# L3 : Anomaly Detection

Jan 13, 2025

Data Mining : Jeff Phillips

# Review

Distribution $D(\Theta)$

$\text{unif}([m])$

- Events # trials until collision

  Median: $\approx \sqrt{2m}$

  Birthday Paradox

- Event: # trials until see all

  Expected # $= (0.577 + \ln(n)) \cdot n$

  $O(n \cdot \log n)$

  Coupon Collectors

# Anomalies

Data $X \underset{iid}{\sim} D(\theta)$ $X = \{x_1, x_2 \dots x_n\}$

$x_i \sim D(\theta)$

What is likely?

If we have $X$, how to generate similar $X$?

Anomalies

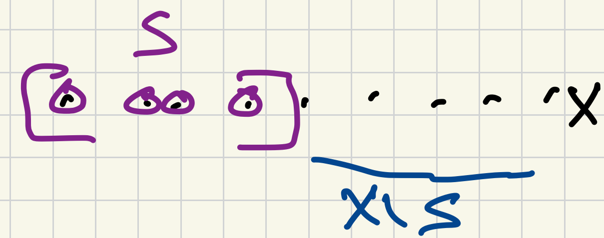
1. what is distribution of data? (model)
2. what would an anomaly look like.
   - score (likelihood, LLR)
   - shape.
3. How interesting? Quantify

**Likelihood** : unnormalized probability

$$L(X;\theta)$$

↑ parameters $\theta$ e.g. $\theta = p \in \mathbb{R}$

potential anomaly $S \subset X$

Data



likelihood of anomaly

$$L(S, X\setminus S; \theta') \qquad \theta' = p, q \in \mathbb{R}$$

score Log-likelihood ratio

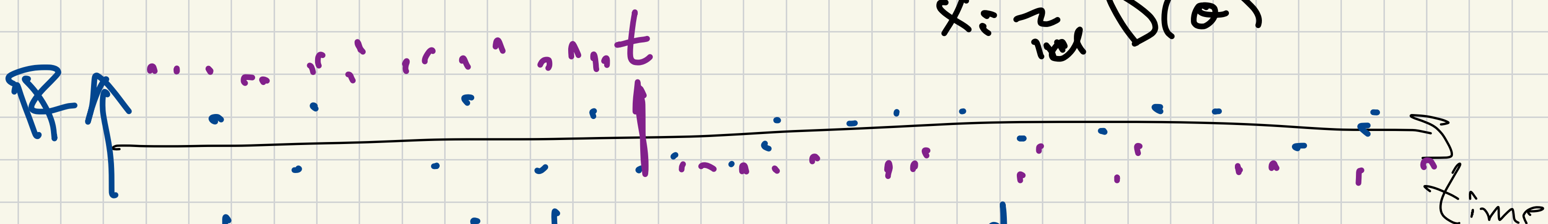
$$LLR(S,X) = \log\left(\frac{L(S, X\setminus S; \theta')}{L(X;\theta)}\right)$$

# Change Point Detection

Data X : sequence n real values

$\langle x_1, x_2, \ldots x_n \rangle$ $\quad x_i \in \mathbb{R}$

$x_i \sim_{iid} D(\theta)$




temp, stock, power used

$D(\theta)$ no change $x_i \sim_{iid} D(\theta)$

mean p Normal noise

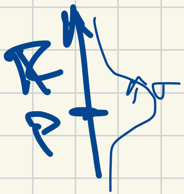
mean

$x_{t+1} \ldots x_n$

$D(\theta')$ : change at time t

$x_1 \ldots x_t \sim_{iid} D(p')$ , $x_{t+1} \ldots x_n \sim_{iid} D(\theta')$

Likelihood $L(x_j; \theta = p)$




$$\Pr[x_i; \theta = p] = \underbrace{\frac{1}{\sqrt{2\sigma\pi}}} \cdot \exp\left(-\frac{(x_i - p)^2}{\sigma^2}\right)$$

$$L(x_i; \theta = p) = \Pr[x_i; \theta = p]$$

proportional to $\rightarrow$ $$\propto \exp\left(\frac{-(x_i - p)^2}{\sigma^2}\right)$$

$$L(X; \theta = p) \propto \prod_{i=1}^{n} \exp\left(\frac{-(x_i - p)^2}{\sigma^2}\right)$$

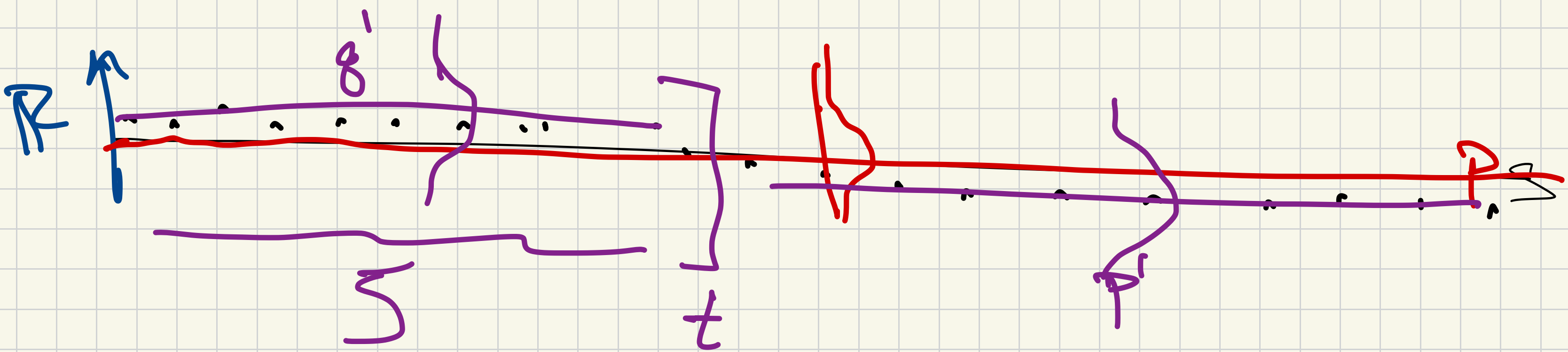
$$\ln(L(X; \theta = p)) = \ln\left(\prod_{i=1}^{n} \exp\left(\frac{-(x_i - p)^2}{\sigma^2}\right)\right)$$

$$= \sum_{i=1}^{n} \cancel{\ln(\exp}\left(\frac{-(x_i - p)^2}{\sigma^2}\right)) = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - p)^2$$

opt $p = \hat{p} = \frac{1}{n} \sum x_i$

$$\ln\left(L(x;p)\right) = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - p)^2$$

$$\ln\left(L(s, x\setminus s; p', g')\right) = -\frac{1}{\sigma^2}\left(\sum_{i=1}^{t}(x_i - g')^2 + \sum_{i=t+1}^{n}(x_i - p')^2\right)$$



score $$\gamma(S) = \max_{p',q'} \ln\left(L(S, X \setminus S; p', q')\right) - \max_{p} \ln\left(L(X; p)\right)$$

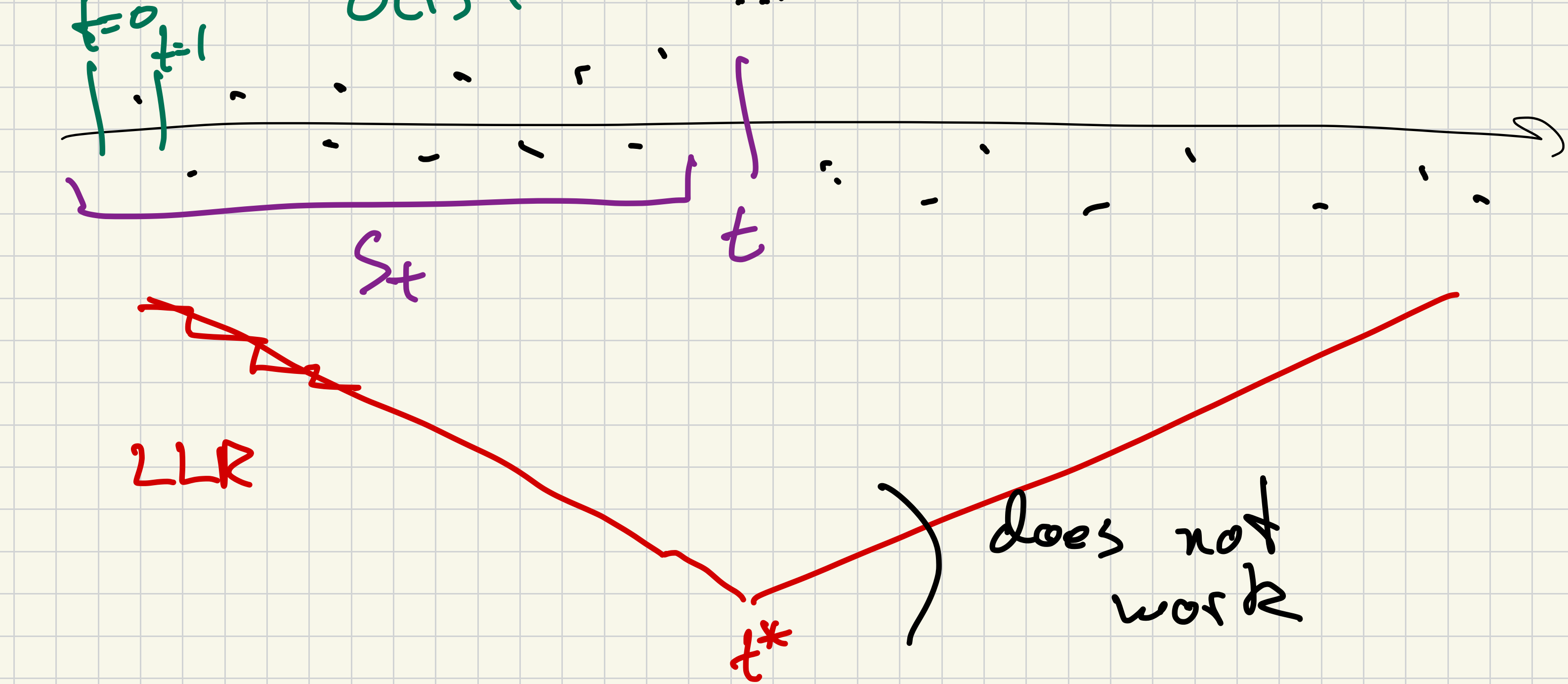
score $\gamma(S)$

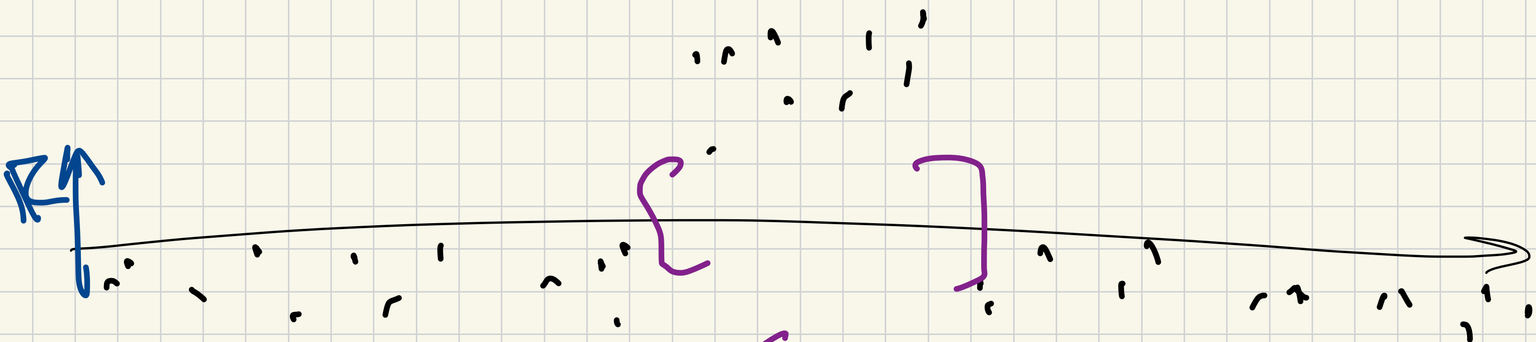
$$LLR(S, X)$$

Find $S^* = \underset{S_t}{\arg\max} \underbrace{LLR(S_t, X)}$

$LLR(0) \rightarrow LLR(t=1)$

$O(1)$ time ... $O(n) + n \cdot O(1) = O(n)$






$S$

$[s, t]$

now $n^2$ anomalies $S_{s,t}$

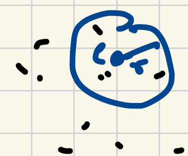
Two distribution $X \sim D(\theta)$ , $X' \sim D(\theta')$

baseline — anomalous

way to measure $(X, \mathcal{R})$ range space

$\mathcal{R} \ni (-\infty, t]$ change point

$(s, t]$ two-sided

 $\mathbb{R}^2$

balls

$$d_{\mathcal{R}}(D(\theta), D(\theta')) = \overset{\max}{\cancel{\sup}}_{R \in \mathcal{R}} \left| \mathop{E}_{x \sim D(\theta)} [R(x)] - \mathop{E}_{x' \sim D(\theta')} [R(x')] \right|$$

integral probability metric

So ... I found $S^* = \arg\max_S LLR(S;X)$

is it interesting?

1. if I totally trust $D(\theta)$, $D(\theta')$

then LLR $\approx$ t-score related to p-values.

$\approx$ if LLR $\approx 2 \rightarrow p \approx 0.05$

What if LLR is just a "score"?

2. Draw more data from $D(\theta)$

$$X_1, X_2, \ldots X_n \sim D(\theta)$$

compute score $\gamma(X_1) = \gamma_1$

$$\gamma_1, \gamma_2, \ldots \gamma_n$$

compare $\gamma(X)$ to $\gamma_1 \ldots \gamma_n$

input

what fraction of $\gamma_1 \ldots \gamma_n > \gamma(X)$

if $p = \frac{\# >}{n} < 0.05$ interesting

3. What if I do not know $D(\theta)$?

Permute existing data

permutation test.

$x_1, x_2 \ldots x_n \in X$ randomly permute order

new $\langle x_7, x_{17}, x_3, \ldots x_8 \rangle = X_1$

$(n \times n)$

$X_2$

$\vdots$

$X_n$

Calc $n$ values $\gamma_i = \gamma(X_i)$

compare $\gamma(X)$ to $\gamma_1, \gamma_2 \ldots \gamma_n$