# One-class logistic regression & friends

## Probabilistic anomaly detection as loss minimisation

Aditya Krishna Menon  Robert C. Williamson

The Australian National University

Jun 28th, 2018

# Anomaly detection

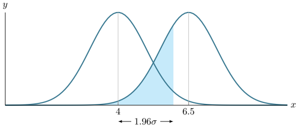
Identify instances that deviate from some systematic pattern

# Anomaly detection

Identify instances that deviate from some systematic pattern

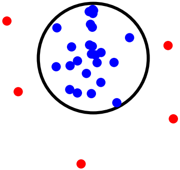
# Anomaly detection landscape

**Statistical test**

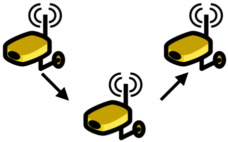



**Structural health monitoring**

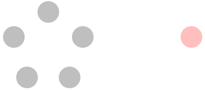
**One-class SVM**



**Network analysis**



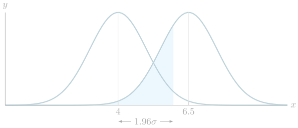
**Nearest neighbour**



**Credit fraud**

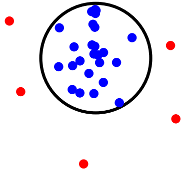
# Anomaly detection landscape

### Statistical test

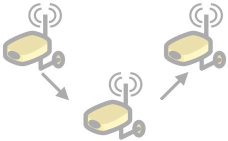



### Structural health monitoring

### One-class SVM



### Network analysis



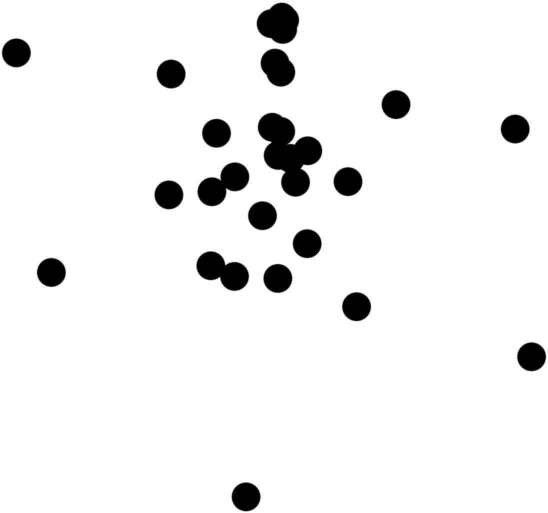
### Nearest neighbour



### Credit fraud

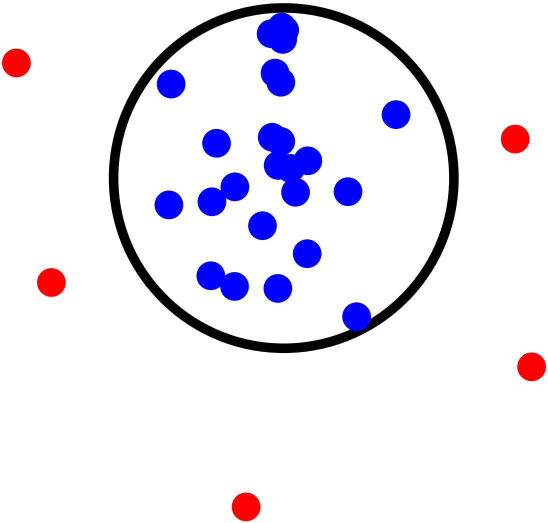
# One-class SVMs: enclosing ball view

Find the smallest ball enclosing most of the data



# One-class SVMs: enclosing ball view

Find the smallest ball enclosing most of the data



# One-class SVMs: origin separation view

Pick an RKHS $\mathcal{H}$ and desired anomaly fraction $\nu \in (0, 1)$

# One-class SVMs: origin separation view

Pick an RKHS $\mathcal{H}$ and desired anomaly fraction $\nu \in (0, 1)$

Find $f \in \mathcal{H}$ and offset $\alpha \in \mathbb{R}$ to separate data from origin

# One-class SVMs: origin separation view

Pick an RKHS $\mathcal{H}$ and desired anomaly fraction $\nu \in (0, 1)$

Find $f \in \mathcal{H}$ and offset $\alpha \in \mathbb{R}$ to separate data from origin

For data distribution $P$,

$$\min_{f,\alpha} \mathop{\mathbb{E}}_{P} [\alpha - f(\mathsf{X})]_+ + \frac{\nu}{2} \cdot \|f\|_{\mathcal{H}}^2 - \nu \cdot \alpha$$

# One-class SVMs: origin separation view

Pick an RKHS $\mathcal{H}$ and desired anomaly fraction $\nu \in (0, 1)$

Find $f \in \mathcal{H}$ and offset $\alpha \in \mathbb{R}$ to separate data from origin

For data distribution $P$,

$$\min_{f,\alpha} \underbrace{\mathbb{E}_{P}[\alpha - f(\mathsf{X})]_+}_{\text{hinge loss}} + \underbrace{\frac{\nu}{2} \cdot \|f\|_{\mathcal{H}}^2}_{\text{regulariser}} - \underbrace{\nu \cdot \alpha}_{\nu-\text{SVM relic}}$$

# One-class SVMs: pros and cons

OC-SVMs inherit the standard SVM's strengths and weaknesses

- ✓ convex objective
- ✓ focus effort on decision boundary

# One-class SVMs: pros and cons

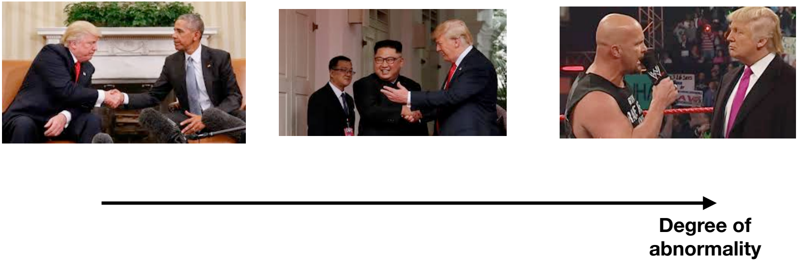
OC-SVMs inherit the standard SVM's strengths and weaknesses

- ✓ convex objective
- ✓ focus effort on decision boundary
- × doesn't focus on probability of instance being anomalous

# One-class SVMs: pros and cons

OC-SVMs inherit the standard SVM's strengths and weaknesses

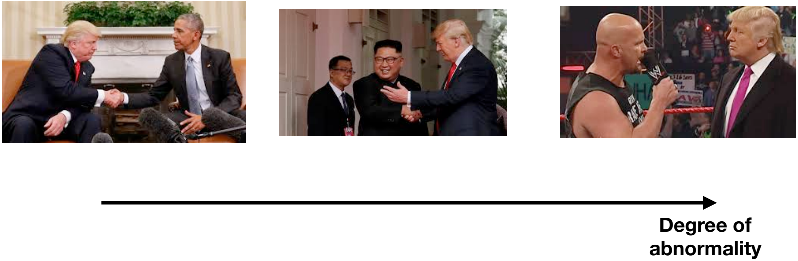
- ✓ convex objective
- ✓ focus effort on decision boundary
- × doesn't focus on probability of instance being anomalous



# One-class SVMs: pros and cons

OC-SVMs inherit the standard SVM's strengths and weaknesses

- ✓ convex objective
- ✓ focus effort on decision boundary
- × doesn't focus on probability of instance being anomalous
- × unclear Bayes-optimal solution



# This talk

**Take-home #1**

Anomaly detection = binary classification

- distinguish samples against an implicit background

# This talk

**Take-home #1**

Anomaly detection = binary classification

- distinguish samples against an implicit background

**Take-home #2**

Probabilistic anomaly detection = class-probability estimation

- can use familiar tools: logistic regression, boosting, ...

# This talk

**Take-home #1**

Anomaly detection = binary classification

- distinguish samples against an implicit background

**Take-home #2**

Probabilistic anomaly detection = class-probability estimation

- can use familiar tools: logistic regression, boosting, ...

**Surprise**

Specific kind of OC-SVM turns out to be a special case!

- gives a different perspective on underlying components

# Deconstructing one-class SVMs

Pick an RKHS $\mathcal{H}$ and desired anomaly fraction $\nu \in (0,1)$

For data distribution $P$, the OC-SVM solves

$$\min_{f,\alpha} \underbrace{\mathbb{E}_{P}[\alpha - f(\mathsf{X})]_+}_{\text{hinge loss}} + \underbrace{\frac{\nu}{2} \cdot \|f\|_{\mathcal{H}}^2}_{\text{regulariser}} - \underbrace{\nu \cdot \alpha}_{\nu-\text{SVM relic}}$$

# Deconstructing one-class SVMs

Pick an RKHS $\mathcal{H}$ and desired anomaly fraction $\nu \in (0, 1)$

For data distribution $P$, the OC-SVM solves

$$\min_{f,\alpha} \underbrace{\mathbb{E}_{P}[\alpha - f(\mathsf{X})]_+}_{\substack{\text{~~hinge loss~~} \\ \textbf{capped proper loss}}} + \underbrace{\frac{\nu}{2} \cdot \|f\|_{\mathcal{H}}^2}_{\substack{\text{regulariser} \\ \textbf{+ background contrast}}} - \underbrace{\nu \cdot \alpha}_{\substack{\nu\text{~~-SVM relic~~} \\ \textbf{pinball loss}}}$$

We give a different interpretation for the OC-SVM's components

# Anomaly detection as classification

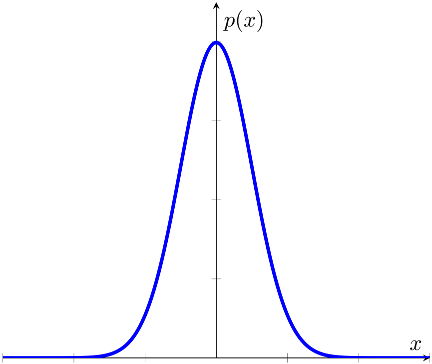
# Density sublevel view of anomaly detection

Pick a reference measure $\mu$ (e.g., Lebesgue)

# Density sublevel view of anomaly detection

Pick a reference measure $\mu$ (e.g., Lebesgue)

Suppose our data distribution $P$ has density $p \doteq \frac{\mathrm{d}P}{\mathrm{d}\mu}$

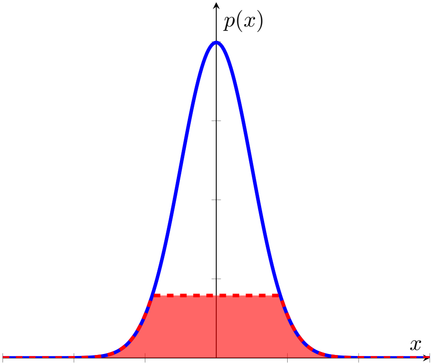


# Density sublevel view of anomaly detection

Pick a reference measure $\mu$ (e.g., Lebesgue)

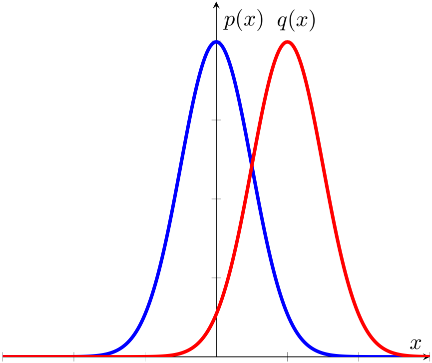
Suppose our data distribution $P$ has density $p \doteq \frac{\mathrm{d}P}{\mathrm{d}\mu}$

Define anomalies to be instances with low density



# Recap: binary classification

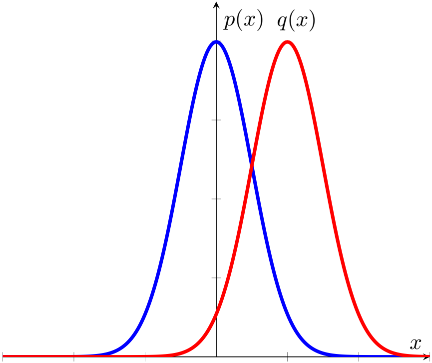
Suppose we have positive and negative data distributions $P, Q$



# Recap: binary classification

Suppose we have positive and negative data distributions $P, Q$

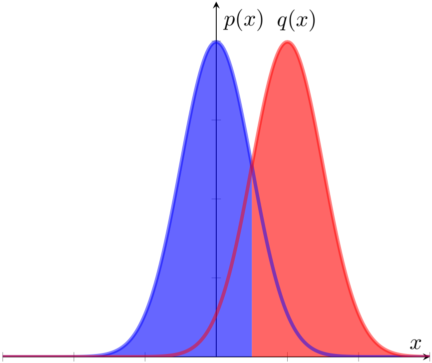
Classify instances based on dominant density



# Recap: binary classification

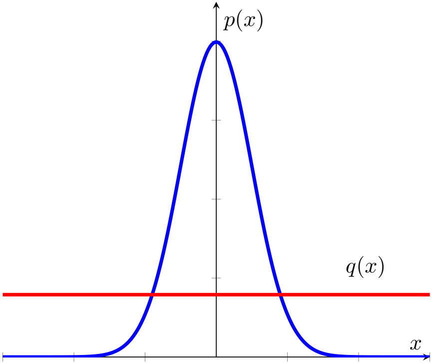
Suppose we have positive and negative data distributions $P, Q$

Classify instances based on dominant density



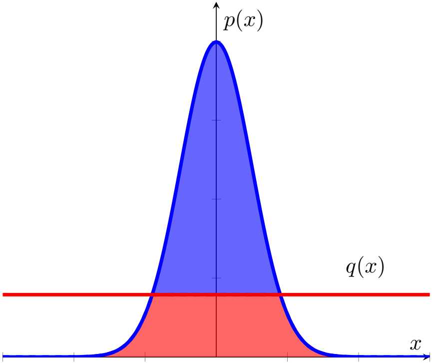
# Anomaly detection as binary classification

Consider classification of data distribution $P$ versus uniform $Q$



# Anomaly detection as binary classification

Consider classification of data distribution $P$ versus uniform $Q$




**Anomaly detection $=$ classification against uniform background!**
**(Steinwart & Scovel, 2005)**

# Anomaly detection as binary classification

Fix some density threshold $\alpha > 0$

Anomaly detection seeks a scorer $f : \mathcal{X} \to \mathbb{R}$, where[1]

$$f(x) > \alpha \iff p(x) > \alpha$$

[1] We assume $P(p(\mathsf{X}) = \alpha) = 0$

# Anomaly detection as binary classification

Fix some density threshold $\alpha > 0$

Anomaly detection seeks a scorer $f \colon \mathcal{X} \to \mathbb{R}$, where[1]

$$f(x) > \alpha \iff p(x) > \alpha$$

(Steinwart & Scovel, 2005): classify data $P$ against background $Q$:

$$\min_f \mathop{\mathbb{E}}_{P} [\![ f(\mathsf{X}) < \alpha ]\!] + \alpha \cdot \mathop{\mathbb{E}}_{Q} [\![ f(\mathsf{X}) > \alpha ]\!]$$

- cost-weighted classification loss

[1] We assume $P(p(\mathsf{X}) = \alpha) = 0$

# Anomaly detection as binary classification

Fix some density threshold $\alpha > 0$

Anomaly detection seeks a scorer $f\colon \mathcal{X} \to \mathbb{R}$, where[1]

$$f(x) > \alpha \iff p(x) > \alpha$$

(Steinwart & Scovel, 2005): classify data $P$ against background $Q$:

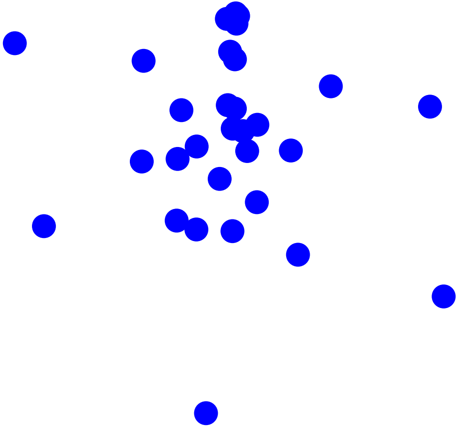
$$\min_f \mathop{\mathbb{E}}_{P} [\![ f(\mathsf{X}) < \alpha ]\!] + \alpha \cdot \mathop{\mathbb{E}}_{Q} [\![ f(\mathsf{X}) > \alpha ]\!]$$

- cost-weighted classification loss

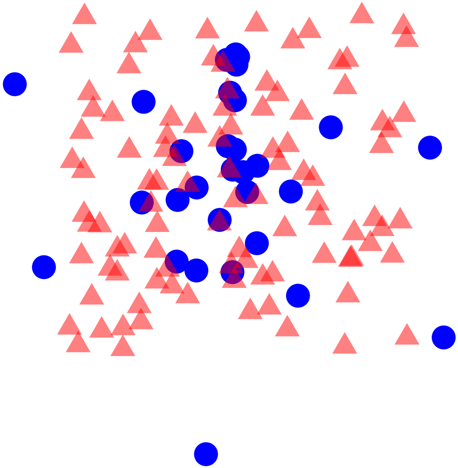
**Anomaly detection as binary classification!**

[1] We assume $P(p(\mathsf{X}) = \alpha) = 0$

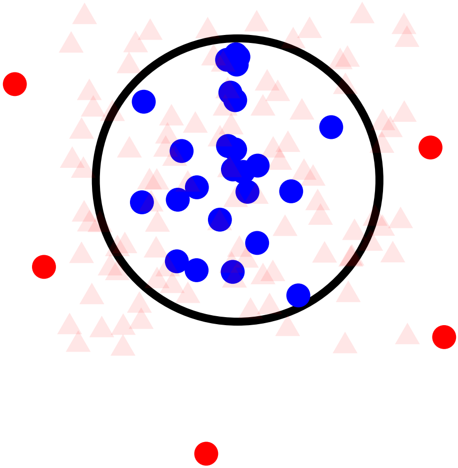
# Anomaly detection as binary classification



# Anomaly detection as binary classification



# Anomaly detection as binary classification



# This talk

**Take-home #1** ✓

Anomaly detection = binary classification

- distinguish samples against an implicit background

# This talk

**Take-home #1** ✓

Anomaly detection = binary classification

- distinguish samples against an implicit background

**Take-home #2** ?

Probabilistic anomaly detection = class-probability estimation

- can use familiar tools: logistic regression, boosting, ...

# Changing the loss function

What if we instead minimise

$$\min_f \mathop{\mathbb{E}}_{P} \ell(+1, f(\mathsf{X})) + \mathop{\mathbb{E}}_{Q} \ell(-1, f(\mathsf{X}))$$

for a generic loss $\ell \colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}$?

# Changing the loss function

What if we instead minimise

$$\min_f \mathop{\mathbb{E}}_{P} \ell(+1, f(\mathsf{X})) + \mathop{\mathbb{E}}_{Q} \ell(-1, f(\mathsf{X}))$$

for a generic loss $\ell\colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}$?

Result will be exactly per discrimination of $P$ versus $Q$

# Changing the loss function

What if we instead minimise

$$\min_f \mathbb{E}_{P} \ell(+1, f(\mathsf{X})) + \mathbb{E}_{Q} \ell(-1, f(\mathsf{X}))$$

for a generic loss $\ell \colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}$?

Result will be exactly per discrimination of $P$ versus $Q$

e.g., for proper losses, we recover $p(x)$

- i.e., we perform density estimation

# A running example

Consider the LSIF loss (Kanamori et al., 2009)

$$\ell(+1, f) = -f \qquad \ell(-1, f) = \frac{1}{2} \cdot f^2$$

# A running example

Consider the LSIF loss (Kanamori et al., 2009)

$$\ell(+1, f) = -f \qquad \ell(-1, f) = \frac{1}{2} \cdot f^2$$

The objective becomes:

# A running example

Consider the LSIF loss (Kanamori et al., 2009)

$$\ell(+1, f) = -f \qquad \ell(-1, f) = \frac{1}{2} \cdot f^2$$

The objective becomes:

$$\text{Risk}(f) = \mathbb{E}_{P} \ell(+1, f(\mathsf{X})) + \mathbb{E}_{Q} \ell(-1, f(\mathsf{X}))$$

# A running example

Consider the LSIF loss (Kanamori et al., 2009)

$$\ell(+1,f) = -f \qquad \ell(-1,f) = \frac{1}{2}\cdot f^2$$

The objective becomes:

$$\begin{aligned}\text{Risk}(f) &= \mathop{\mathbb{E}}_{P} \ell(+1, f(\mathsf{X})) + \mathop{\mathbb{E}}_{Q} \ell(-1, f(\mathsf{X})) \\ &= \mathop{\mathbb{E}}_{P} -f(\mathsf{X}) + \mathop{\mathbb{E}}_{Q} \frac{1}{2}\cdot f(\mathsf{X})^2\end{aligned}$$

# A running example

Consider the LSIF loss (Kanamori et al., 2009)

$$\ell(+1, f) = -f \qquad \ell(-1, f) = \frac{1}{2} \cdot f^2$$

The objective becomes:

$$\begin{aligned}
\text{Risk}(f) &= \mathop{\mathbb{E}}_{P} \ell(+1, f(\mathsf{X})) + \mathop{\mathbb{E}}_{Q} \ell(-1, f(\mathsf{X})) \\
&= \mathop{\mathbb{E}}_{P} -f(\mathsf{X}) + \mathop{\mathbb{E}}_{Q} \frac{1}{2} \cdot f(\mathsf{X})^2 \\
&= \mathop{\mathbb{E}}_{Q} \left[ -p(\mathsf{X}) \cdot f(\mathsf{X}) + \frac{1}{2} \cdot f(\mathsf{X})^2 \right]
\end{aligned}$$

# A running example

Consider the LSIF loss (Kanamori et al., 2009)

$$\ell(+1, f) = -f \qquad \ell(-1, f) = \frac{1}{2} \cdot f^2$$

The objective becomes:

$$\begin{aligned} \mathrm{Risk}(f) &= \mathop{\mathbb{E}}_{P} \ell(+1, f(\mathsf{X})) + \mathop{\mathbb{E}}_{Q} \ell(-1, f(\mathsf{X})) \\ &= \mathop{\mathbb{E}}_{P} -f(\mathsf{X}) + \mathop{\mathbb{E}}_{Q} \frac{1}{2} \cdot f(\mathsf{X})^2 \\ &= \mathop{\mathbb{E}}_{Q} \left[ -p(\mathsf{X}) \cdot f(\mathsf{X}) + \frac{1}{2} \cdot f(\mathsf{X})^2 \right] \\ &= \mathop{\mathbb{E}}_{Q} (f(\mathsf{X}) - p(\mathsf{X}))^2 + \text{constant}. \end{aligned}$$

# A running example

Consider the LSIF loss (Kanamori et al., 2009)

$$\ell(+1, f) = -f \qquad \ell(-1, f) = \frac{1}{2} \cdot f^2$$

The objective becomes:

$$\begin{aligned}
\mathrm{Risk}(f) &= \mathop{\mathbb{E}}_{P} \ell(+1, f(\mathsf{X})) + \mathop{\mathbb{E}}_{Q} \ell(-1, f(\mathsf{X})) \\
&= \mathop{\mathbb{E}}_{P} -f(\mathsf{X}) + \mathop{\mathbb{E}}_{Q} \frac{1}{2} \cdot f(\mathsf{X})^2 \\
&= \mathop{\mathbb{E}}_{Q} \left[ -p(\mathsf{X}) \cdot f(\mathsf{X}) + \frac{1}{2} \cdot f(\mathsf{X})^2 \right] \\
&= \mathop{\mathbb{E}}_{Q} (f(\mathsf{X}) - p(\mathsf{X}))^2 + \text{constant}.
\end{aligned}$$

**LSIF loss minimisation = least squares density fitting!**

# State of play

The general objective

$$\min_f \mathbb{E}_{P} \ell(+1, f(\mathsf{X})) + \mathbb{E}_{Q} \ell(-1, f(\mathsf{X}))$$

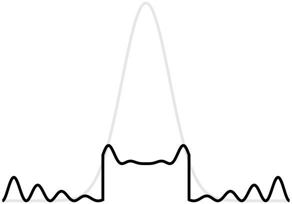
captures two distinct problem settings

# State of play

The general objective

$$\min_f \mathbb{E}_{P} \ell(+1, f(\mathsf{X})) + \mathbb{E}_{Q} \ell(-1, f(\mathsf{X}))$$

captures two distinct problem settings



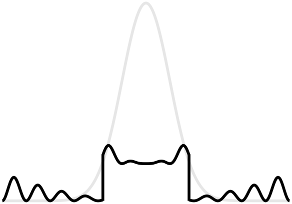
Density sublevel estimation
$\ell =$ cost-sensitive loss

# State of play

The general objective

$$\min_f \mathop{\mathbb{E}}_{P} \ell(+1, f(\mathsf{X})) + \mathop{\mathbb{E}}_{Q} \ell(-1, f(\mathsf{X}))$$

captures two distinct problem settings



Density sublevel estimation

$\ell =$ cost-sensitive loss

Density estimation

$\ell =$ proper loss

# State of play

The general objective

$$\min_f \mathop{\mathbb{E}}_{P} \ell(+1, f(\mathsf{X})) + \mathop{\mathbb{E}}_{Q} \ell(-1, f(\mathsf{X}))$$

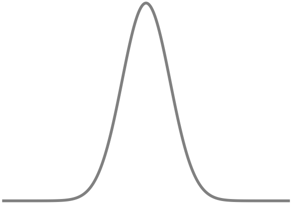
captures two distinct problem settings



Density sublevel estimation
$\ell =$ cost-sensitive loss

?

**Partial density estimation**
$\ell =$ ?



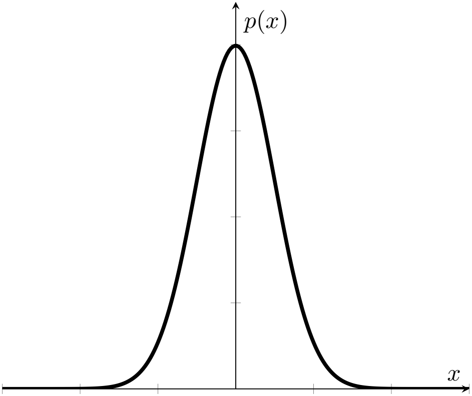
Density estimation
$\ell =$ proper loss

**What problem lives in between?**

# Partially proper losses

# Partial density estimation

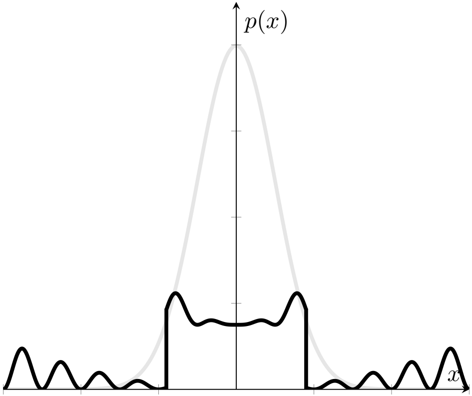
The targets for the two problem settings we've seen are:




The full $p(x)$ for density estimation

# Partial density estimation

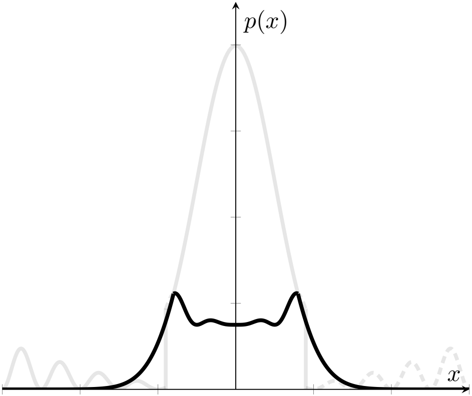
The targets for the two problem settings we've seen are:




The full $p(x)$ for density estimation and a thresholded version for sublevel estimation

# Partial density estimation

The targets for the two problem settings we've seen are:

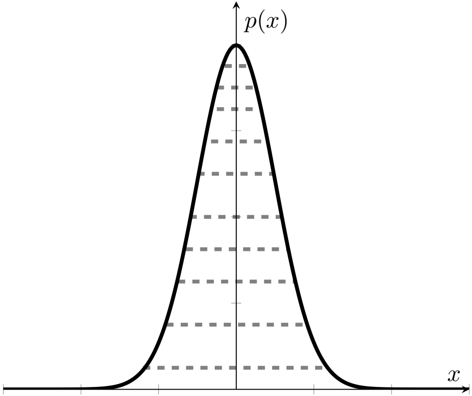



The full $p(x)$ for density estimation and a thresholded version for sublevel estimation

Natural intermediary: model the tail only

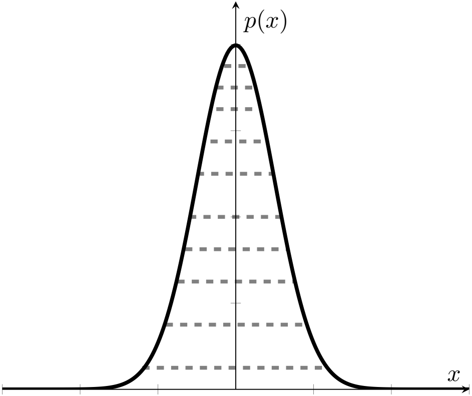
# An ensemble of cost-sensitive losses

Density estimation seeks the entire family of sublevel sets



# An ensemble of cost-sensitive losses

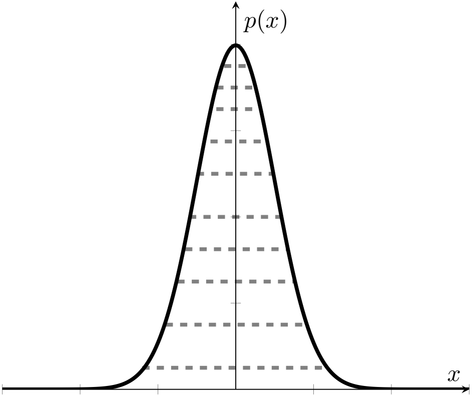
Density estimation seeks the entire family of sublevel sets




Each set is attainable with the $\alpha$ cost-sensitive loss

# An ensemble of cost-sensitive losses

Density estimation seeks the entire family of sublevel sets




Each set is attainable with the $\alpha$ cost-sensitive loss

**Combine losses for various values of $\alpha$?**

# Weight functions for proper losses

Consider the cost-sensitive loss

$$\ell_{\text{CS}}(+1, f; c) = (1 - c) \cdot [\![f < c]\!] \qquad \ell_{\text{CS}}(-1, f; c) = c \cdot [\![f > c]\!]$$

# Weight functions for proper losses

Consider the cost-sensitive loss

$$\ell_{\mathrm{CS}}(+1, f; c) = (1 - c) \cdot [\![f < c]\!] \qquad \ell_{\mathrm{CS}}(-1, f; c) = c \cdot [\![f > c]\!]$$

Every proper loss is a mixture of cost-sensitive losses:

$$\ell(y, f) = \int_0^1 w(c) \cdot \ell_{\mathrm{CS}}(y, f; c)\, \mathrm{d}c.$$

# Weight functions for proper losses

Consider the cost-sensitive loss

$$\ell_{\text{CS}}(+1, f; c) = (1 - c) \cdot [\![f < c]\!] \qquad \ell_{\text{CS}}(-1, f; c) = c \cdot [\![f > c]\!]$$

Every proper loss is a mixture of cost-sensitive losses:

$$\ell(y, f) = \int_0^1 w(c) \cdot \ell_{\text{CS}}(y, f; c) \, \mathrm{d}c.$$

The weight function $w$ determines modelling effort

# Weight functions for proper losses

Consider the cost-sensitive loss

$$\ell_{\text{CS}}(+1, f; c) = (1 - c) \cdot [\![f < c]\!] \qquad \ell_{\text{CS}}(-1, f; c) = c \cdot [\![f > c]\!]$$

Every proper loss is a mixture of cost-sensitive losses:

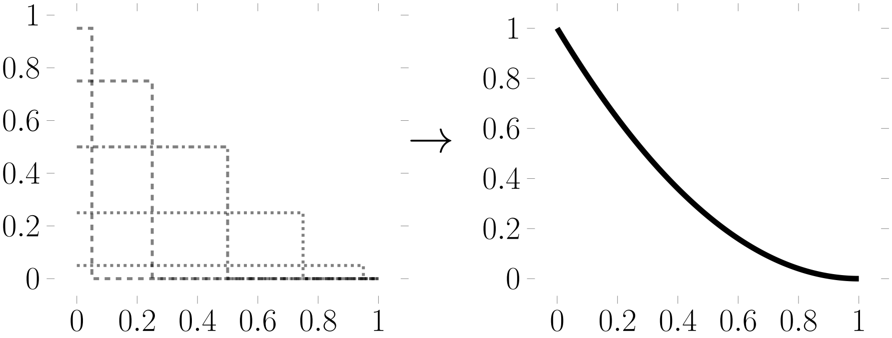
$$\ell(y, f) = \int_0^1 w(c) \cdot \ell_{\text{CS}}(y, f; c) \, \mathrm{d}c.$$

The weight function $w$ determines modelling effort

**Choose a weight which emphasises small $c$ values**

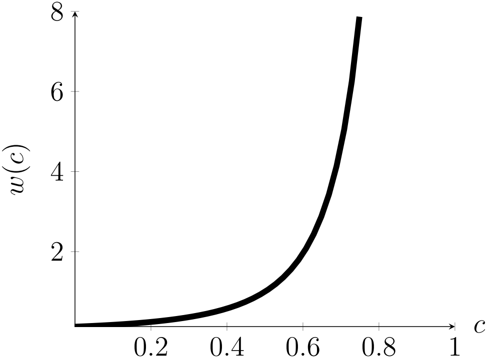
# Weight functions for proper losses

For square loss, $w(c) = 1$, i.e., all costs are equal



# Weight functions for proper losses

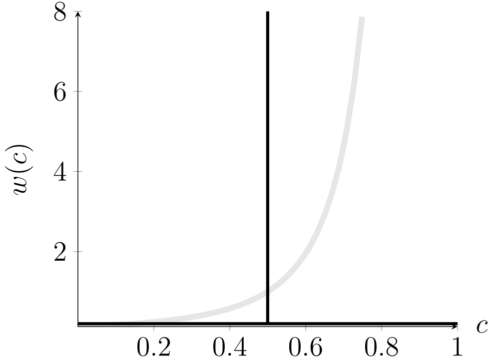
For the LSIF loss, we have smooth $w(c) = (1-c)^{-3}$



# Weight functions for proper losses

For the LSIF loss, we have smooth $w(c) = (1-c)^{-3}$

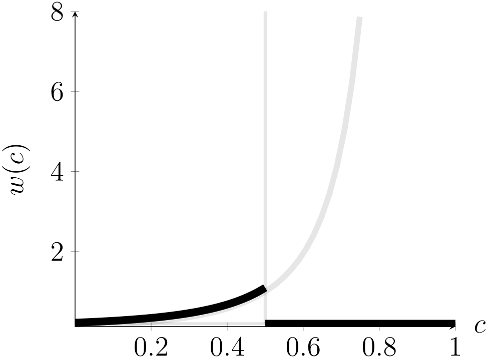
cost-sensitive loss, we have delta-function $w(c) = \delta_{c_0}(c)$



# Weight functions for proper losses

For the LSIF loss, we have smooth $w(c) = (1-c)^{-3}$

cost-sensitive loss, we have delta-function $w(c) = \delta_{c_0}(c)$




Natural intermediary: weight with partial support

# Partially supported weight functions

Fix a proper loss $\ell$ with weight function $w$

# Partially supported weight functions

Fix a proper loss $\ell$ with weight function $w$

Suppose for $c_0 \in (0,1)$, we modify the weight to

$$\bar{w}(c) = [\![ c \leq c_0 ]\!] \cdot w(c)$$

# Partially supported weight functions

Fix a proper loss $\ell$ with weight function $w$

Suppose for $c_0 \in (0,1)$, we modify the weight to

$$\bar{w}(c) = [\![ c \leq c_0 ]\!] \cdot w(c)$$

**Fact**

For $\alpha = \frac{c_0}{1-c_0}$, the loss corresponding to $\bar{w}$ is

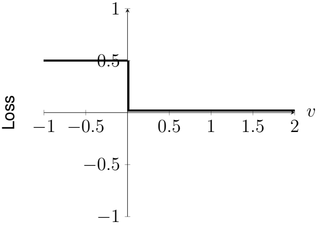
$$\bar{\ell}(+1, f) = \ell(+1, f \wedge \alpha) \qquad \bar{\ell}(-1, f) = \ell(-1, f \wedge \alpha)$$

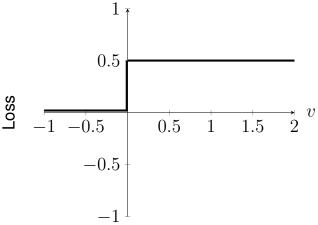
Effect is to saturate the losses

# Partially supported weight functions

Consider the cost-sensitive loss with $c_0 = \frac{1}{2}$,

$$\ell(+1, f) = \frac{1}{2} \cdot [\![ f < 0 ]\!] \qquad \ell(-1, f) = \frac{1}{2} \cdot [\![ f > 0 ]\!]$$

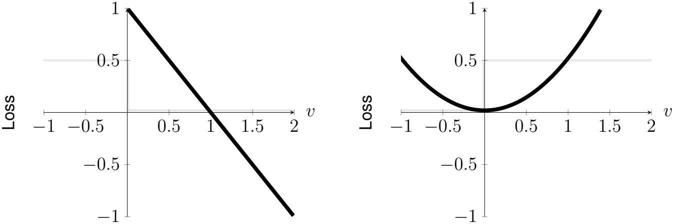


# Partially supported weight functions

Consider the LSIF loss

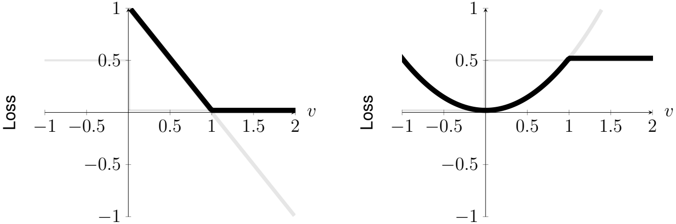
$$\ell(+1,f) = 1 - f \qquad \ell(-1,f) = \frac{1}{2} \cdot f^2$$



# Partially supported weight functions

Consider the modified LSIF loss

$$\ell(+1,f) = 1 - (f \wedge 1) \qquad \ell(-1,f) = \frac{1}{2} \cdot (f \wedge 1)^2$$



# Partially proper losses

For the LSIF loss, the modified version

$$\bar{\ell}(+1,f) = [\alpha - f]_+ \qquad \bar{\ell}(-1,f) = \frac{1}{2} \cdot (f \wedge \alpha)^2$$

is partially proper in the following sense

# Partially proper losses

For the LSIF loss, the modified version

$$\bar{\ell}(+1, f) = [\alpha - f]_+ \qquad \bar{\ell}(-1, f) = \frac{1}{2} \cdot (f \wedge \alpha)^2$$

is partially proper in the following sense

**Fact**

The optimal prediction under $\bar{\ell}$ is

$$f(x) \in \begin{cases} [\alpha, +\infty) & \text{if } p(x) > \alpha \\ p(x) & \text{if } p(x) < \alpha \end{cases}$$

Exactly as desired for partial density estimation!

# Partially proper losses

For the LSIF loss, consider a further modification

$$\tilde{\ell}(+1, f) = [\alpha - f]_+ \qquad \tilde{\ell}(-1, f) = \frac{1}{2} \cdot f^2$$

- only saturate the loss on positives

# Partially proper losses

For the LSIF loss, consider a further modification

$$\tilde{\ell}(+1,f) = [\alpha - f]_+ \qquad \tilde{\ell}(-1,f) = \frac{1}{2} \cdot f^2$$

- only saturate the loss on positives

**Fact**

The optimal prediction under $\tilde{\ell}$ is

$$f(x) \in \begin{cases} \alpha & \text{if } p(x) > \alpha \\ p(x) & \text{if } p(x) < \alpha \end{cases}$$

# Partially proper losses

For the LSIF loss, consider a further modification

$$\tilde{\ell}(+1, f) = [\alpha - f]_+ \qquad \tilde{\ell}(-1, f) = \frac{1}{2} \cdot f^2$$

- only saturate the loss on positives

**Fact**

The optimal prediction under $\tilde{\ell}$ is

$$f(x) \in \begin{cases} \alpha & \text{if } p(x) > \alpha \\ p(x) & \text{if } p(x) < \alpha \end{cases}$$

Perform **capped density estimation**

- no longer have full flexibility for high density area

# Comparison to one-class SVMs

For data distribution $P$, the OC-SVM solves

$$\min_{f,\alpha} \underbrace{\mathbb{E}_{P}[\alpha - f(\mathsf{X})]_{+}}_{\text{hinge loss}} + \underbrace{\frac{\nu}{2} \cdot \|f\|_{\mathcal{H}}^2}_{\text{regulariser}} - \underbrace{\nu \cdot \alpha}_{\nu-\text{SVM relic}}$$

# Comparison to one-class SVMs

For data distribution $P$, the OC-SVM solves

$$\min_{f,\alpha} \underbrace{\mathbb{E}_{P}[\alpha - f(\mathsf{X})]_+}_{\text{hinge loss}} + \underbrace{\frac{\nu}{2} \cdot \|f\|_{\mathcal{H}}^2}_{\text{regulariser}} - \underbrace{\nu \cdot \alpha}_{\nu-\text{SVM relic}}$$

while we solve

$$\min_{f} \underbrace{\mathbb{E}_{P}[\alpha - f(\mathsf{X})]_+}_{\textbf{capped proper loss}} + \underbrace{\mathbb{E}_{Q} \frac{1}{2} \cdot f(\mathsf{X})^2}_{\textbf{background contrast}}$$

# This talk

**Take-home #1** ✓

Anomaly detection = binary classification

- distinguish samples against an implicit background

**Take-home #2** ✓

Probabilistic anomaly detection = class-probability estimation

- can use familiar tools: logistic regression, boosting, . . .

# This talk

**Take-home #1** ✓

Anomaly detection = binary classification

- distinguish samples against an implicit background

**Take-home #2** ✓

Probabilistic anomaly detection = class-probability estimation

- can use familiar tools: logistic regression, boosting, ...

**Surprise** ?

Specific kind of OC-SVM turns out to be a special case!

- gives a different perspective on underlying components

# Kernel absorption

# Partial density estimation

To obtain tail density probabilities, we propose to minimise

$$\min_f \mathbb{E}_P [\alpha - f(\mathsf{X})]_+ + \mathbb{E}_Q \frac{1}{2} \cdot f(\mathsf{X})^2$$

# Partial density estimation

To obtain tail density probabilities, we propose to minimise

$$\min_f \mathop{\mathbb{E}}_{P} [\alpha - f(\mathsf{X})]_+ + \mathop{\mathbb{E}}_{Q} \frac{1}{2} \cdot f(\mathsf{X})^2$$

Practically, we may pick $f$ from an RKHS $\mathcal{H}$ via

$$\min_f \mathop{\mathbb{E}}_{P} [\alpha - f(\mathsf{X})]_+ + \mathop{\mathbb{E}}_{Q} \frac{1}{2} \cdot f(\mathsf{X})^2 + \frac{\gamma}{2} \cdot \|f\|_{\mathcal{H}}^2$$

# Partial density estimation

To obtain tail density probabilities, we propose to minimise

$$\min_f \mathop{\mathbb{E}}_{P} [\alpha - f(\mathsf{X})]_+ + \mathop{\mathbb{E}}_{Q} \frac{1}{2} \cdot f(\mathsf{X})^2$$

Practically, we may pick $f$ from an RKHS $\mathcal{H}$ via

$$\min_f \mathop{\mathbb{E}}_{P} [\alpha - f(\mathsf{X})]_+ + \mathop{\mathbb{E}}_{Q} \frac{1}{2} \cdot f(\mathsf{X})^2 + \frac{\gamma}{2} \cdot \|f\|_{\mathcal{H}}^2$$

Convex, but requires computing a high-dimensional integral

# Partial density estimation

To obtain tail density probabilities, we propose to minimise

$$\min_f \mathbb{E}_P [\alpha - f(\mathsf{X})]_+ + \mathbb{E}_Q \frac{1}{2} \cdot f(\mathsf{X})^2$$

Practically, we may pick $f$ from an RKHS $\mathcal{H}$ via

$$\min_f \mathbb{E}_P [\alpha - f(\mathsf{X})]_+ + \mathbb{E}_Q \frac{1}{2} \cdot f(\mathsf{X})^2 + \frac{\gamma}{2} \cdot \|f\|_{\mathcal{H}}^2$$

Convex, but requires computing a high-dimensional integral

A simple trick lets us side-step this

# A kernel trick

Observe that

$$\mathop{\mathbb{E}}_{Q} \frac{1}{2} \cdot f(\mathsf{X})^2 + \frac{\gamma}{2} \cdot \|f\|_{\mathcal{H}}^2 = \|f\|_{L_2(\mu)}^2 + \frac{\gamma}{2} \cdot \|f\|_{\mathcal{H}}^2$$

- standard plus Hilbert-space square norm

# A kernel trick

Observe that

$$\mathbb{E}_{Q} \frac{1}{2} \cdot f(\mathsf{X})^2 + \frac{\gamma}{2} \cdot \|f\|_{\mathcal{H}}^2 = \|f\|_{L_2(\mu)}^2 + \frac{\gamma}{2} \cdot \|f\|_{\mathcal{H}}^2$$

- standard plus Hilbert-space square norm

Fortuitously, we can write (McCullagh and Møller, 2006)

$$\|f\|_{L_2(\mu)}^2 + \gamma \cdot \|f\|_{\mathcal{H}}^2 = \|f\|_{\bar{\mathcal{H}}(\gamma,\mu)}^2$$

for some modified RKHS $\bar{\mathcal{H}}(\gamma, \mu)$

- corresponding kernel $\bar{k}$ modifies eigenvalues of $k$

# A kernel trick

Observe that

$$\mathop{\mathbb{E}}_{Q} \frac{1}{2} \cdot f(\mathsf{X})^2 + \frac{\gamma}{2} \cdot \|f\|_{\mathcal{H}}^2 = \|f\|_{L_2(\mu)}^2 + \frac{\gamma}{2} \cdot \|f\|_{\mathcal{H}}^2$$

- standard plus Hilbert-space square norm

Fortuitously, we can write (McCullagh and Møller, 2006)

$$\|f\|_{L_2(\mu)}^2 + \gamma \cdot \|f\|_{\mathcal{H}}^2 = \|f\|_{\bar{\mathcal{H}}(\gamma,\mu)}^2$$

for some modified RKHS $\bar{\mathcal{H}}(\gamma, \mu)$

- corresponding kernel $\bar{k}$ modifies eigenvalues of $k$

This obviates the need for approximating the expectation!

# A kernel trick: comments

Connection to point processes is unsurprising

- latter is scaled density estimation (Fithian & Hastie, 2013)

# A kernel trick: comments

Connection to point processes is unsurprising

- latter is scaled density estimation (Fithian & Hastie, 2013)

Penalty $\|f\|^2_{\mathcal{H}(\gamma,\mu)}$ bakes in measure $\mu$ and regulariser

- model complexity plus discrimination

# A kernel trick: comments

Connection to point processes is unsurprising

- latter is scaled density estimation (Fithian & Hastie, 2013)

Penalty $\|f\|^2_{\bar{\mathcal{H}}(\gamma,\mu)}$ bakes in measure $\mu$ and regulariser

- model complexity plus discrimination

New kernel $\bar{k}$ may not have analytic form

- can approximate with Nyström method

# Comparison to one-class SVMs

For data distribution $P$, the OC-SVM solves

$$\min_{f,\alpha} \underbrace{\mathbb{E}_{P}[\alpha - f(\mathsf{X})]_+}_{\text{hinge loss}} + \underbrace{\frac{\nu}{2} \cdot \|f\|_{\mathcal{H}}^2}_{\text{regulariser}} - \underbrace{\nu \cdot \alpha}_{\nu-\text{SVM relic}}$$

# Comparison to one-class SVMs

For data distribution $P$, the OC-SVM solves

$$\min_{f,\alpha} \underbrace{\mathop{\mathbb{E}}_{P}[\alpha - f(\mathsf{X})]_+}_{\text{hinge loss}} + \underbrace{\frac{\nu}{2} \cdot \|f\|^2_{\mathcal{H}}}_{\text{regulariser}} - \underbrace{\nu \cdot \alpha}_{\nu-\text{SVM relic}}$$

while we solve

$$\min_{f} \underbrace{\mathop{\mathbb{E}}_{P}[\alpha - f(\mathsf{X})]_+}_{\textbf{capped proper loss}} + \underbrace{\frac{1}{2} \cdot \|f\|^2_{\bar{\mathcal{H}}(\gamma,\mu)}}_{\text{regulariser}}$$

**+ background contrast**

# Comparison to one-class SVMs

For data distribution $P$, the OC-SVM solves

$$\min_{f,\alpha} \underbrace{\mathbb{E}_{P}[\alpha - f(\mathsf{X})]_+}_{\text{hinge loss}} + \underbrace{\frac{\nu}{2} \cdot \|f\|_{\mathcal{H}}^2}_{\text{regulariser}} - \underbrace{\nu \cdot \alpha}_{\nu-\text{SVM relic}}$$

while we solve

$$\min_{f} \underbrace{\mathbb{E}_{P}[\alpha - f(\mathsf{X})]_+}_{\textbf{capped proper loss}} + \underbrace{\frac{1}{2} \cdot \|f\|_{\bar{\mathcal{H}}(\gamma,\mu)}^2}_{\substack{\text{regulariser} \\ \textbf{+ background contrast}}}$$

How do we control the threshold $\alpha$?

# Alarm rate control

# Parametrising anomaly level

To obtain tail density probabilities, we propose to minimise

$$\min_f \mathbb{E}_{P}[\alpha - f(\mathsf{X})]_+ + \frac{1}{2} \cdot \|f\|^2_{\mathcal{H}(\gamma,\mu)}$$

# Parametrising anomaly level

To obtain tail density probabilities, we propose to minimise

$$\min_f \mathbb{E}_P [\alpha - f(\mathsf{X})]_+ + \frac{1}{2} \cdot \|f\|^2_{\mathcal{H}(\gamma,\mu)}$$

Choice of $\alpha$ determines density threshold

# Parametrising anomaly level

To obtain tail density probabilities, we propose to minimise

$$\min_f \mathbb{E}_P [\alpha - f(\mathsf{X})]_+ + \frac{1}{2} \cdot \|f\|^2_{\mathcal{H}(\gamma,\mu)}$$

Choice of $\alpha$ determines density threshold

More intuitive: given $\nu \in (0,1)$, implicitly use $\alpha_\nu$ such that

$$P(p(\mathsf{X}) < \alpha_\nu) = \nu$$

- quantile of the random variable $p(\mathsf{X})$
- $\nu$ specifies the alarm rate of our predictor

# Pinball loss

Recall that the median $\alpha_{1/2}$ of a distribution $P$ is

$$\alpha_{1/2} = \operatorname*{argmin}_{\alpha \in \mathbb{R}} \mathop{\mathbb{E}}_{P} |\mathsf{X} - \alpha|$$

# Pinball loss

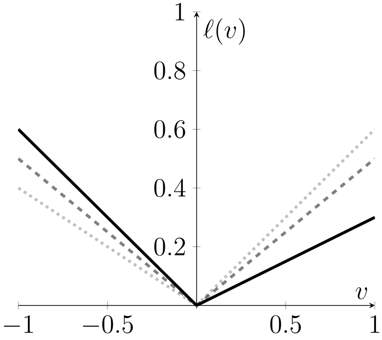
Recall that the median $\alpha_{1/2}$ of a distribution $P$ is

$$\alpha_{1/2} = \operatorname*{argmin}_{\alpha \in \mathbb{R}} \mathop{\mathbb{E}}_{P} |\mathsf{X} - \alpha|$$

More generally, the $\nu$th quantile of a distribution $P$ is

$$\alpha_\nu = \operatorname*{argmin}_{\alpha \in \mathbb{R}} \mathop{\mathbb{E}}_{P} \left[\phi_{\text{pin}}(\mathsf{X} - \alpha; \nu)\right]$$

for the pinball loss $\phi_{\text{pin}}$



# Relating the hinge and pinball loss

**Fact**

The pinball loss is equivalently

$$\phi_{\text{pin}}(z; \nu) = [z]_+ + \nu \cdot z$$

# Relating the hinge and pinball loss

**Fact**

The pinball loss is equivalently

$$\phi_{\text{pin}}(z; \nu) = [z]_+ + \nu \cdot z$$

Thus, we have

$$\mathop{\mathbb{E}}_{P}[\alpha - f(\mathsf{X})]_+ = \mathop{\mathbb{E}}_{P}\left[\phi_{\text{pin}}(f(\mathsf{X}) - \alpha; \nu)\right] - \nu \cdot \mathop{\mathbb{E}}_{P}[f(\mathsf{X})] + \nu \cdot \alpha$$

# Relating the hinge and pinball loss

**Fact**

The pinball loss is equivalently

$$\phi_{\text{pin}}(z; \nu) = [z]_+ + \nu \cdot z$$

Thus, we have

$$\mathop{\mathbb{E}}_{P}[\alpha - f(\mathsf{X})]_+ - \nu \cdot \alpha = \mathop{\mathbb{E}}_{P}\left[\phi_{\text{pin}}(f(\mathsf{X}) - \alpha; \nu)\right] - \nu \cdot \mathop{\mathbb{E}}_{P}[f(\mathsf{X})]$$

# Relating the hinge and pinball loss

**Fact**

The pinball loss is equivalently

$$\phi_{\text{pin}}(z; \nu) = [z]_+ + \nu \cdot z$$

Thus, we have

$$\mathop{\mathbb{E}}_{P}[\alpha - f(\mathsf{X})]_+ - \nu \cdot \alpha = \mathop{\mathbb{E}}_{P}\left[\phi_{\text{pin}}(f(\mathsf{X}) - \alpha; \nu)\right] - \nu \cdot \mathop{\mathbb{E}}_{P}[f(\mathsf{X})]$$

Thus, we may jointly minimise

$$\min_{f, \alpha} \mathop{\mathbb{E}}_{P}\left[\phi_{\text{pin}}(f(\mathsf{X}) - \alpha; \nu)\right] - \nu \cdot \mathop{\mathbb{E}}_{P}[f(\mathsf{X})] + \frac{1}{2} \cdot \|f\|_{\mathcal{H}}^2$$

# Relating the hinge and pinball loss

**Fact**

The pinball loss is equivalently

$$\phi_{\text{pin}}(z; \nu) = [z]_+ + \nu \cdot z$$

Thus, we have

$$\mathop{\mathbb{E}}_{P}[\alpha - f(\mathsf{X})]_+ - \nu \cdot \alpha = \mathop{\mathbb{E}}_{P}\left[\phi_{\text{pin}}(f(\mathsf{X}) - \alpha; \nu)\right] - \nu \cdot \mathop{\mathbb{E}}_{P}[f(\mathsf{X})]$$

Thus, we may jointly minimise

$$\min_{f, \alpha} \mathop{\mathbb{E}}_{P}[\alpha - f(\mathsf{X})]_+ - \nu \cdot \alpha + \frac{1}{2} \cdot \|f\|_{\mathcal{H}}^2$$

# Relating the hinge and pinball loss

**Fact**

The pinball loss is equivalently

$$\phi_{\text{pin}}(z; \nu) = [z]_+ + \nu \cdot z$$

Thus, we have

$$\mathbb{E}_{P}[\alpha - f(\mathsf{X})]_+ - \nu \cdot \alpha = \mathbb{E}_{P}\left[\phi_{\text{pin}}(f(\mathsf{X}) - \alpha; \nu)\right] - \nu \cdot \mathbb{E}_{P}[f(\mathsf{X})]$$

Thus, we may jointly minimise

$$\min_{f, \alpha} \mathbb{E}_{P}[\alpha - f(\mathsf{X})]_+ - \nu \cdot \alpha + \frac{1}{2} \cdot \|f\|_{\mathcal{H}}^2$$

and obtain $\alpha^*$ as the $\nu$th quantile of $f^*(\mathsf{X})$!

# Summary: deconstructing one-class SVMs

For data distribution $P$, the OC-SVM solves

$$\min_{f,\alpha} \underbrace{\mathop{\mathbb{E}}_{P}[\alpha - f(\mathsf{X})]_+}_{\text{hinge loss}} + \underbrace{\frac{\nu}{2} \cdot \|f\|_{\mathcal{H}}^2}_{\text{regulariser}} - \underbrace{\nu \cdot \alpha}_{\nu-\text{SVM relic}}$$

# Summary: deconstructing one-class SVMs

For data distribution $P$, the OC-SVM solves

$$\min_{f,\alpha} \underbrace{\mathbb{E}_P [\alpha - f(\mathsf{X})]_+}_{\text{hinge loss}} + \underbrace{\frac{\nu}{2} \cdot \|f\|^2_{\mathcal{H}}}_{\text{regulariser}} - \underbrace{\nu \cdot \alpha}_{\nu-\text{SVM relic}}$$

while we solve

$$\min_{f,\alpha} \underbrace{\mathbb{E}_P [\alpha - f(\mathsf{X})]_+}_{\textbf{capped proper loss}} + \underbrace{\frac{1}{2} \cdot \|f\|^2_{\bar{\mathcal{H}}(\mu,\gamma)}}_{\substack{\textbf{regulariser} \\ +\textbf{background contrast}}} - \underbrace{\nu \cdot \alpha}_{\textbf{pinball loss}}$$

# Summary: deconstructing one-class SVMs

For data distribution $P$, the OC-SVM solves

$$\min_{f,\alpha} \underbrace{\mathbb{E}_{P}[\alpha - f(\mathsf{X})]_+}_{\text{hinge loss}} + \underbrace{\frac{\nu}{2} \cdot \|f\|_{\mathcal{H}}^2}_{\text{regulariser}} - \underbrace{\nu \cdot \alpha}_{\nu-\text{SVM relic}}$$

while we solve

$$\min_{f,\alpha} \underbrace{\mathbb{E}_{P}[\alpha - f(\mathsf{X})]_+}_{\textbf{capped proper loss}} + \underbrace{\frac{1}{2} \cdot \|f\|_{\bar{\mathcal{H}}(\mu,\gamma)}^2}_{\substack{\textbf{regulariser} \\ +\textbf{background contrast}}} - \underbrace{\nu \cdot \alpha}_{\textbf{pinball loss}}$$

Note this is just one special case of our framework

# Empirical illustration

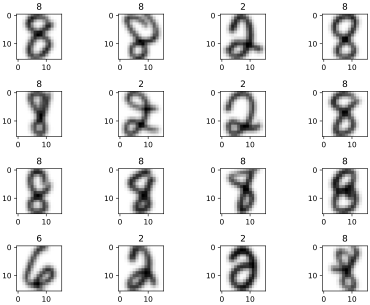
# Qualitative results

Augment `usps` test instances with one-hot encoding of label

# Qualitative results

Augment `usps` test instances with one-hot encoding of label

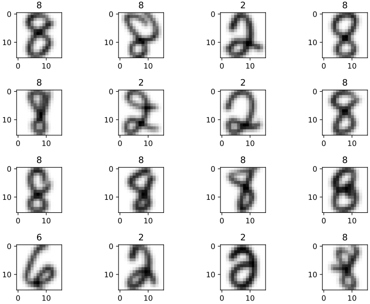
Identify inliers

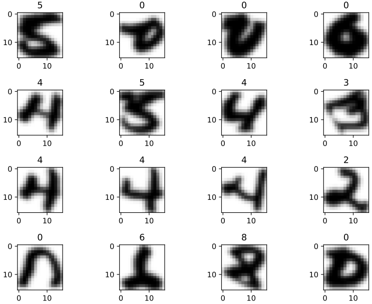


# Qualitative results

Augment `usps` test instances with one-hot encoding of label

Identify inliers and outliers



# Quantitative results

We fit our model to a "normal" sample on three datasets

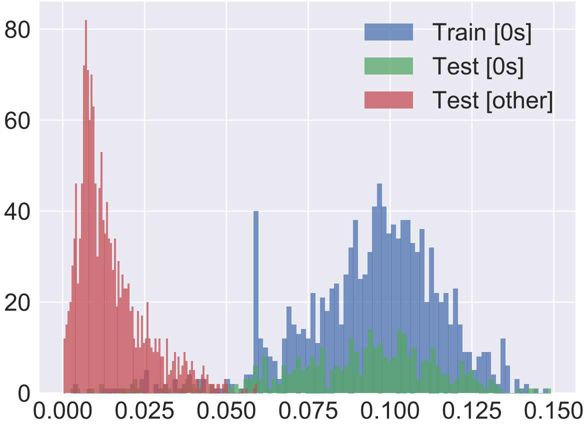
- `usps`: digit 0
- `sat`: largest 3 classes
- `art`: $\sim$ mixture of Gaussians

Evaluate classification performance on a test sample of normal and anomalous instances

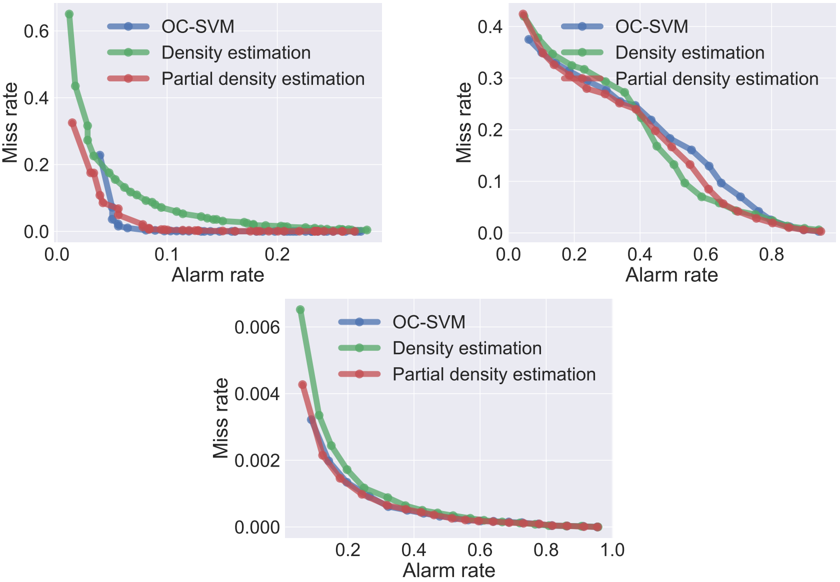
# Quantitative results: `usps` score distribution

Scores for digit $0$ on train and test set largely agree

Scores for digit $1$–$9$ distinct, despite being unseen at train time



# Quantitative results: alarm-miss curves



# Summary

# This talk

**Take-home #1**

Anomaly detection = binary classification

- distinguish samples against an implicit background

**Take-home #2**

Probabilistic anomaly detection = class-probability estimation

- can use familiar tools: logistic regression, boosting, ...

**Surprise**

Specific kind of OC-SVM turns out to be a special case!

- gives a different perspective on underlying components

# Deconstructing one-class SVMs

Pick an RKHS $\mathcal{H}$ and desired anomaly fraction $\nu \in (0, 1)$

For data distribution $P$, the OC-SVM solves

$$\min_{f,\alpha} \underbrace{\mathop{\mathbb{E}}_{P}[\alpha - f(\mathsf{X})]_+}_{\text{hinge loss}} + \underbrace{\frac{\nu}{2} \cdot \|f\|_{\mathcal{H}}^2}_{\text{regulariser}} - \underbrace{\nu \cdot \alpha}_{\nu-\text{SVM relic}}$$

# Deconstructing one-class SVMs

Pick an RKHS $\mathcal{H}$ and desired anomaly fraction $\nu \in (0,1)$

For data distribution $P$, the OC-SVM solves

$$\min_{f,\alpha} \underbrace{\mathop{\mathbb{E}}_{P}[\alpha - f(\mathsf{X})]_+}_{\substack{\text{~~hinge loss~~} \\ \textbf{capped proper loss}}} + \underbrace{\frac{\nu}{2} \cdot \|f\|^2_{\mathcal{H}}}_{\substack{\text{regulariser} \\ +\textbf{background contrast}}} - \underbrace{\nu \cdot \alpha}_{\substack{\nu\text{~~-SVM relic~~} \\ \textbf{pinball loss}}}$$

# Deconstructing one-class SVMs

Pick an RKHS $\mathcal{H}$ and desired anomaly fraction $\nu \in (0,1)$

For data distribution $P$, the OC-SVM solves

$$\min_{f,\alpha} \underbrace{\mathbb{E}_P[\alpha - f(\mathsf{X})]_+}_{\substack{\text{~~hinge loss~~} \\ \textbf{capped proper loss}}} + \underbrace{\frac{\nu}{2} \cdot \|f\|^2_{\mathcal{H}}}_{\substack{\text{regulariser} \\ +\textbf{background contrast}}} - \underbrace{\nu \cdot \alpha}_{\substack{\nu\text{~~-SVM relic~~} \\ \textbf{pinball loss}}}$$

Questions nonetheless remain:

- implicit $\mu, \gamma$ for Gaussian kernel?
- avoiding need for density for minimum volume sets?
- link interpretation of robust versions of loss?

# Thanks!