Topics:

interactivity

qo's

1. Market Basket Analysis
2. Association Rule Mining

RecSys

3. Apriori Algo

4. Introduction to Time-series Analysis
   - ✓ ↳ Forecasting — missing values ✓
   - ✓ → Resampling — outliers / Anamolies

Market Basket (90's)

DMart / walmart

items → Milk, bread, jam, butter - - - → few 100's of products

$\mathcal{D} = \{1, 2, 3, \ldots n\}$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:10

Transactions:

basket

T1: {1, 3, 6, 8}

T2: {1, 3, 7, 12}

T3: {1, 7, 3, 16}

T

patterns:

itemset

{1, 3}

bread

Milk

{1} → recommend {3}

RecSys

~100

Task: (n) $\mathcal{D}$: set of all items

{T: Set of transaction

(m) $T_i \subseteq \mathcal{D}$

m >> n

V. large

Find $\subseteq \mathcal{D}$ item-sets } that occur very frequently in T

Find frequent itemsets

Q

1

for each Ti

for each Tj j≠i

Ti ∩ Tj

[incomplete]

2

X K

value

cnt

itemset

{i,j} ∀ i,j

{i,j,k}

⋮

(2) all subsets of $x = \{1, 2, \dots n\}$

x: $2 \; 2 \; 2 \cdots 2$

2^n

$\{ \dots \}$

= Too slow

$\{1,2,4\} \rightarrow T_1, T_2 \dots T_M$

m = # transactions

Time complx: $O(2^n \times m)$

anecdote: 5PM - 7PM 90's

{Beer, diapers}

Key idea:

$\#\{2,4\} < 100$ (let) ← Threshold

↳ minimum support (c)

$\#\{2,4,6\} < 100$

↳ superset of $\{2,4\}$

C = 100 min support

50 <

{1}^100 {2}^200 - {3}^150 {4}^50

X

✓105 ✓110 60

① → {1,2} {1,3} {2,3}

X

② → {X}

⊄ {2,3}

⊄ {4}

$O(2^n \times m)$ — $m \to 10^6$

reduce this significantly

$n = 100$

2^{10}

$2^{10} \approx 10^3$

$2^{20} \approx 10^6$

$2^{30} \approx 10^9$

$2^{40} \approx 10^{12}$

not used in ecommerce

Apriori algorithm (1994-95)

n increases (~1000)

Very Costly

Modifications → FP-growth

Specialized Ds: Tries

- Frequent itemset mining
  - Used $n = |\Sigma|$ is small

Q other applications ? (beyond retail)

→ bio-informatics

{ $\widetilde{c_1}$, $\widetilde{c_3}$ } { $\widetilde{AGTC}$, $\widetilde{ATTC}$ }

→ medicine

{ M_1, M_2, M_3 }

→ web usage mining } ✓

{w1, w2, w3}

→ Similar words

{w1, w2, w3}

You need
2
T

Association Rule Mining

set of items

$D = \{1, 2, 3, \ldots n\}$

Let

$X = \{1, 2, 3\}$

$Y = \{4, 6\}$

$\{1, 2, 3, 4, 6\}$: frequent itemset

$X \Rightarrow Y$

↓

Apriori algo

e.g: X {beer} ⇒ Y {diapers}

{milk, bread} ⇒ {Jam, eggs}

$$\text{Support}(X) = \frac{\#\text{ Transactions that contain } X}{\text{Total transactions}}$$

$$= \frac{100}{10{,}000} = P(\underline{X} \text{ in } T)$$

Confidence $(X \Rightarrow Y) = \dfrac{\#(X \text{ and } Y)}{\#X}$

$= P(Y \mid X)$ (logically)

90% +

freq.
X ∪ Y is a itemset

$$\text{Lift}(\tilde{X} \Rightarrow \tilde{Y}) = \frac{\text{Support}(X \cap Y)}{\text{Supp}(X) \times \text{Supp}(Y)}$$

$$= \frac{P(X \text{ and } Y)}{P(X)\ P(Y)}$$

intuitively

{ X & Y are indep

$P(X \text{ and } Y) = P(X)\,P(Y)$ if X & Y are indep

$\text{Lift}(X \Rightarrow Y) = 1$ if X & Y are indep

> 1

Market Basket → freq. itemsets

Apriori / FP-growth

Association Rule Mining

freq. itemsets

{milk, bread} ⟶ freq itemset

X Y

↓

(Apriori)

[Supp(x)
Supp(Y)
Conf(X⇒Y) ✓

Lift(X⇒Y) ✓]

Q

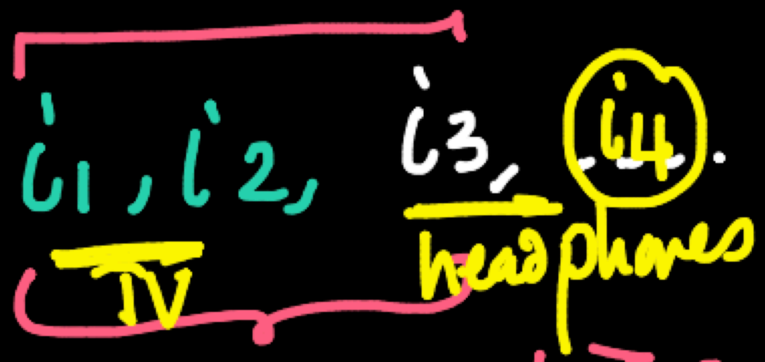
View → purchased [MF]

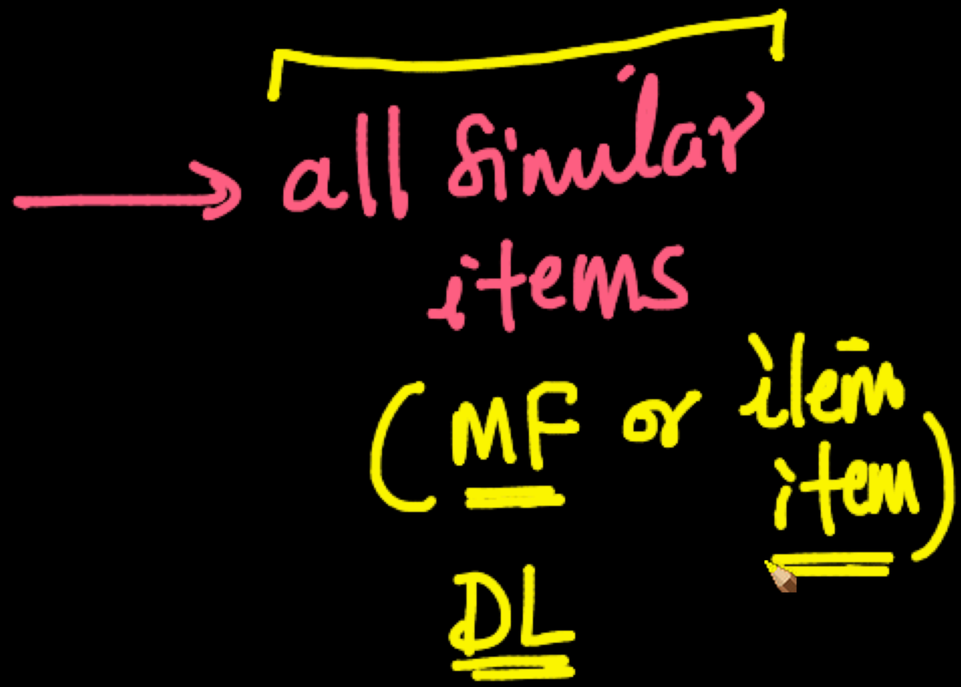
e-commerce!

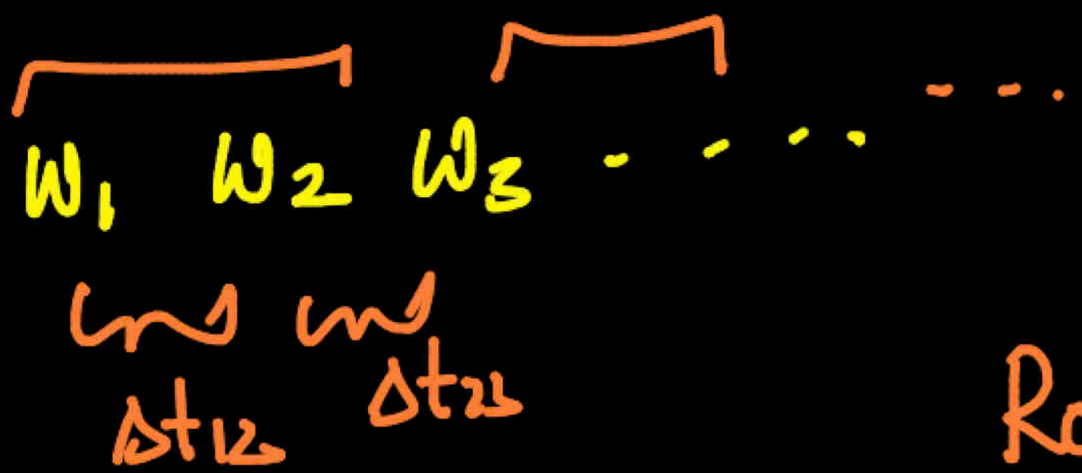
i_1 ✓, i_2 ✓, i_3 ✓ .... i_{10} ✓ ← Samsung TV

V V V B

$P(i_{10} \mid i_1, i_2 \ldots)$ ← Naive Bayes

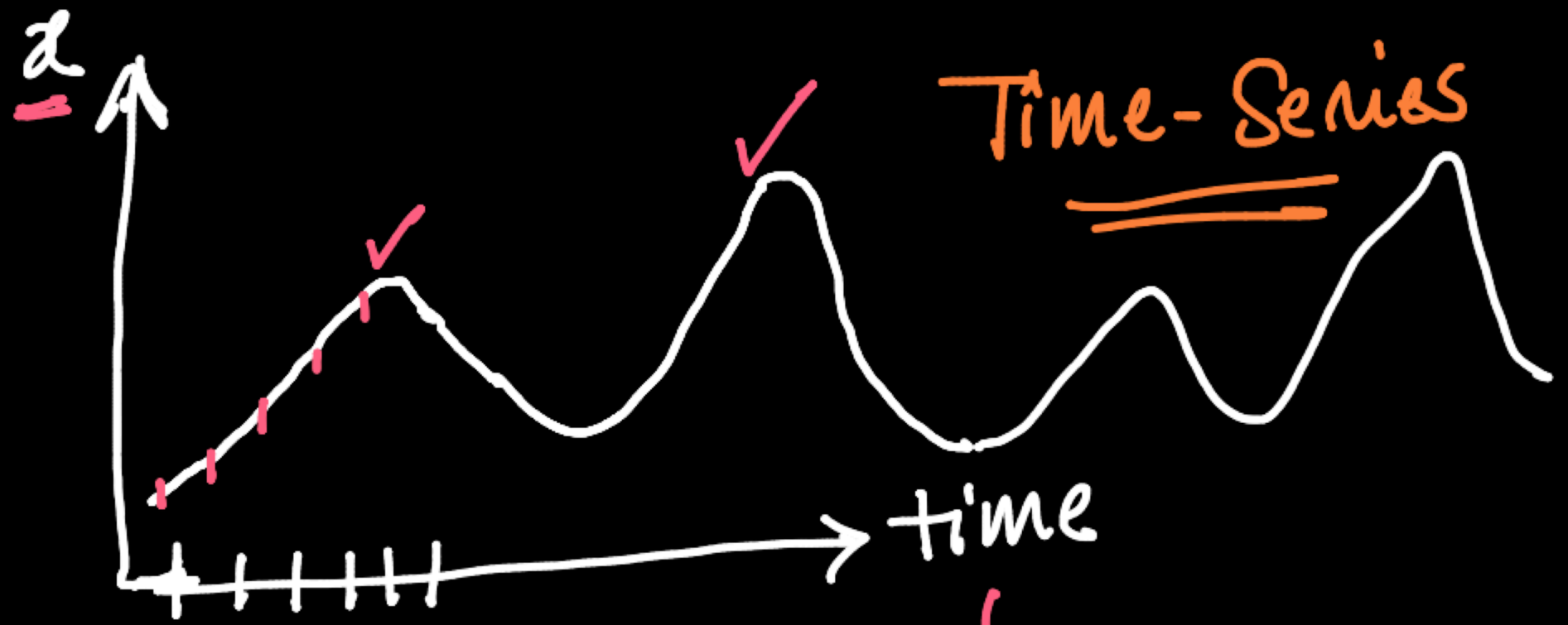






Time-Series

#items sold on Amazon

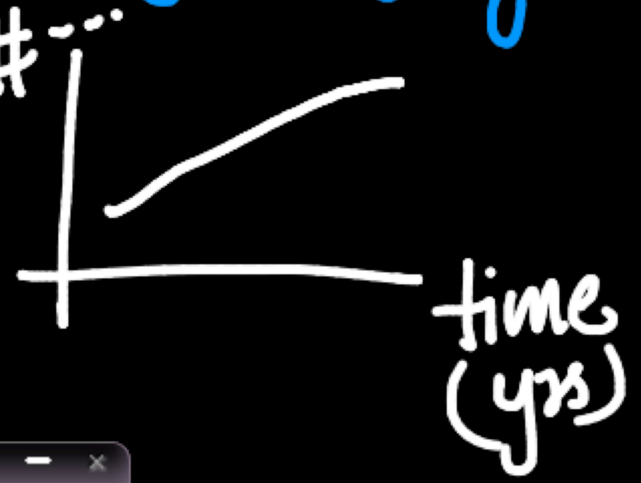


min / hrs / weeks / months / years

$x(t)$

items :

- hourly pattern ← morning, night-
- weekly pattern ← weekend & weekday
- holiday season ...
- weather-pattern .. (category)
- Yearly pattern



Forecasting

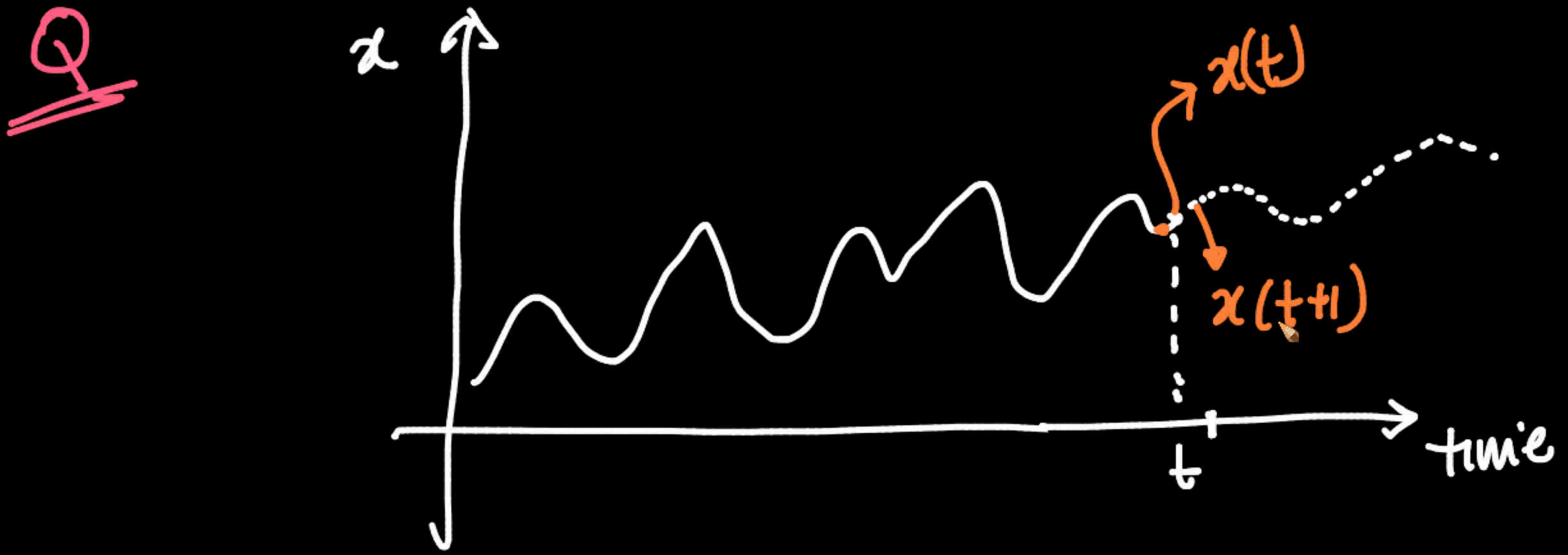
Task:

Given

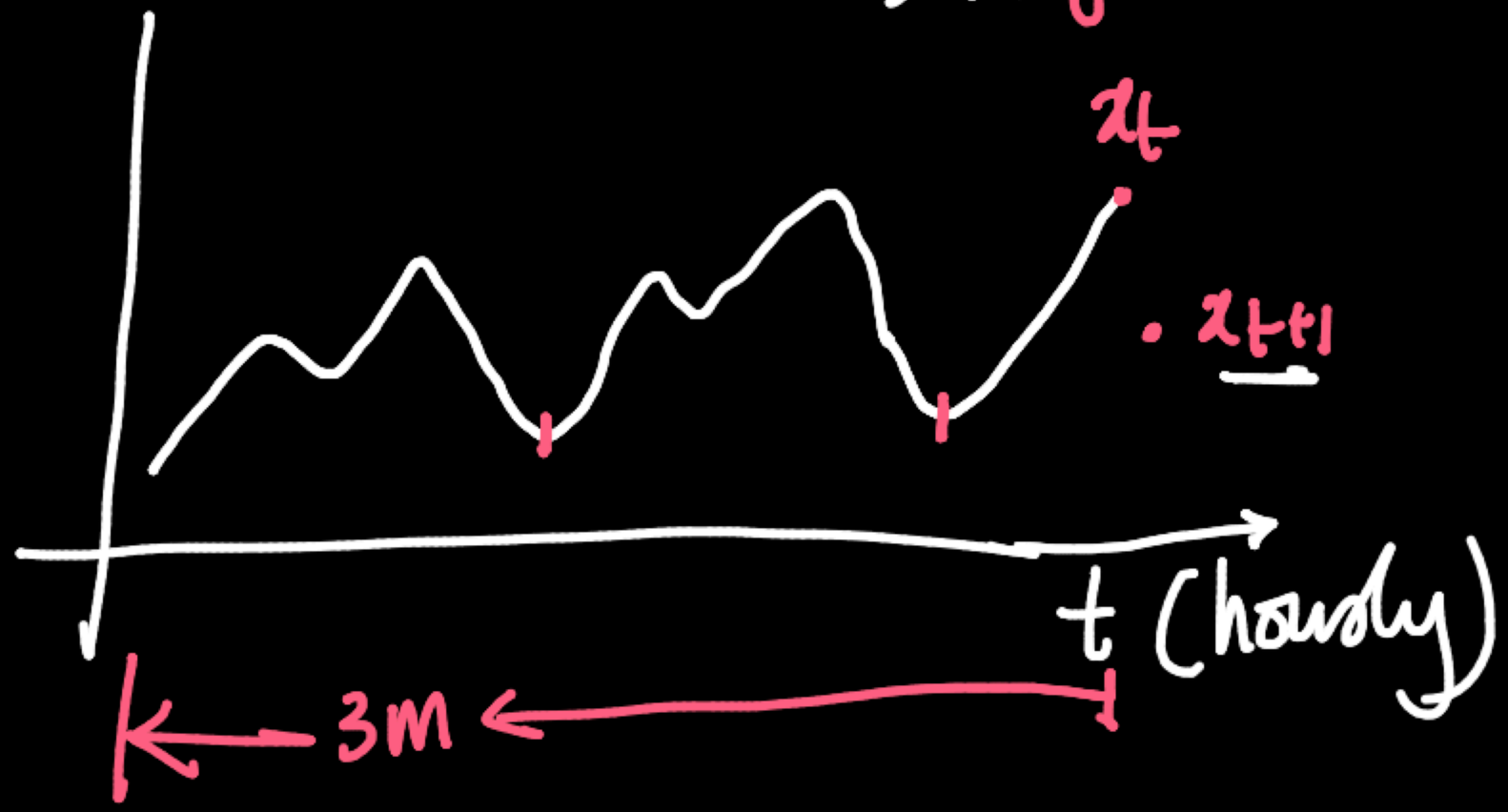
x_1 x_2 ---- x_{t-1} x_t ← historical data

predict

x_{t+1} x_{t+2} ...

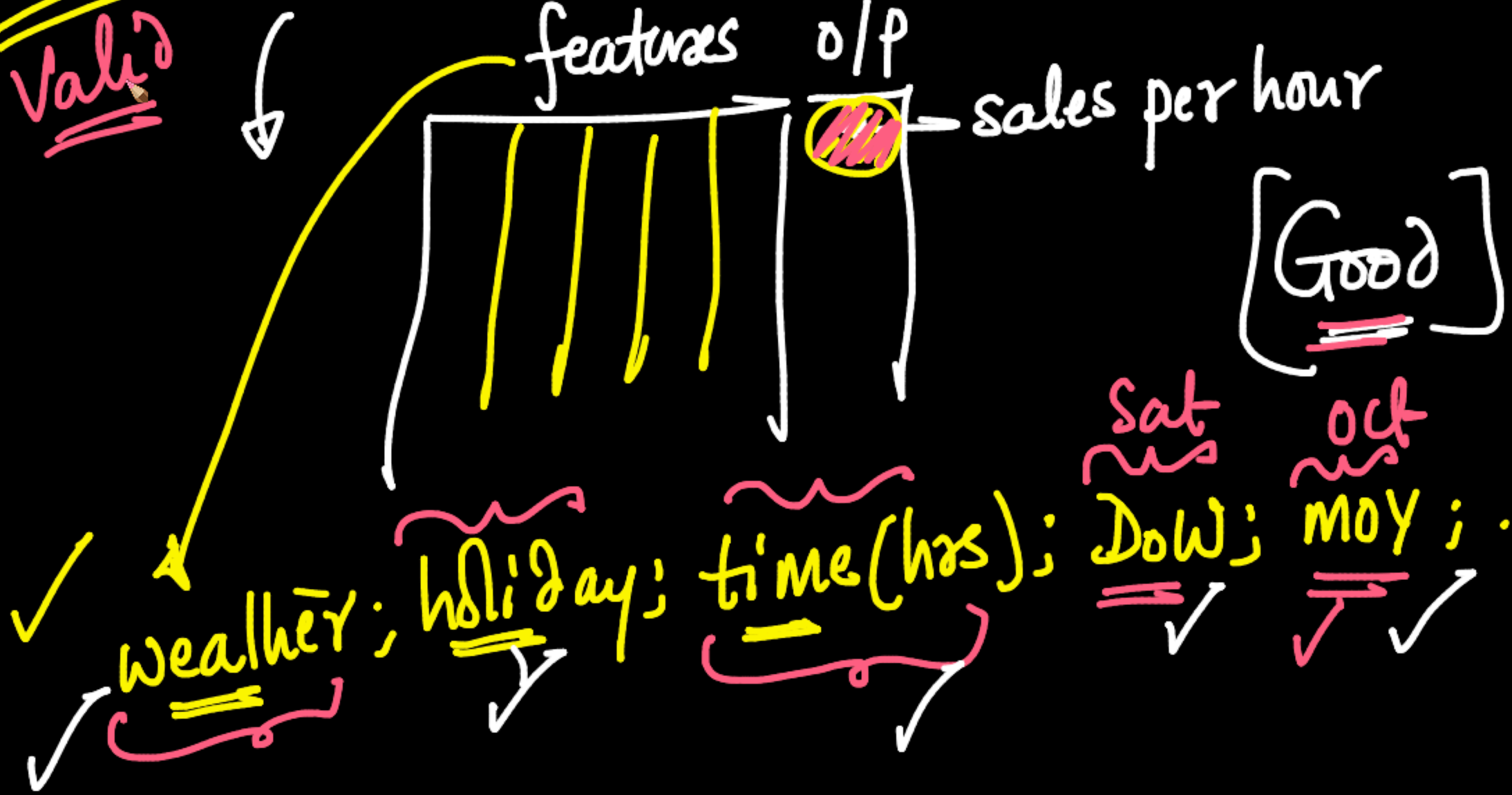


① average of last ~~3 months~~ ~~daily~~ may not work



(2) regression / curve fitting

Valid?



flaw: may not capture trend over years

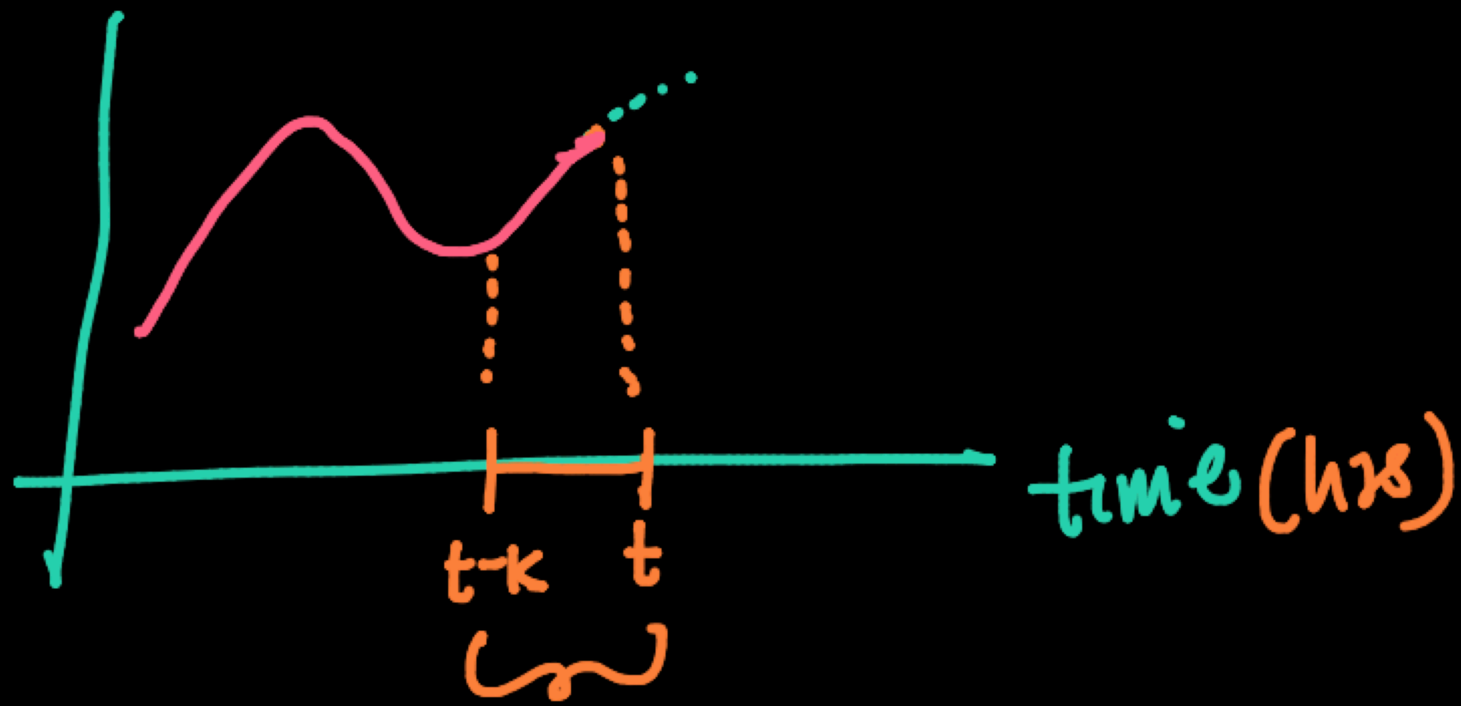
soln: feature: ✓

2012 → 0
2013 → 1
2014 → 2
⋮

↳ Year of operation

2012

(3) Weighted moving average → classical TS



window-size = K

$$\hat{x}_{t+1} = \frac{\alpha_0 \vec{x}_t + \alpha_1 \vec{x}_{t-1} + \alpha_2 \vec{x}_{t-2} \cdots + \alpha_k \vec{x}_{t-k}}{\alpha_0 + \alpha_1 + \alpha_2 \cdots + \alpha_k}$$

Weighted avg

$$= \frac{\sum_{i=0}^{k} \alpha_i x_{t-i}}{\sum_{i=0}^{k} \alpha_i}$$

logically $\alpha_0 > \alpha_1 > \alpha_2 \cdots > \alpha_k$

$x_{t-k} \cdots x_t \longrightarrow \hat{x}_{t+1}$ ✓

$x_{t-k+1} \quad \hat{x}_{t+1} \longrightarrow \hat{x}_{t+2}$

WMA: Type A. regression with manual weights

$$x_{t+1} = \frac{\sum_{i=0}^{r} \alpha_i \, x_{t-i}}{\sum_{i=0}^{r} \alpha_i}$$

WMA with computed weights

↓

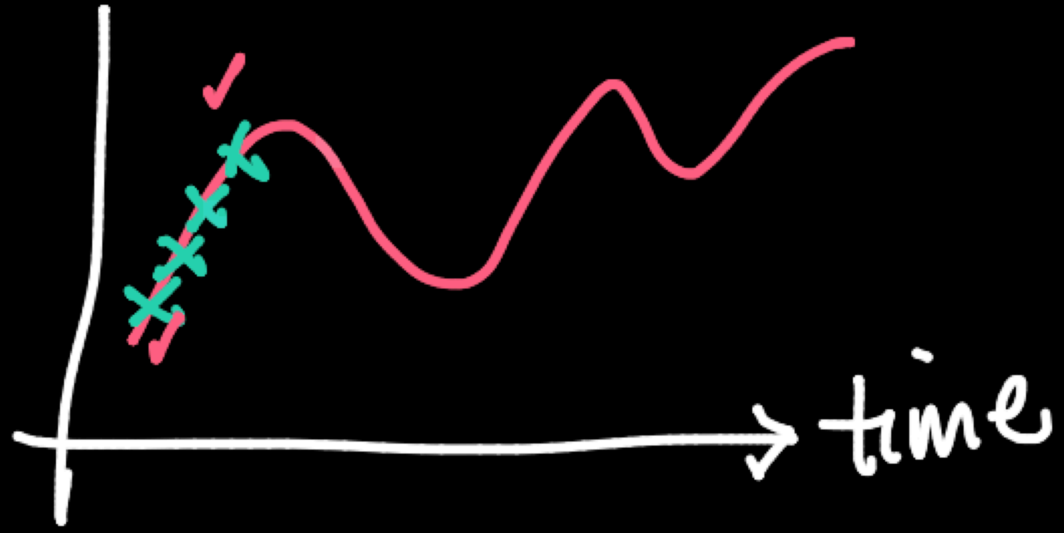
d_i: weights

(Linear-reg)

Split your data

{ Train - Val - Test

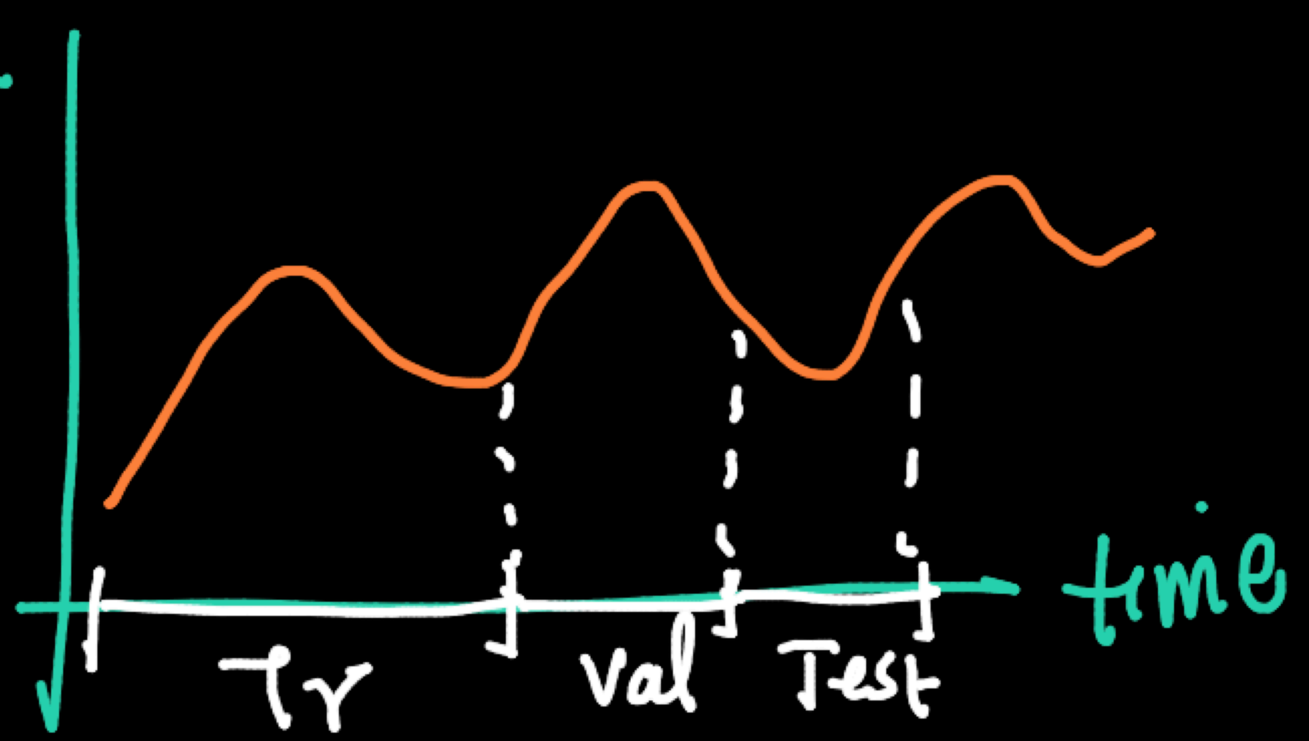
random - splitting ✗



Time-based Splitting

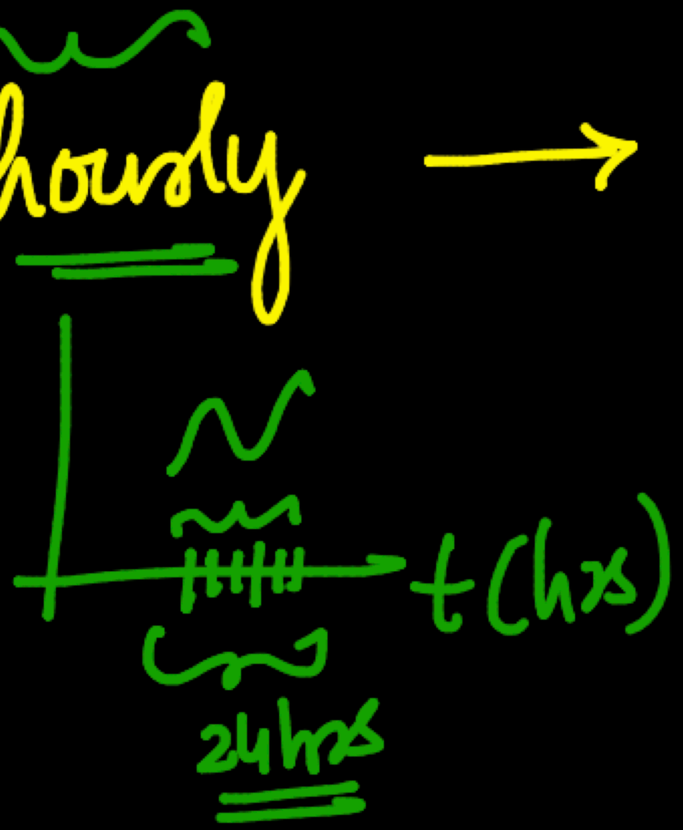
WMA: hyperparam

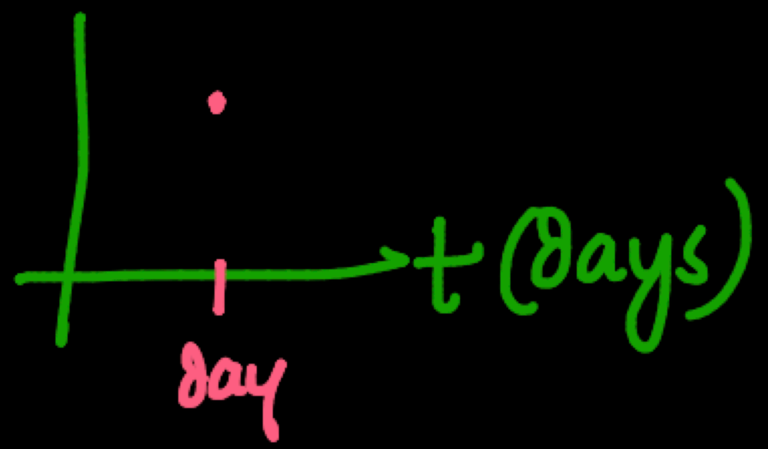
K



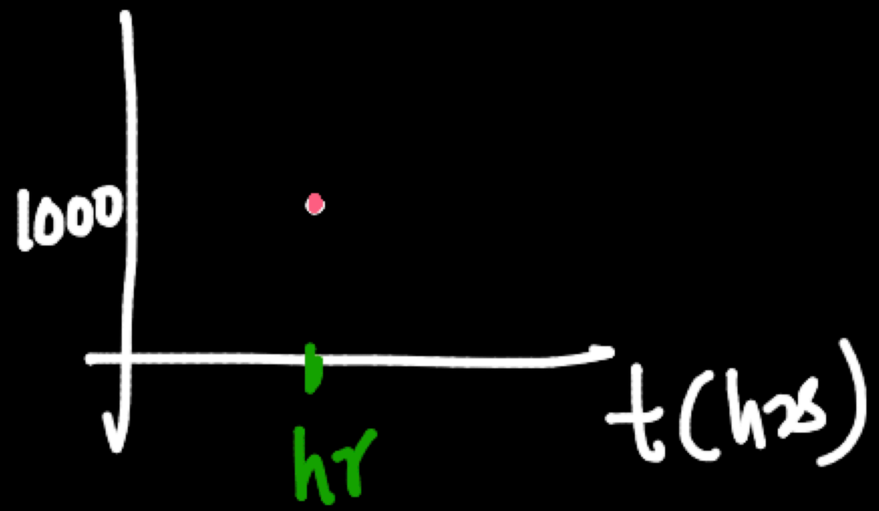
Resampling

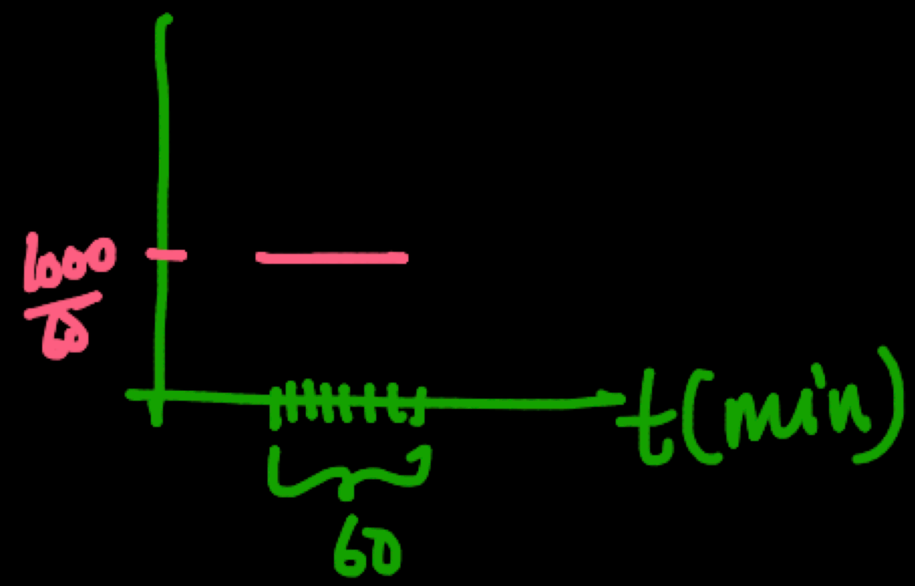
① hourly → daily





② hourly → minutely

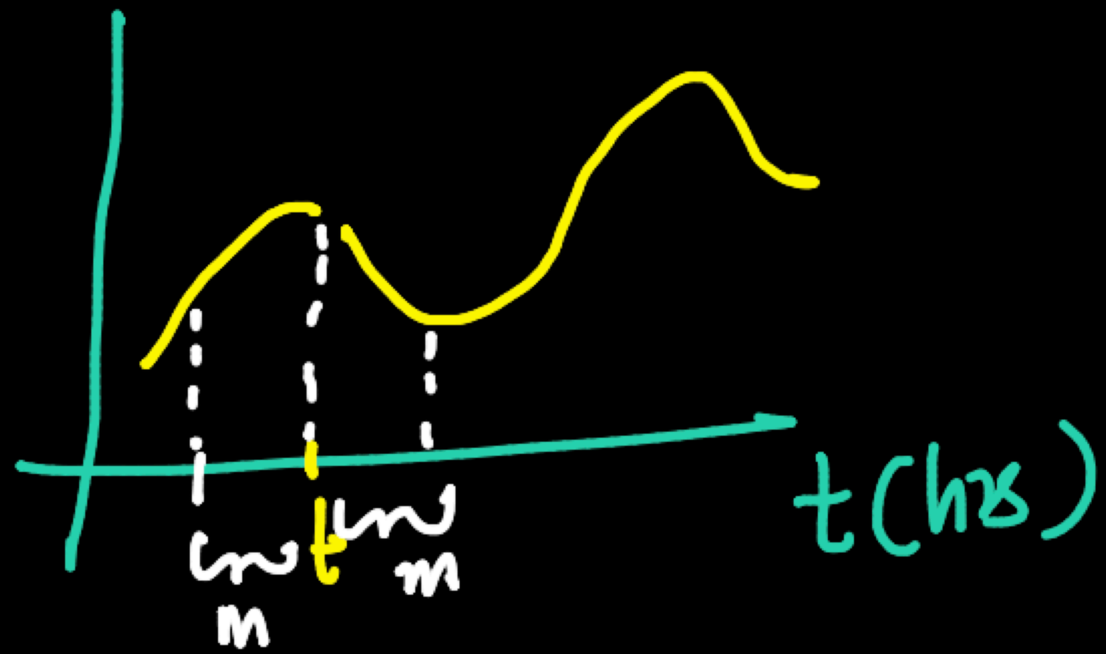




assume uniform spread across each min

Missing - data Imputē

→ Cannot drop



→ $x_t = x_{t-1}$ ① ②

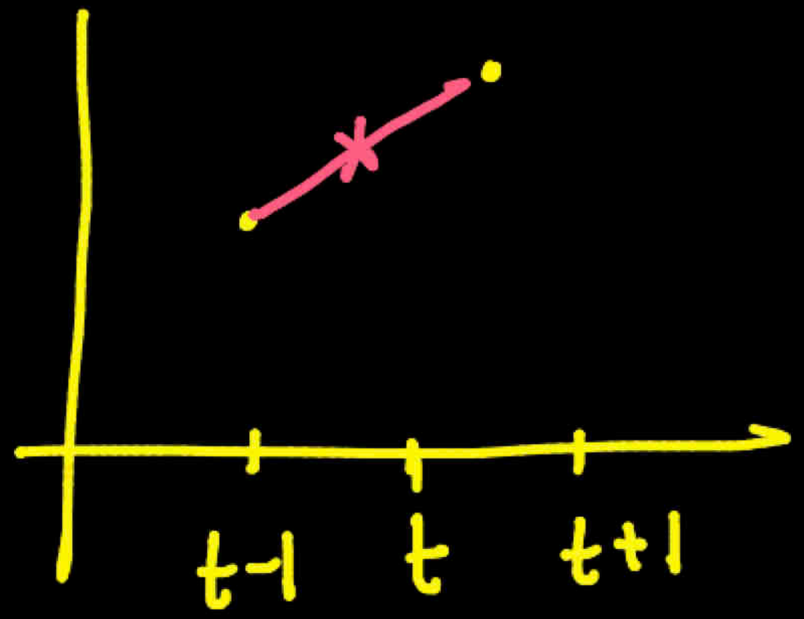
✓ $x_t = \frac{x_{t-1} + x_{t+1}}{2}$

✓ ③ Centered MA

$$x_t = \frac{1}{2m+1} \sum_{i=t-m}^{t+m} x_i \cdot 1$$

generalizzabile
a looking back
& forward

interpolation : $x_t = \frac{x_{t-1} + x_{t+1}}{2}$ ≈ Centered MA of width = 1



Outliers / anamolies

#ulenas

t(hrs)

(1)

$$\frac{x_t - x_{t-1}}{x_{t-1}} = \Delta_t$$

Δ_t ← relative change

$\Delta_1 \; \Delta_2 \cdots \Delta_t \longrightarrow$ IQR ↙

(2) iForest / one class SVM / DBScan

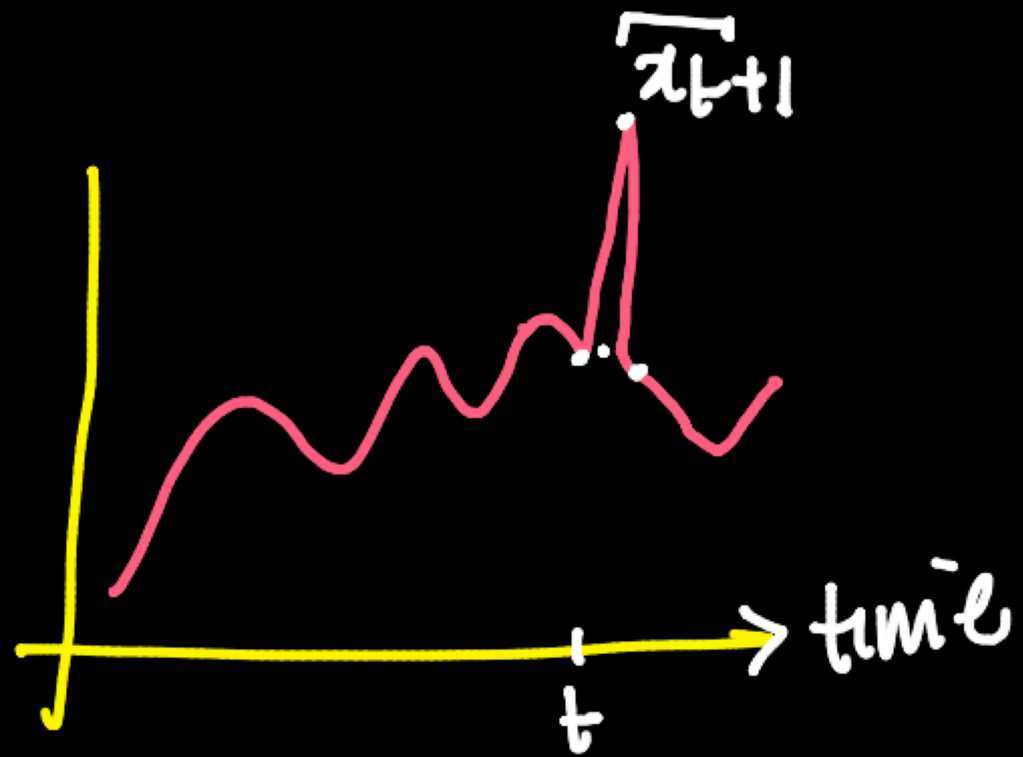
↳ Tabular data

↑

Time Series

sequence / series

(3)



assumption
- historical data has no outliers
- pattern..

$\hat{x}_{t+1}$: regression / WMA ...
DL model

✓ $\frac{x_{t+1} - \hat{x}_{t+1}}{\hat{x}_{t+1}} = \delta_t$ is high/low

↳ 200%

Next - class

- code
- Trend, Seasonality
- ARIMA, SARIMA ...

Classical ML → time series

Q

Classical-ML < DL

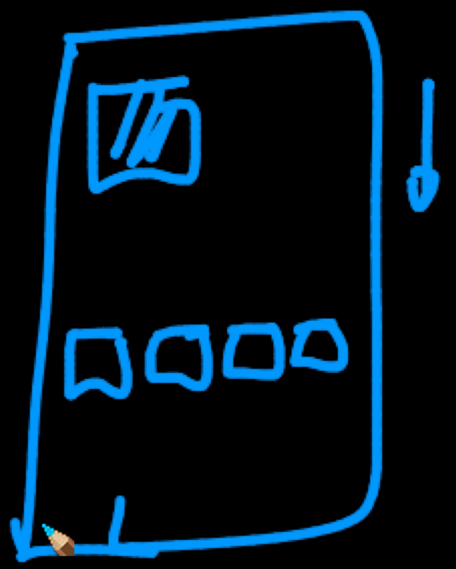
Classical-TS-models < DL (LSTM/GRUs)
(ARIMA)

Q

RecSys → RMSE; log-loss; ...

↳ click-rate

→ click-purchase-rate



Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:17

Market basket - Wikipedia | Association rule learning - Wiki | Apriori algorithm - Wikipedia | New Tab

Search Google or type a URL

Gmail Images

Google

Search Google or type a URL

Colaboratory YouTube My Drive InterviewBit S... Learning

GitHub Reduce the fil... Scaler Acade... Amazon.in Add shortcut

Photo by NASA Image Libra

54 / 54