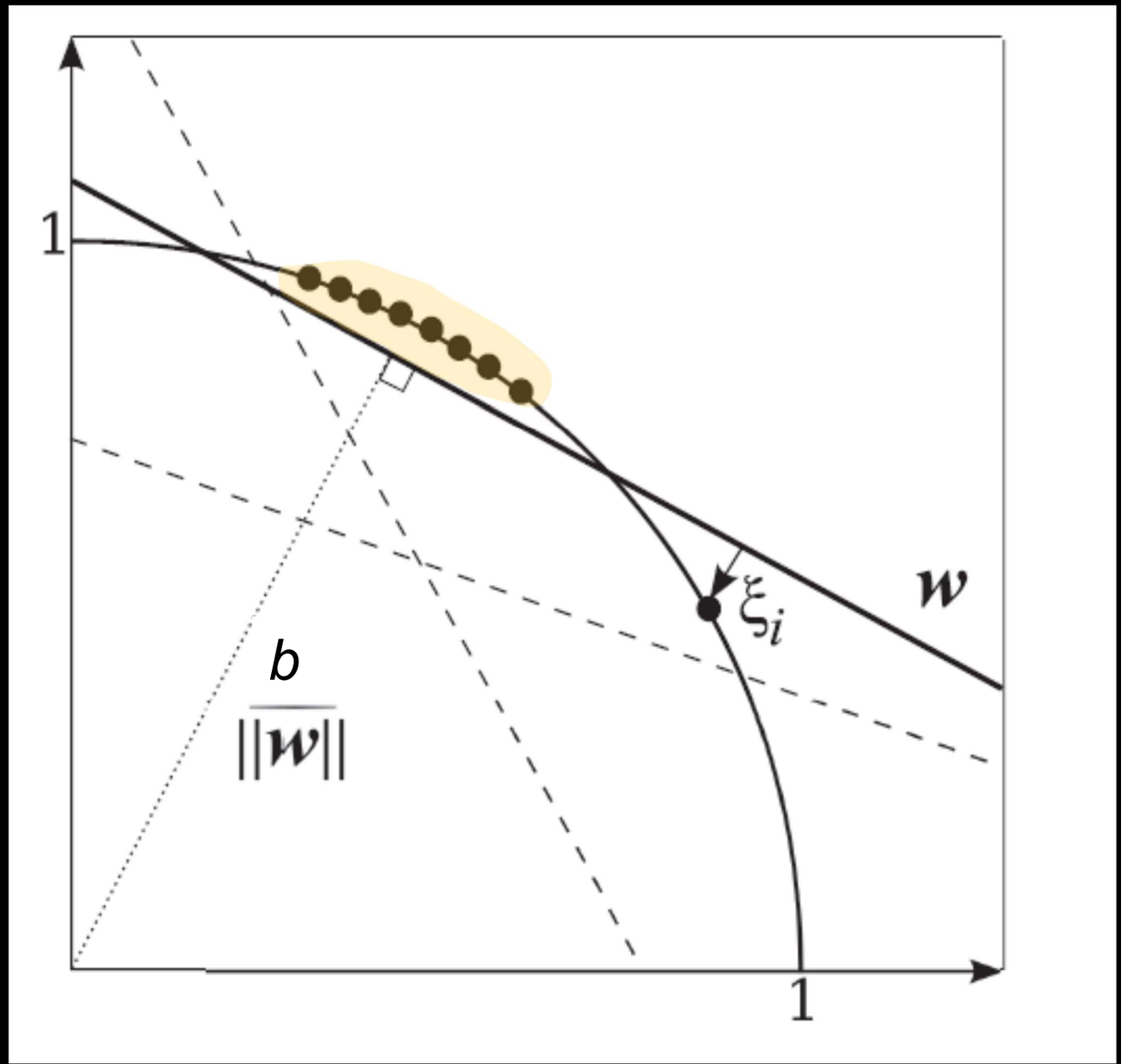


POINT ANOMALY DETECTION: TWO DIMENSIONAL ONE CLASS SVM

- the hyperplane should be as distant from the origin as possible
- this ensures a tight representation of points on the on the upper side of the hyperplane (corresponding to the normal class).
- the distance of the hyperplane from the origin is

$$\frac{b}{||\boldsymbol{w}||}$$

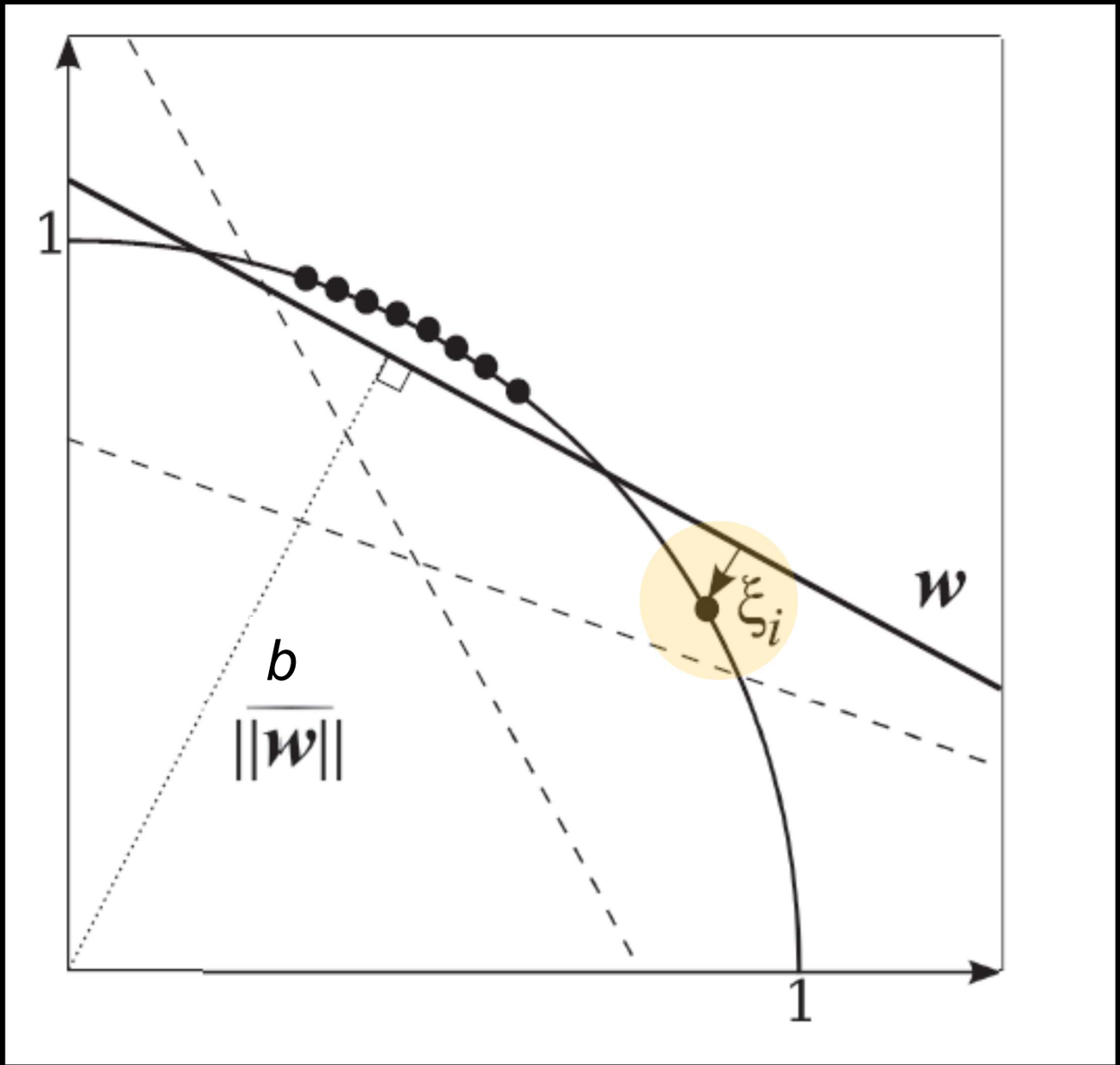
- maximizing b translates to maximizing the distance of the hyperplane from the origin.



POINT ANOMALY DETECTION: TWO DIMENSIONAL ONE CLASS SVM

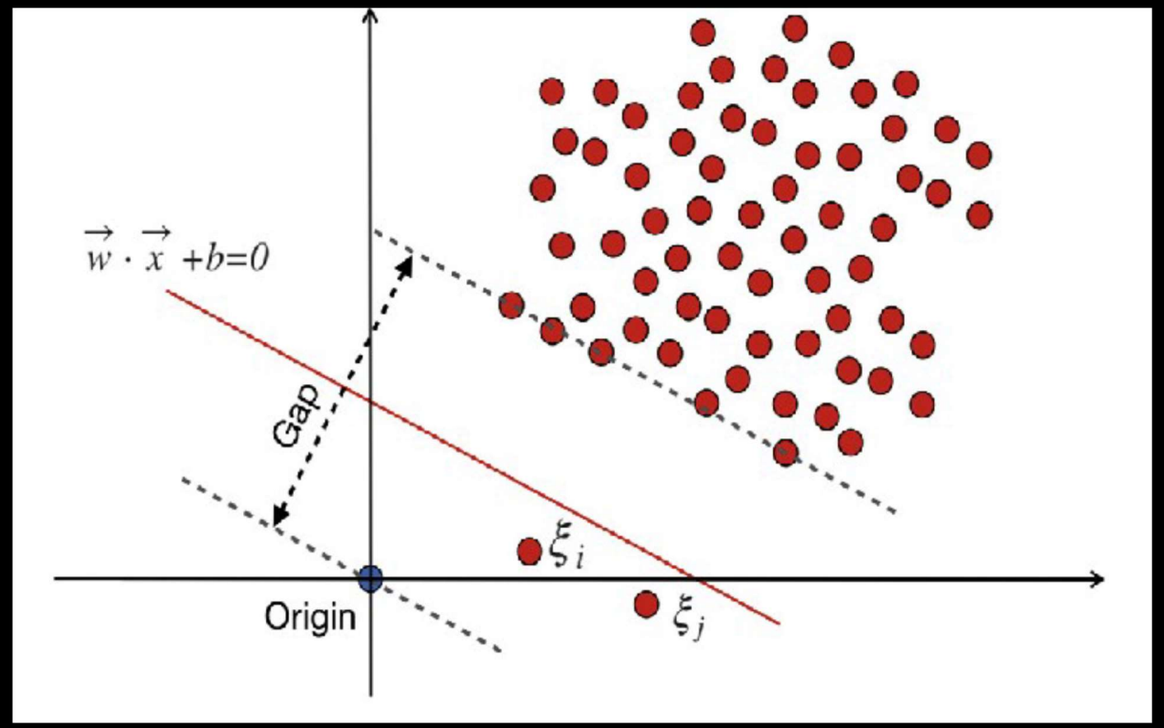
- In the style of soft margin SVMs, if some of the training instances lie on the opposite side of the hyperplane (corresponding to the anomaly class), then the distance of such points from the hyperplane should be minimized

- It is important for an anomaly detection algorithm to be robust to a small number of outliers in the training set



ONE CLASS SVM: EQUATIONS

- equation of hyperplane $\langle \mathbf{w}, \phi(\mathbf{x}) \rangle = b$
- ϕ is the mapping to high dimensional space
- weight vector is $\mathbf{w} = \sum_{i=1}^{n} \alpha_i \, \phi(\mathbf{x}_i)$
- υ is fraction of outliers
- solving the following mathematical programming problem

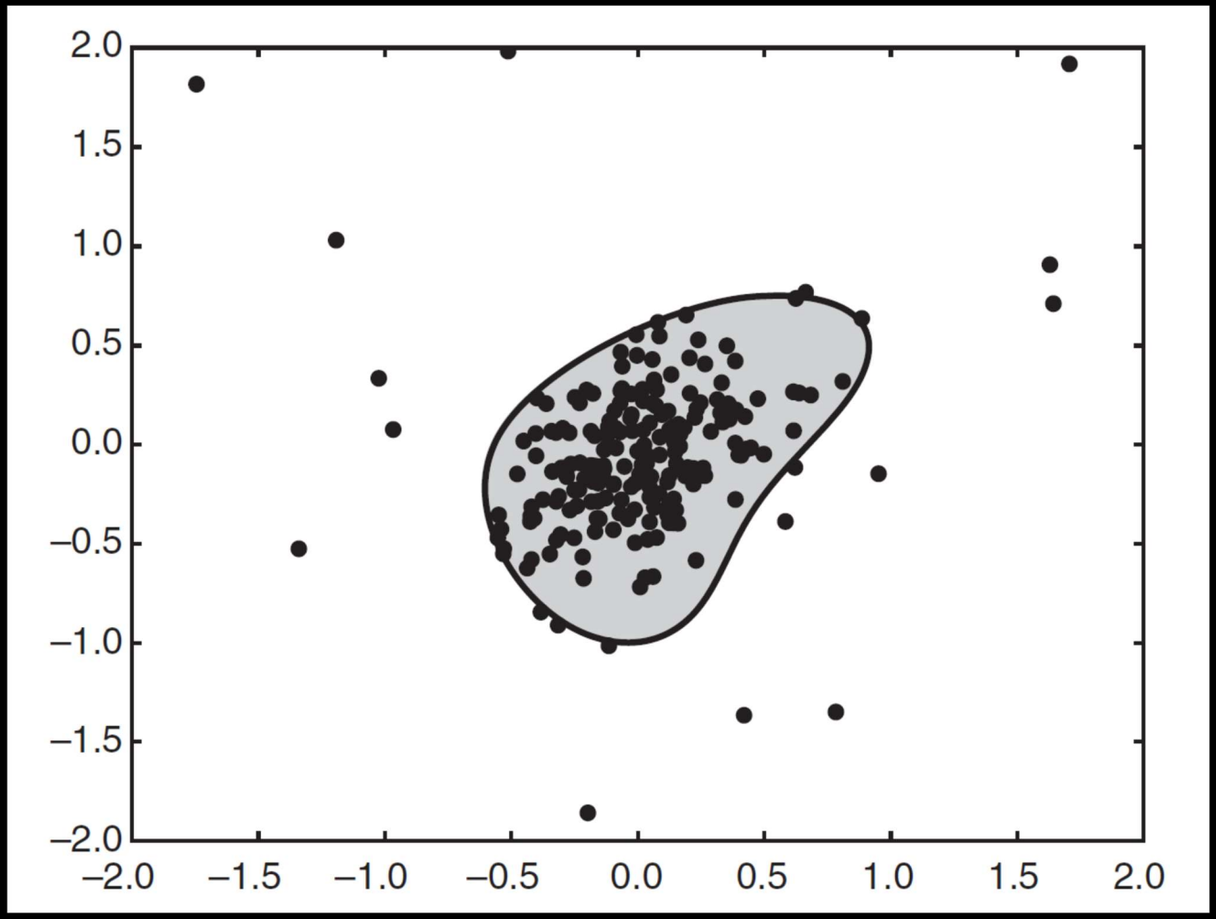


$$\min_{\mathbf{w},b,\xi} \frac{1}{2} \|\mathbf{w}\|^2 - b + \frac{1}{n\upsilon} \sum_{i=1}^{n} \xi_i$$

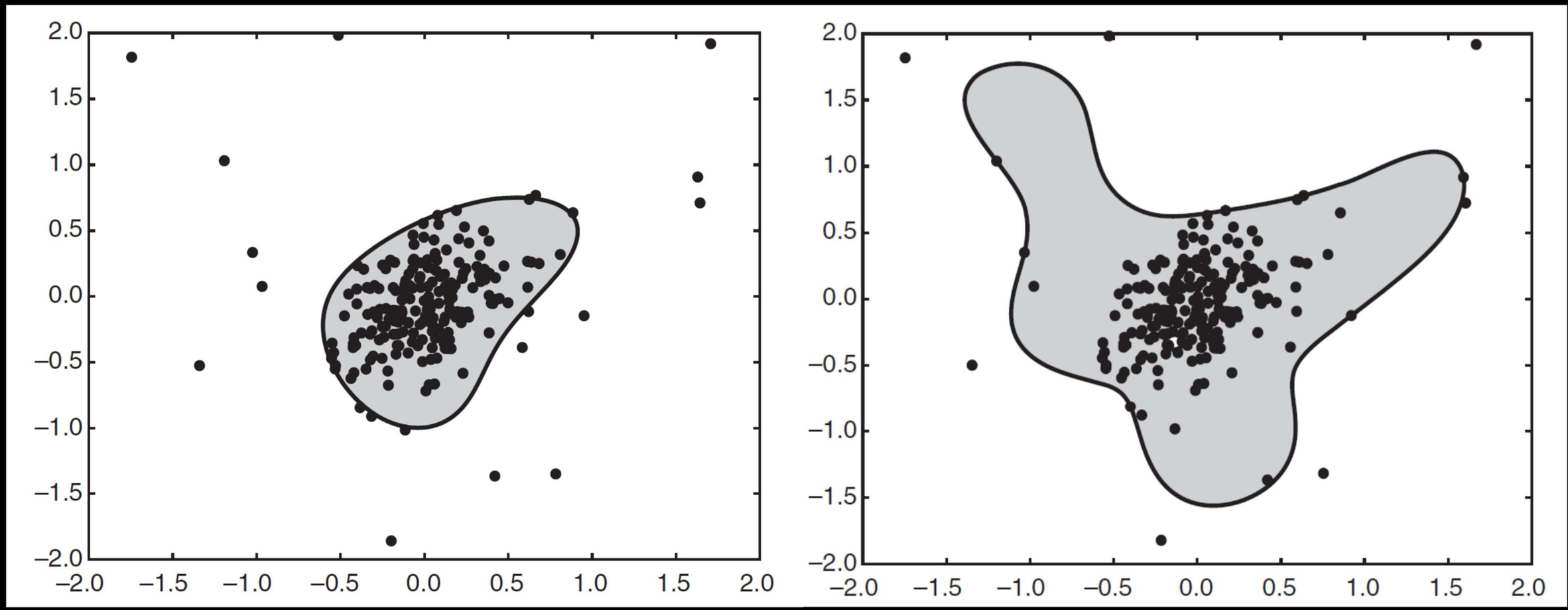
$$\langle \mathbf{w}, \phi(\mathbf{x}) \rangle \geq b - \xi_i \qquad i = 1, \ldots, n$$

$$\xi_i \geq 0$$

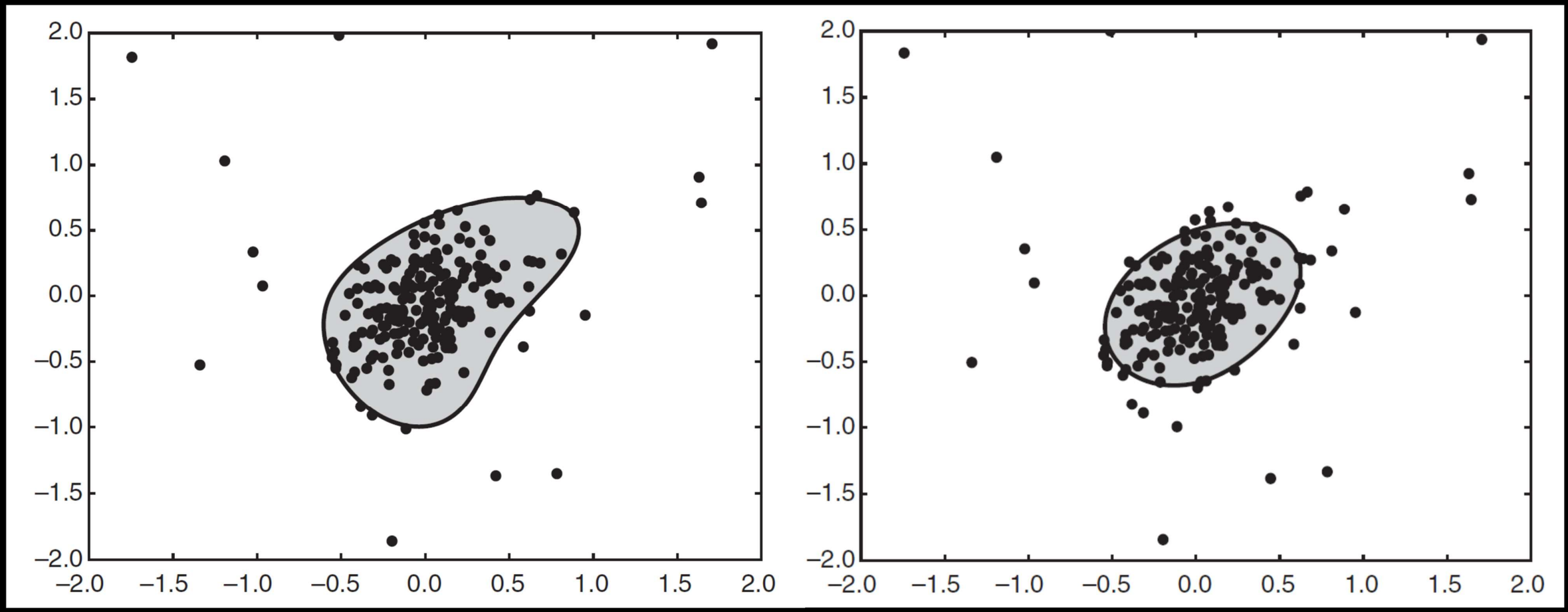
ONE CLASS SVM: DECISION BOUNDARY WITH $\upsilon = 0.1$



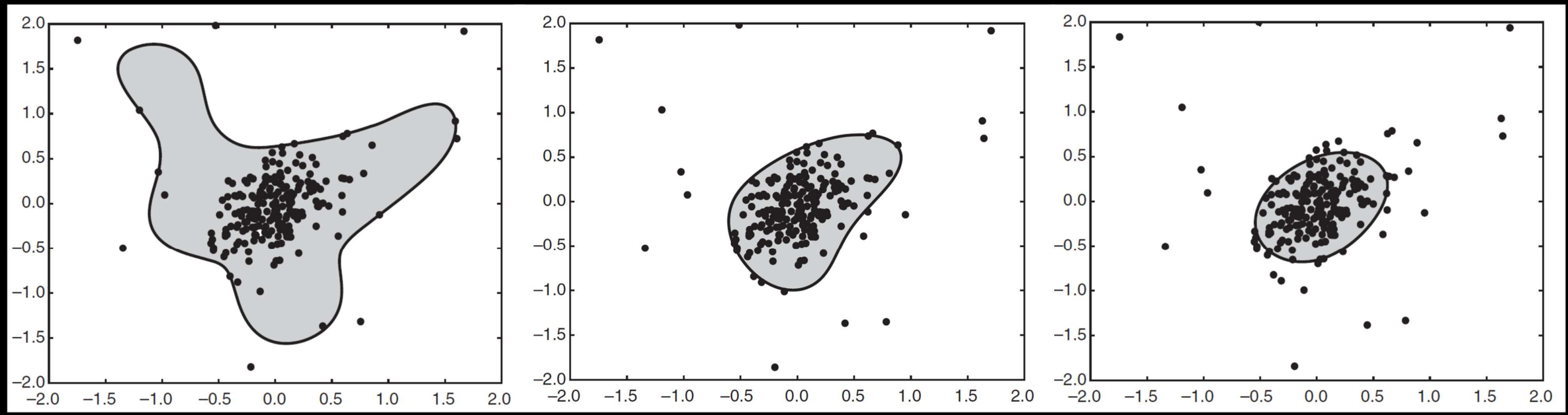
ONE CLASS SVM: DECISION BOUNDARY $\upsilon = 0.1$ VS $\upsilon = 0.05$



ONE CLASS SVM: DECISION BOUNDARY $\upsilon = 0.1$ VS $\upsilon = 0.2$



ONE CLASS SVM: DECISION BOUNDARY $v = 0.05 \; - \; v = 0.1 - v = 0.2$



ONE CLASS SVM: STRENGHTS AND WEAKNESSES

- Strong theoretical foundation
- Choice of υ is difficult
- Computationally expensive

INFORMATION THEORETIC APPROACHES

- key idea is to measure how much information decreases when you delete an observation

$$Gain(\mathbf{x}) = Info(\mathbf{D}) - Info(\mathbf{D}\backslash\mathbf{x})$$

- anomalies $\mathbf{x}$ should show higher gain
- normal points $\mathbf{x}$ should have less gain

$$Info(\mathbf{D}) = Entropy(\mathbf{D}) = \sum_{i=1}^{n} -p_i \, log_2(p_i)$$

weight	height	Frequency
low	low	20
low	medium	15
medium	medium	40
high	high	20
high	low	5

p_i	$-p_i log_2(p_i)$
0,20	0,46
0,15	0,41
0,40	0,53
0,20	0,46
0,05	0,22

$2,08$

INFORMATION THEORETIC APPROACHES

- key idea is to measure how much information decreases when you delete an observation

$$Gain(\mathbf{x}) = Info(\mathbf{D}) - Info(\mathbf{D}\backslash\mathbf{x})$$

- anomalies $\mathbf{x}$ should show higher gain
- normal points $\mathbf{x}$ should have less gain

$$Info(\mathbf{D}) = Entropy(\mathbf{D}) = \sum_{i=1}^{n} -p_i \, log_2(p_i)$$

weight	height	Frequency
low	low	20
low	medium	15
medium	medium	40
high	high	20
~~high~~	~~low~~	~~5~~

p_i	$-p_i log_2(p_i)$
0,21	0,47
0,16	0,42
0,42	0,53
0,21	0,47
	1,89

INFORMATION THEORETIC APPROACHES

- key idea is to measure how much information decreases when you delete an observation

$$Gain(\mathbf{x}) = Info(\mathbf{D}) - Info(\mathbf{D}\backslash\mathbf{x})$$

- anomalies $\mathbf{x}$ should show higher gain
- normal points $\mathbf{x}$ should have less gain

$$Info(\mathbf{D}) = Entropy(\mathbf{D}) = \sum_{i=1}^{n} -p_i \, log_2(p_i)$$

weight	height	Frequency
low	low	20
low	medium	15
medium	medium	40
high	high	20
high	low	5

p_i	$-p_i log_2(p_i)$
0,20	0,46
0,15	0,41
0,40	0,53
0,20	0,46
0,05	0,22
	$2,08$

p_i	$-p_i log_2(p_i)$
0,21	0,47
0,16	0,42
0,42	0,53
0,21	0,47
	$1,89$

$$Gain(\mathbf{x}) = 2{,}08 - 1{,}89 = 0.19$$

INFORMATION THEORETIC APPROACHES: STRENGHTS AND WEAKNESSES

- Solid theoretical foundation
- Theoretically applicable to all kinds of data
- Difficult and computationally expensive to implement in practice

EVALUATION OF ANOMALY DETECTION

- If class labels are present, then use standard evaluation approaches for rare class such as precision, recall, or false positive rate
  - FPR is also know as false alarm rate
- For unsupervised anomaly detection use measures provided by the anomaly method
  - e.g. reconstruction error or gain
- Can also look at histograms of anomaly scores

EVALUATION OF ANOMALY DETECTION

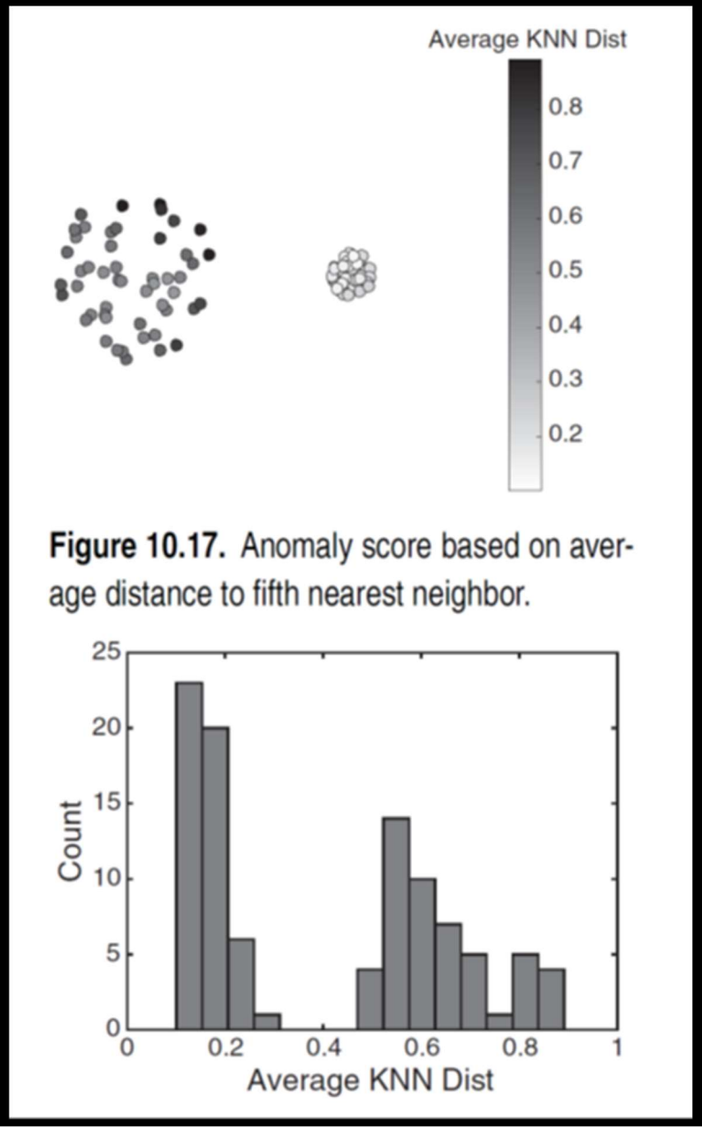


Figure 10.17. Anomaly score based on average distance to fifth nearest neighbor.

EVALUATION OF ANOMALY DETECTION

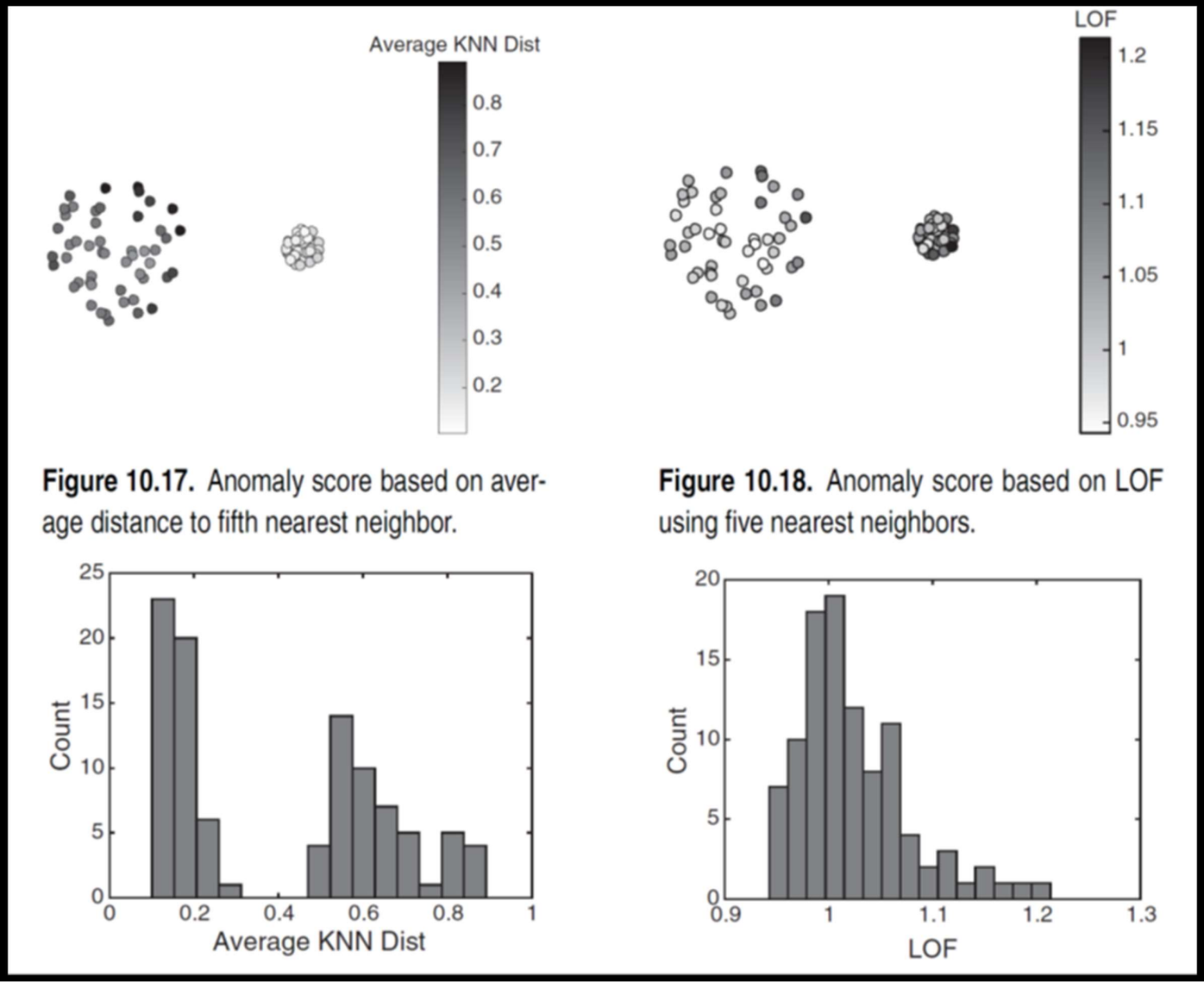


Figure 10.17. Anomaly score based on average distance to fifth nearest neighbor.

Figure 10.18. Anomaly score based on LOF using five nearest neighbors.

- The majority of anomaly scores should be relatively low, with a smaller fraction of scores toward the higher end.
  - we assume a higher score indicates an instance is more anomalous
- By looking at the distribution of the scores, via histogram or density plot, we can assess whether the approach we are using generates scores that behave in a reasonable manner.

EVALUATION OF ANOMALY DETECTION

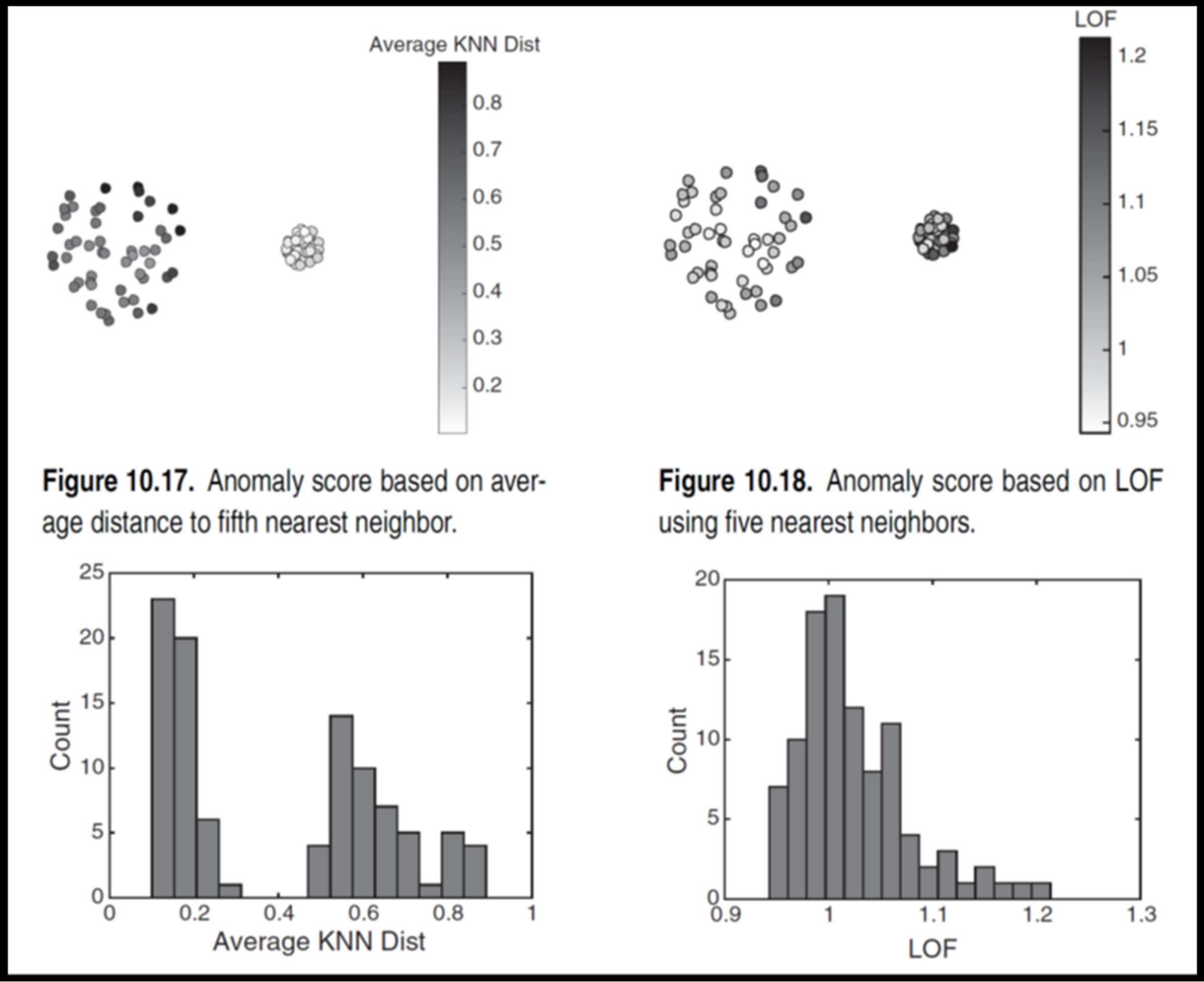


Figure 10.17. Anomaly score based on average distance to fifth nearest neighbor.

Figure 10.18. Anomaly score based on LOF using five nearest neighbors.

- The distribution of anomaly scores should look similar to that of the LOF scores in the example to the left.
- There may be one or more secondary peaks in the distribution as one moves to the right, but these secondary peaks should only contain a relatively small fraction of the points, and not a large fraction of the points as with the average KNN data approach.

RECAP

- One Class SVM
- Information Theoretic Approaches
- Evaluation of Anomaly Detection