Data Stream Mining- Lecture 2

Basics of stream mining

Chandresh Kumar Maurya, Research Assistant Professor

Eötvös Loránd University, Budapest, Hungary

September 19, 2019

Data Synopsis

Need to compute an estimate of the stream due to 1) low memory, 2) fast computation. Data synopsis can be done in two ways:

- Sliding Window
- Data Reduction

Sliding window

Why we need sliding window?

Sliding window

Why we need sliding window? For capturing recent data

Types of Sliding window

- **Sequence based**: they contain sequences of data and size of the window is decided based on the number of data sequences they contain.
- **Timestamp based**: The size of the window is decided based on the time interval considered.

Sequence based window: Examples

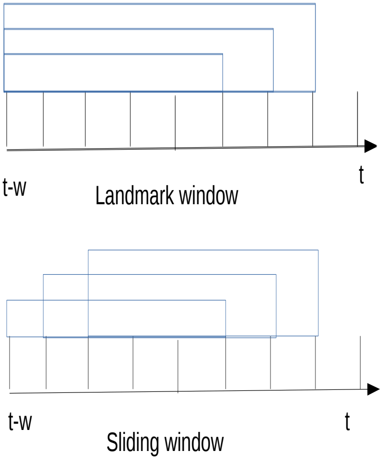

Figure: Sequence based windows. Top figure: Landmark window and bottom figure is sliding window (used in packet transmission)

Timestamp based window: Examples

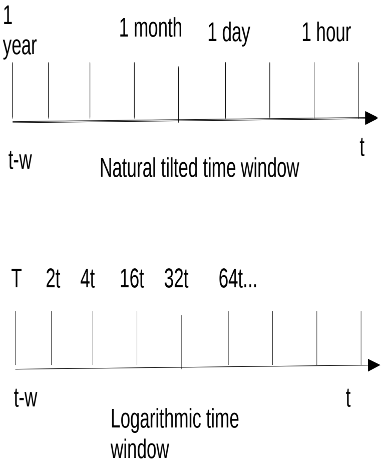

Figure: Timestamp based windows. Top figure: natural tilted window and bottom figure is logarithmic time window

Computing Statistics over Sliding Window: The ADWIN algorithm

Why we need to estimate statistics over window? Because can not store all items in the window or want to perform some operation. Solution: Adaptive Sliding Window Algorithm (ADWIN)[Bifet and Gavalda, 2007] .

ADWIN

A change detector and estimator algorithm using an adaptive size sliding window

Computing Statistics over Sliding Window: The ADWIN algorithm

```
Algorithm 1: ADWIN
Input         : Sequence $\{x_t\}$ and confidence value $\delta$
Initialization: Window $W$
for $t > 0$ do
    $W \longrightarrow W \cup x_t$ (add items to the head of $W$)
    do
        Drop elements from the tail of $W$
    while $|\hat{\mu}_{W_0} - \hat{\mu}_{W_1}| < \epsilon_{cut}$ holds for all split of $W$ into $W_0$
     and $W_1$;
end
Output: $\hat{\mu}_W$
```

Where ϵ_{cut} is given by:

$\epsilon_{cut} = \sqrt{\frac{1}{2m}.\ln\frac{4|W|}{\delta}}$ and m is the harmonic mean of W_0 and W_1.

Bibliography I

Bifet, A. and Gavalda, R. (2007).
Learning from time-changing data with adaptive windowing.
In *Proceedings of the 2007 SIAM international conference on data mining*, pages 443–448. SIAM.