A loss framework for calibrated anomaly detection

Aditya Krishna Menon   Robert C. Williamson

Australian National University

Dec 5th, 2018

Anomaly detection

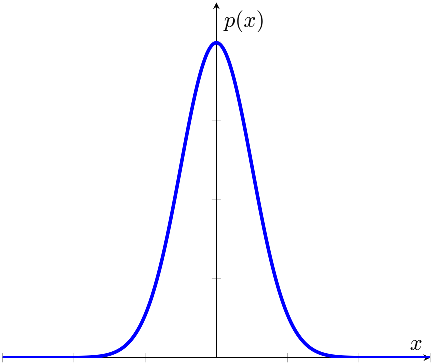
Identify instances that deviate from some systematic pattern

Anomaly detection

Identify instances that deviate from some systematic pattern

A density sublevel view

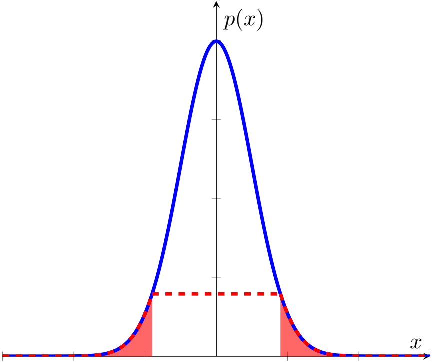
Suppose our data distribution P has density $p \doteq \frac{\mathrm{d}P}{\mathrm{d}\mu}$



A density sublevel view

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

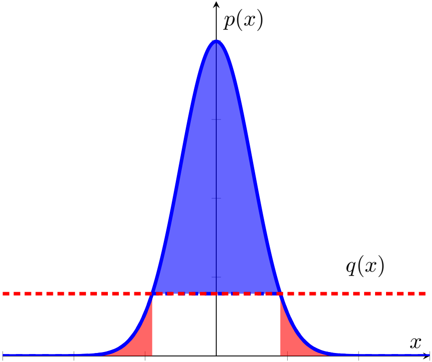
Anomalies are instances with low density



A density sublevel view

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

Anomalies are instances with low density relative to uniform Q



Classify data against background (Steinwart & Scovel, '05)

A classification view

Pick density threshold $\alpha > 0$, and classify data P vs background Q:

$$\min_f \mathop{\mathbb{E}}_P \ell_{\mathrm{CS}}(+1, f; c) + \mathop{\mathbb{E}}_Q \ell_{\mathrm{CS}}(-1, f; c)$$

for cost-sensitive loss ℓ_{CS} with cost-ratio $c = \alpha/(1+\alpha)$

A classification view

Pick density threshold $\alpha > 0$, and classify data P vs background Q:

$$\min_f \mathop{\mathbb{E}}_P \ell_{\mathrm{CS}}(+1, f; c) + \mathop{\mathbb{E}}_Q \ell_{\mathrm{CS}}(-1, f; c)$$

for cost-sensitive loss ℓ_{CS} with cost-ratio $c = \alpha/(1+\alpha)$

Appealing, but with limitations:

Issue
Need sampling for $\mathop{\mathbb{E}}_Q f(\mathsf{X}) = \int_{\mathcal{X}} f(x)\, \mathrm{d}Q(x)$

A classification view

Pick density threshold $\alpha > 0$, and classify data P vs background Q:

$$\min_f \mathop{\mathbb{E}}_P \ell_{\mathrm{CS}}(+1, f; c) + \mathop{\mathbb{E}}_Q \ell_{\mathrm{CS}}(-1, f; c)$$

for cost-sensitive loss ℓ_{CS} with cost-ratio $c = \alpha/(1+\alpha)$

Appealing, but with limitations:

Issue
Need sampling for $\mathop{\mathbb{E}}_Q f(\mathsf{X}) = \int_{\mathcal{X}} f(x)\,\mathrm{d}Q(x)$
Scale of $\alpha \to$ scale of $p(\cdot)$

A classification view

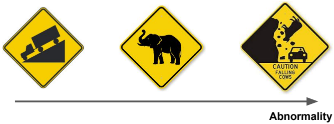
Pick density threshold $\alpha > 0$, and classify data P vs background Q:

$$\min_f \mathop{\mathbb{E}}_P \ell_{\mathrm{CS}}(+1, f; c) + \mathop{\mathbb{E}}_Q \ell_{\mathrm{CS}}(-1, f; c)$$

for cost-sensitive loss ℓ_{CS} with cost-ratio $c = \alpha/(1+\alpha)$

Appealing, but with limitations:

Issue
Need sampling for $\mathop{\mathbb{E}}_Q f(\mathsf{X}) = \int_{\mathcal{X}} f(x)\,\mathrm{d}Q(x)$
Scale of $\alpha \to$ scale of $p(\cdot)$
Doesn't yield confidence scores



A classification view

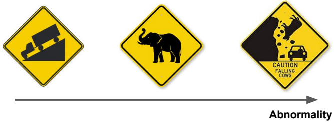
Pick density threshold $\alpha > 0$, and classify data P vs background Q:

$$\min_f \mathop{\mathbb{E}}_{P} \ell_{\text{CS}}(+1, f; c) + \mathop{\mathbb{E}}_{Q} \ell_{\text{CS}}(-1, f; c)$$

for cost-sensitive loss ℓ_{CS} with cost-ratio $c = \alpha/(1+\alpha)$

Appealing, but with limitations:

Issue	**Resolution**
Need sampling for $\mathop{\mathbb{E}}_{Q} f(\mathsf{X}) = \int_{\mathcal{X}} f(x)\, \mathrm{d}Q(x)$	A kernel trick
Scale of $\alpha \to$ scale of $p(\cdot)$	Pinball loss
Doesn't yield confidence scores	Capped proper loss



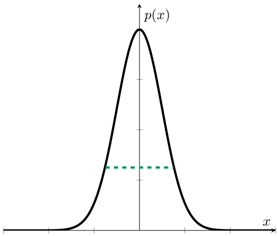
Capped proper losses

Intuitively, confidence scores are $\propto p(\cdot)^{-1}$

To obtain a single sublevel set of $p(\cdot)$, use

$$\min_f \mathbb{E}_P \ell(+1,f) + \mathbb{E}_Q \ell(-1,f)$$

$$\ell(y,f) = \ell_{\mathrm{CS}}(y,f;c)$$



$\times$No confidences

Capped proper losses

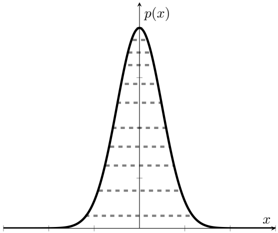
Intuitively, confidence scores are $\propto p(\cdot)^{-1}$

To obtain all sublevel sets of $p(\cdot)$, use

$$\min_f \mathbb{E}_P \, \ell(+1, f) + \mathbb{E}_Q \, \ell(-1, f)$$

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

for positive weight function w; yields proper losses



✓Confidences for **all** instances

Capped proper losses

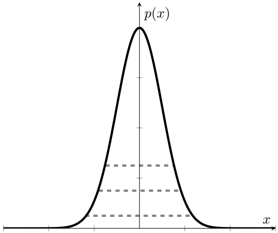
Intuitively, confidence scores are $\propto p(\cdot)^{-1}$

To obtain **tail** sublevel sets of $p(\cdot)$, use

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

$$\ell(y, f) = \int_0^1 [\![c \leq c_0]\!] \cdot w(c) \cdot \ell_{\mathrm{CS}}(y, f; c)\, \mathrm{d}c$$

for positive weight function w; yields **capped** proper losses



✓Confidences for **anomalous** instances

Capped proper losses

Fact

Focussing on the tail sublevel sets results in **capping** the loss

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

Capped proper losses

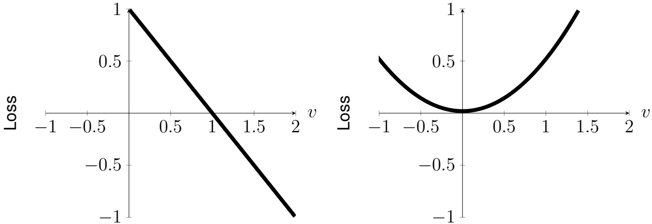
Fact

Focussing on the tail sublevel sets results in **capping** the loss

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

An admissible example is

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



Capped proper losses

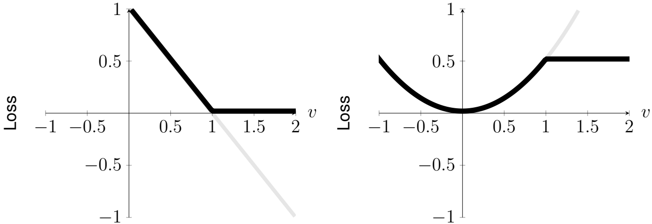
Fact

Focussing on the tail sublevel sets results in **capping** the loss

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

An admissible example is

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



Quantile control

One can remove cap on $\ell(-1, \cdot)$, yielding e.g.

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

for fixed density threshold $\alpha > 0$

Quantile control

One can remove cap on $\ell(-1, \cdot)$, yielding e.g.

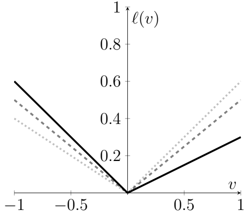
$$\min_{f} \mathbb{E}_{P} [\alpha - f(\mathsf{X})]_+ + \frac{1}{2} \cdot \mathbb{E}_{Q} f(\mathsf{X})^2$$

for fixed density threshold $\alpha > 0$

Can learn α: for anomaly fraction $\nu \in (0, 1)$, find

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

- last term arises from pinball loss
- α^* will be the νth quantile of $f^*(\mathsf{X})$



A (different) kernel trick

The background loss can be written

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

A (different) kernel trick

The background loss can be written

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

$$= \min_{f,\alpha} \mathop{\mathbb{E}}_{P} [\alpha - f(\mathsf{X})]_+ + \frac{1}{2} \cdot \int_{\mathcal{X}} f(x)^2 \, \mathrm{d}Q(x) - \nu \cdot \alpha$$

A (different) kernel trick

The background loss can be written

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

$$= \min_{f,\alpha} \mathbb{E}_{P} [\alpha - f(\mathsf{X})]_+ + \frac{1}{2} \cdot \int_{\mathcal{X}} f(x)^2 \,\mathrm{d}Q(x) - \nu \cdot \alpha$$

$$= \min_{f,\alpha} \mathbb{E}_{P} [\alpha - f(\mathsf{X})]_+ + \frac{1}{2} \cdot \|f\|^2_{L_2(Q)} - \nu \cdot \alpha$$

A (different) kernel trick

The background loss can be written

$$\begin{aligned}
&\min_{f,\alpha} \mathop{\mathbb{E}}_{P} [\alpha - f(\mathsf{X})]_+ + \frac{1}{2} \cdot \mathop{\mathbb{E}}_{Q} f(\mathsf{X})^2 - \nu \cdot \alpha \\
&= \min_{f,\alpha} \mathop{\mathbb{E}}_{P} [\alpha - f(\mathsf{X})]_+ + \frac{1}{2} \cdot \int_{\mathcal{X}} f(x)^2 \, \mathrm{d}Q(x) - \nu \cdot \alpha \\
&= \min_{f,\alpha} \mathop{\mathbb{E}}_{P} [\alpha - f(\mathsf{X})]_+ + \frac{1}{2} \cdot \|f\|^2_{L_2(Q)} - \nu \cdot \alpha
\end{aligned}$$

Suppose we commit to using kernelised f:

$$\min_{f \in \mathcal{H}, \alpha \in \mathbb{R}} \mathop{\mathbb{E}}_{P} [\alpha - f(\mathsf{X})]_+ + \frac{1}{2} \cdot \|f\|^2_{L_2(Q)} + \frac{\gamma}{2} \cdot \|f\|^2_{\mathcal{H}} - \nu \cdot \alpha$$

A (different) kernel trick

The background loss can be written

$$\begin{aligned}
&\min_{f,\alpha} \mathbb{E}_{P} [\alpha - f(\mathsf{X})]_+ + \frac{1}{2} \cdot \mathbb{E}_{Q} f(\mathsf{X})^2 - \nu \cdot \alpha \\
&= \min_{f,\alpha} \mathbb{E}_{P} [\alpha - f(\mathsf{X})]_+ + \frac{1}{2} \cdot \int_{\mathcal{X}} f(x)^2 \, \mathrm{d}Q(x) - \nu \cdot \alpha \\
&= \min_{f,\alpha} \mathbb{E}_{P} [\alpha - f(\mathsf{X})]_+ + \frac{1}{2} \cdot \|f\|^2_{L_2(Q)} - \nu \cdot \alpha
\end{aligned}$$

Suppose we commit to using kernelised f:

$$\min_{f \in \mathcal{H}, \alpha \in \mathbb{R}} \mathbb{E}_{P} [\alpha - f(\mathsf{X})]_+ + \frac{1}{2} \cdot \|f\|^2_{L_2(Q)} + \frac{\gamma}{2} \cdot \|f\|^2_{\mathcal{H}} - \nu \cdot \alpha$$

Observed in point processes (McCullagh and Møller, '06) that

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

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

Drop by poster **#766**!

We propose to minimise, for proper loss ℓ,

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

This gives a framework for anomaly detection which:

- avoids sampling for background Q
- provides quantile control
- yields calibrated confidence scores

See paper for experiments