L21: Anomaly Detection

Nov 12, 2025
Data Mining
Jeff M. Phillips

Anomalies

Data X

anomaly $\leq cX$

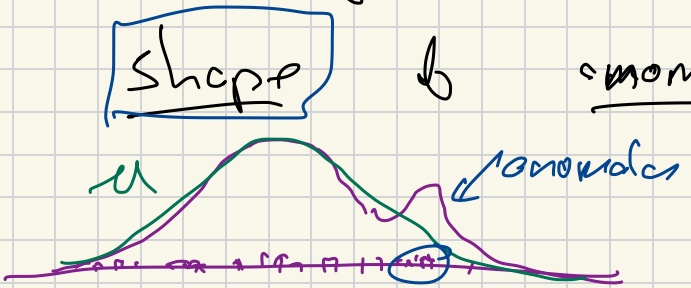
1. What to expect from your data X.

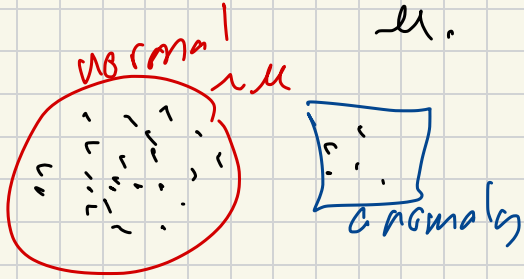
unsupervised.

$X \sim_{iid} u$

2. How what you observe can differ from u.

algos → as shape & anomaly





3. Determine if the most anomalous subset found, is interesting or not.

1. What to expect from data?

$$X = \{x_1, x_2, \ldots x_n\} \overset{iid}{\sim} \mathcal{U}(\theta)$$

parameter(s)

Likelihood $L(X; D(\theta)) = \prod_{x_i \in X} L(x_i; D(\theta))$

$$L(X) = \max_{\theta} L(X; D(\theta))$$

Goal: Define subset $S \subset X$

$$\frac{L(S) \cdot L(X \setminus S)}{L(X)}$$

likelihood ratio

$D(\theta) = \mathcal{N}(\mu, \sigma)$ ← σ is known $\theta = \mu$

$$Pr[x_i \mid D(\theta)] = \frac{1}{\sqrt{2\pi}\cdot\sigma} \cdot \exp\left((\mu - x_i)^2 / 2\sigma^2\right)$$

$\theta = \mu$

$$Pr[X \mid \mu] = \prod_{x_i \in X} Pr[x_i \mid \mu] = \prod_{x_i \in X} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp(x_i - \mu)^2 / 2\sigma^2$$

proportional to

$$\propto \prod_{x_i \in X} \exp\left((x_i - \mu)^2 / 2\sigma^2\right) = L(X; \mu)$$

log Likelihood $\ln(L(x;p))$

$$\prod_{x_i \in X} \exp\left(-(x_i - p)^2/2\sigma^2\right) = L(x;p)$$

$$\ln(L(x;p)) = \ln\left(\prod_{x_i \in X} \exp\left(-(x_i - p)^2/2\sigma^2\right)\right)$$

$$\operatorname{argmax} \ln(L(x;p)) = \sum_{x_i \in X} \ln\left(\exp\left(-(x_i - p)^2/2\sigma^2\right)\right)$$

$$= \operatorname{argmin} -\ln(L(x;p)) = \sum_{x_i \in X} -(x_i - p)^2/2\sigma^2$$

$$= \operatorname{argmin} \sum_{x_i \in X} (x_i - p)^2 = -\frac{1}{2\sigma^2} \sum_{x_i \in X} (x_i - p)^2$$

$$= \bar{p} = \frac{1}{|x|} \sum_{x_i \in X} x_i$$

$X \overset{iid}{\sim} N(\mu, \sigma)$

$L(x; N(\mu, \sigma)) = \dots$

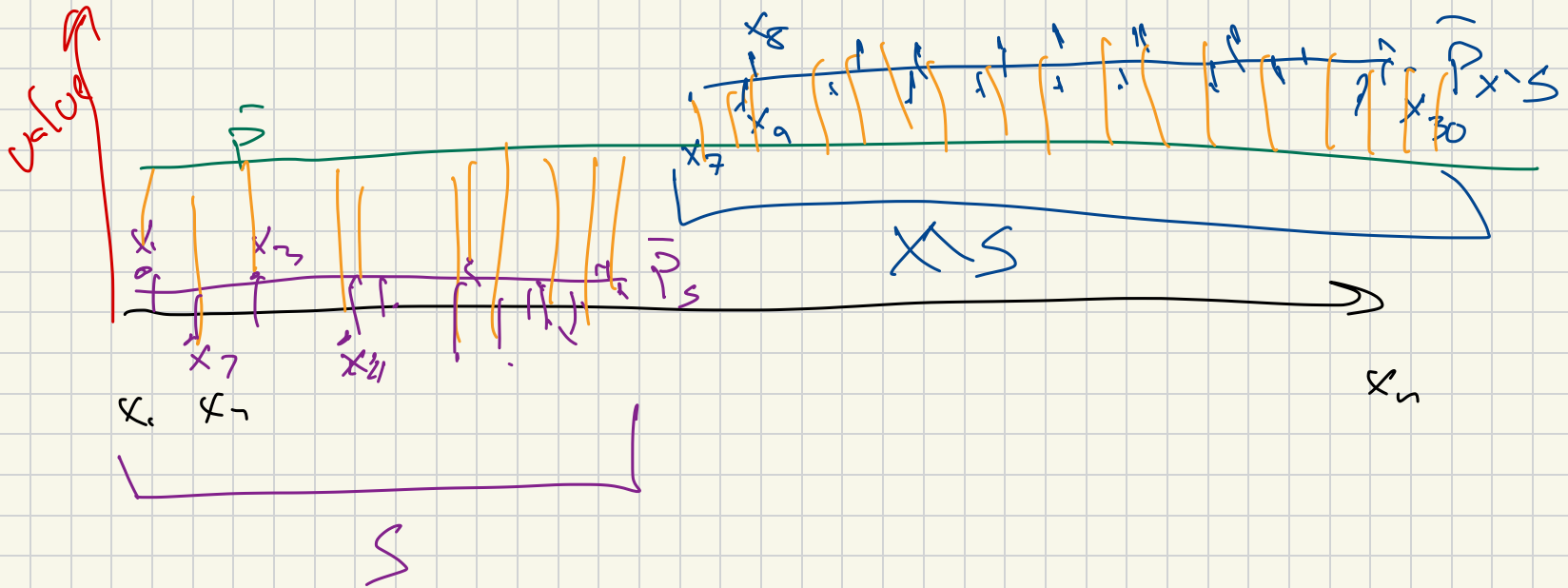
$L(x) = \max_{\mu \in \mathbb{R}} L(x; N(\mu, \sigma))$

$\mu^* = \bar{x} = \frac{1}{n} \sum_i x_i$

$LR = \frac{L(s) \cdot L(x \setminus s)}{L(x)} = \frac{L(s; \bar{x}_s) \cdot L(x \setminus s; \bar{x}_{x \setminus s})}{L(x; \bar{x})}$

$LLR = \ln\left(\frac{L(s) \cdot L(x \setminus s)}{L(x)}\right) = \ln(L(s)) + \ln(L(x \setminus s)) - \ln(L(x))$

$LLR(s, x)$



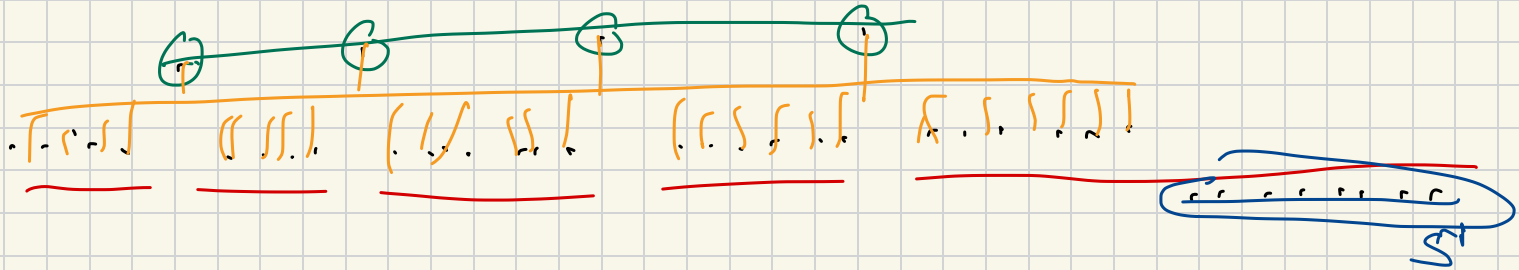
Most anomalous region S

$$S^* = \arg\max_S LLR(S, X)$$

Most anomalous region S

$$S^* = \arg\max_S LLR(S, X)$$

S - larger LLR score.



- more points
- closer together

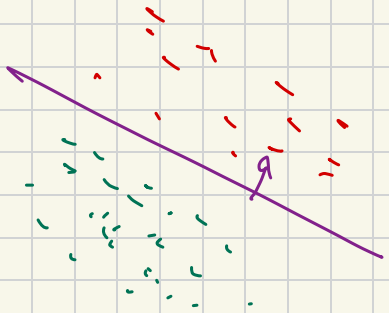
What shape of anomaly to consider?

S' - interval subset

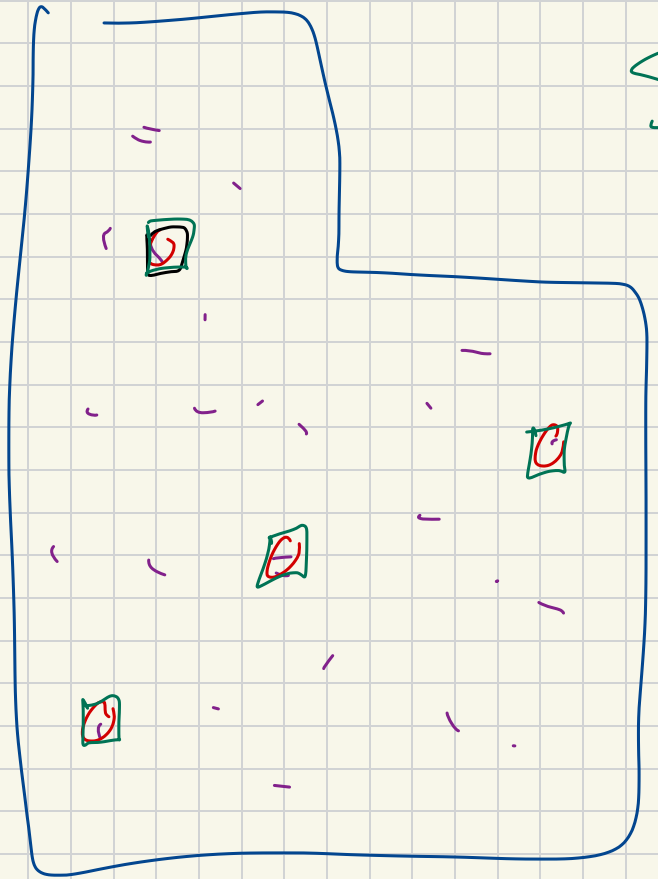
2. Finding Anomalies

Define class of possible anomaly shapes S

$$S^* = \arg\max_{S \in \mathcal{S}} LLR(S, X)$$



- intervals in time series
- days of the week
- spatial regions (circles, rectangles, halfspaces)



$S \subset X$

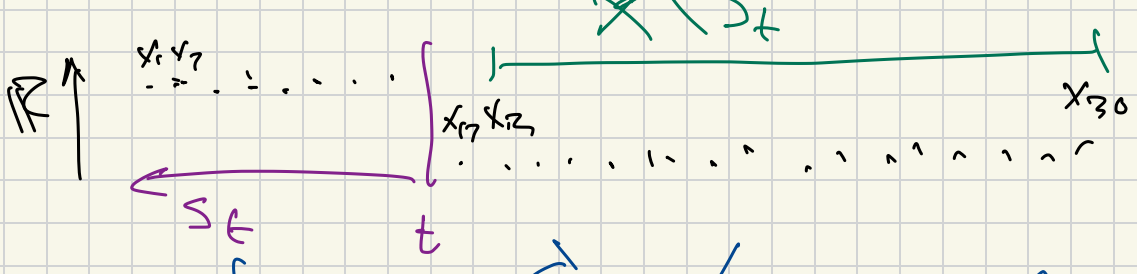
if no restriction to class S

↳ does not provide useful understanding.

Example $X = \{x_1, x_2 \dots x_n\} \subset \mathbb{R}$

time $\quad <$ time

$x_i \quad < \quad x_{i+1}$



Choose $S = \{ S_t \text{ s.t. } S_t = i \leq t \}$

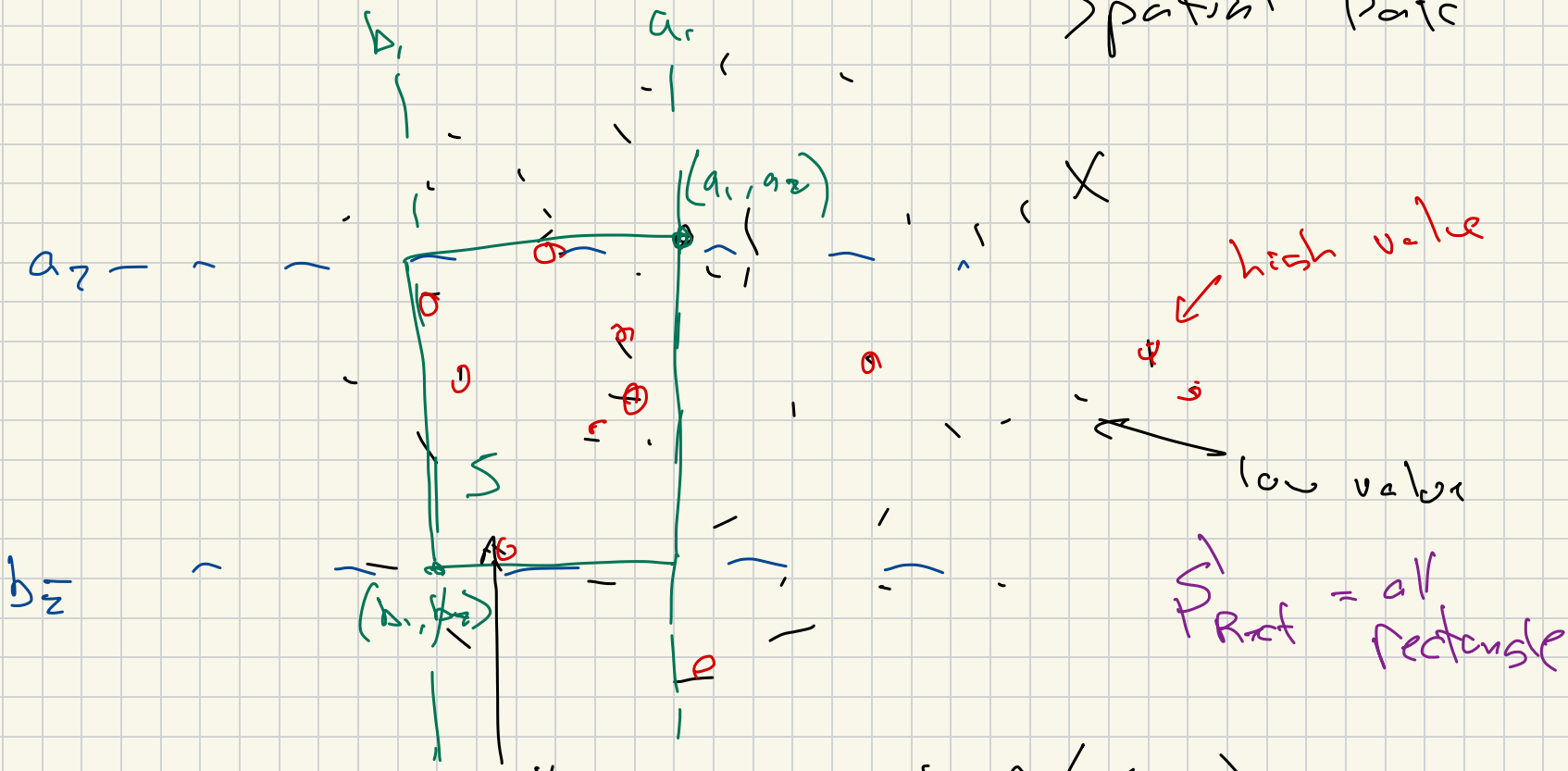
one-sided interval $-\infty \dots t$ $\quad X \setminus S$

Solve $S^* = \underset{S_t \in S}{\operatorname{argmax}} \, LLR(S_t, X)$ in $O(n)$ time.

- streaming algo
- mean variance

change point model

Spatial Data



S'_{Rect} = all rectangle

Find $S^* = \arg\max_{S \in S'_{Rect}} LLR(S, X)$

3. How to Quantify in Anomaly is interesting

$S^* = \left[\underset{S \in \mathcal{S}}{\operatorname{argmax}} \; LLR(S, X) \right]$ code — always exists!

Found something S^*, is it interesting?

option A: trust the LLR.

$LLR \geq 2.5$ or $LLR \leq 0.2$ → interesting.

→ make $\mathcal{S}$ larger → LLR → larger.

option B : Resample $X \sim u$

if I know u

generate m more datasets

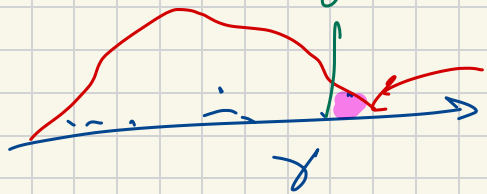
$X, X_1, \dots X_m \sim u$

total $m+n$ pts

$$S_j = \underset{S \in \mathcal{S}}{\operatorname{argmax}} \, LLR(S, X_j)$$

$$\gamma_j = LLR(S_j, X_j) \Rightarrow \{\gamma, \gamma_2, \dots \gamma_m\}$$

$$\gamma^* = LLR(S^*, X)$$



func heavier than γ^*

$\Rightarrow$ p-value