Rolling window features: part 1

Window features

Rolling window features

Date	Sales
2020-02-12	35
2020-02-13	30
2020-02-14	23
2020-02-15	30
2020-02-16	34
2020-02-17	?

Rolling window features

Date	Sales
2020-02-12	**35**
2020-02-13	**30**
2020-02-14	**23**
2020-02-15	30
2020-02-16	34
2020-02-17	?

- Apply a window to the time series.

Rolling window features

Date	Sales	Sales rolling mean	Sales rolling min
2020-02-12	**35**		
2020-02-13	**30**		
2020-02-14	**23**		
2020-02-15	30	**29.3**	**23**
2020-02-16	34		
2020-02-17	?		

- Apply a window to the time series.
- Compute statistics from data inside the window (e.g., mean, min).

Rolling window features

Date	Sales	Sales rolling mean	Sales rolling min
2020-02-12	**35**		
2020-02-13	**30**		
2020-02-14	**23**		
2020-02-15	30	**29.3**	**23**
2020-02-16	34		
2020-02-17	?		

- Apply a window to the time series.
- Compute statistics from data inside the window (e.g., mean, min).
- To avoid data leakage, assign the statistics to timestamp after window.

Rolling window features

Date	Sales	Sales rolling mean	Sales rolling min	
2020-02-12	**35**			
2020-02-13	**30**			
2020-02-14	**23**	**29.3**	**23**	✖
2020-02-15	30			
2020-02-16	34			
2020-02-17	?			

- Apply a window to the time series.
- Compute statistics from data inside the window (e.g., mean, min).
- To avoid data leakage, assign the statistics to timestamp after window.

Rolling window features

Date	Sales	Sales rolling mean	Sales rolling min	
2020-02-12	**35**			
2020-02-13	**30**			
2020-02-14	**23**			
2020-02-15	30	**29.3**	**23**	✓
2020-02-16	34			
2020-02-17	?			

- Apply a window to the time series.
- Compute statistics from data inside the window (e.g., mean, min).
- To avoid data leakage, assign the statistics to timestamp after window.

Rolling window features

Date	Sales	Sales rolling mean	Sales rolling min
2020-02-12	35		
2020-02-13	**30**		
2020-02-14	**23**		
2020-02-15	**30**	29.3	23
2020-02-16	34	**27.7**	**30**
2020-02-17	?		

- Apply a window to the time series
- Compute statistics from data inside the window (e.g., mean, min)
- To avoid data leakage, assign the statistics to timestamp after window
- Move window and iterate (i.e., roll) across the time series

Rolling window features

Date	Sales	Sales rolling mean	Sales rolling min
2020-02-12	35		
2020-02-13	30		
2020-02-14	**23**		
2020-02-15	**30**	29.3	23
2020-02-16	**34**	27.7	30
2020-02-17	?	**29**	**30**

- Apply a window to the time series
- Compute statistics from data inside the window (e.g., mean, min)
- To avoid data leakage, assign the statistics to timestamp after window
- Move window and iterate (i.e., roll) across the time series

What about edge cases?

Date	Sales	Sales rolling mean	Sales rolling min
2020-02-12	35		
2020-02-13	30		
2020-02-14	23		
2020-02-15	30	29.3	23
2020-02-16	34	27.7	30
2020-02-17	?	29	30

- Option 1: Treat as missing data.

What about edge cases?

Date	Sales	Sales rolling mean	Sales rolling min
2020-02-12	35	NaN	NaN
2020-02-13	30	NaN	NaN
2020-02-14	23	NaN	NaN
2020-02-15	30	29.3	23
2020-02-16	34	27.7	30
2020-02-17	?	29	30

- Option 1: Treat as missing data.

- Drop the rows with missing data.
- Impute the missing data.

- Pros:
  - All rolling statistics have the same window size.
  - Simple.

- Cons:
  - Reduces the amount of data if dropping rows.
  - Edge cases could be quite different to rest of data.

What about edge cases?

Date	Sales	Sales rolling mean	Sales rolling min
2020-02-12	35		
2020-02-13	30		
2020-02-14	23		
2020-02-15	30	29.3	23
2020-02-16	34	27.7	30
2020-02-17	?	29	30

- Option 2: Use smaller window sizes at the edges

What about edge cases?

Date	Sales	Sales rolling mean	Sales rolling min
2020-02-12	**35**		
2020-02-13	30	**35**	**35**
2020-02-14	23		
2020-02-15	30	29.3	23
2020-02-16	34	27.7	30
2020-02-17	?	29	30

- Option 2: Use smaller window sizes at the edges

What about edge cases?

Date	Sales	Sales rolling mean	Sales rolling min
2020-02-12	**35**		
2020-02-13	**30**	35	35
2020-02-14	23	**32.5**	**30**
2020-02-15	30	29.3	23
2020-02-16	34	27.7	30
2020-02-17	?	29	30

- Option 2: Use smaller window sizes at the edges

What about edge cases?

Date	Sales	Sales rolling mean	Sales rolling min
2020-02-12	**35**		
2020-02-13	**30**	35	35
2020-02-14	**23**	32.5	30
2020-02-15	30	**29.3**	**23**
2020-02-16	34	27.7	30
2020-02-17	?	29	30

- Option 2: Use smaller window sizes at the edges

What about edge cases?

Date	Sales	Sales rolling mean	Sales rolling min
2020-02-12	35	NaN	NaN
2020-02-13	30	35	35
2020-02-14	23	32.5	30
2020-02-15	30	29.3	23
2020-02-16	34	27.7	30
2020-02-17	?	29	30

- Option 2: Use smaller window sizes at the edges

- Drop the rows with missing data
- Impute the missing data

- Pros:
  - Less missing data
  - Simple

- Cons:
  - Statistics at edges are based on smaller window sizes