CSE 590
Data Science Fundamentals

Time Series Data

Klaus Mueller

Computer Science Department
Stony Brook University and SUNY Korea
<table>
<thead>
<tr>
<th>Lecture</th>
<th>Topic</th>
<th>Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intro, schedule, and logistics</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Data Science components and tasks</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Data types</td>
<td>Project #1 out</td>
</tr>
<tr>
<td>4</td>
<td>Introduction to R, statistics foundations</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Introduction to D3, visual analytics</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Data preparation and reduction</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Data preparation and reduction</td>
<td>Project #1 due</td>
</tr>
<tr>
<td>8</td>
<td>Similarity and distances</td>
<td>Project #2 out</td>
</tr>
<tr>
<td>9</td>
<td>Similarity and distances</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Cluster analysis</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Cluster analysis</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Pattern miming</td>
<td>Project #2 due</td>
</tr>
<tr>
<td>13</td>
<td>Pattern mining</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Outlier analysis</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Outlier analysis</td>
<td>Final Project proposal due</td>
</tr>
<tr>
<td>16</td>
<td>Classifiers</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Midterm</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Classifiers</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Optimization and model fitting</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Optimization and model fitting</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Causal modeling</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Streaming data</td>
<td>Final Project preliminary report due</td>
</tr>
<tr>
<td>23</td>
<td>Text data</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td><strong>Time series data</strong></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Graph data</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Scalability and data engineering</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Data journalism</td>
<td>Final Project slides and final report due</td>
</tr>
</tbody>
</table>

*Bolded topics indicate additional assignments or due dates.*
Time-series database

- consists of sequences of values or events changing with time
- data is recorded at regular intervals
- characteristic time-series components
  - trends, cycles, seasonal, irregular

Applications

- financial: stock price, inflation
- industry: power consumption
- scientific: experiment results
- meteorological: precipitation
A time series can be illustrated as a time series graph which describes a point moving with the passage of time.
Categories of Time-Series Movements

- long-term or trend movements (trend curve): general direction in which a time series is moving over a long interval of time
- cyclic movements or cycle variations: long term oscillations about a trend line or curve
  - e.g., business cycles, may or may not be periodic
- seasonal movements or seasonal variations
  - i.e., almost identical patterns that a time series appears to follow during corresponding months of successive years.
- irregular or random movements

Time series analysis: decomposition of a time series into these four basic movements

- additive model: $TS = T + C + S + I$
- multiplicative model: $TS = T \times C \times S \times I$
Time series are in some sense similar to discrete sequences
  ▪ but differences apply
  ▪ discrete sequence data are not always temporal
  ▪ for example, gene data
  ▪ many of the similarity measures used for time series and discrete sequences can be reused across either domain
  ▪ but some of the measures are more suited to one of the domains

Distinguish between
  ▪ temporal (or placement) attribute
  ▪ behavioral attribute
May be subject to
- scaling
- translation
- noise

May show similar patterns of movements, but the absolute values may be very different
- mean and standard deviation may be different
- but patterns are similar
- difficult to compare when standard metrics are used
Also called *contextual* attribute

- in some applications different (simultaneous) time series may represent the same period of time (e.g., stocks)
- in other applications the time stamp is not important (e.g., medical data)
- in this case the time series need to be shifted for comparisons

Temporal (contextual) attribute scaling

- series may need to be stretched or compressed along the temporal axis to allow more effective matching
- may need to use different warp functions depending on time
Behavioral attribute translation:
- the behavioral attribute is mean centered during preprocessing

Behavioral attribute scaling:
- the standard deviation of the behavioral attribute is scaled to 1 unit

Normalization is generally easier for the behavioral attribute
- can typically be done during pre-processing
Standard pairwise distance

\[ \text{Dist}(\overline{X}, \overline{Y}) = \left( \sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \]

Shortcomings:
- designed for time series of equal length
- cannot address distortions on the temporal (contextual) attributes
Dynamic Time Warping Distance

Can better accommodate local mismatches

Three constraints
- no skipping of beginning or ends of either sequence
- continuity – no jumps
- monotonicity – can’t go back in time
DTW – FIND THE MINIMUM COST PATH

Euclidian

DTW
DTW – Find The Minimum Cost Path

DTW

Compute using dynamic programming
DTW Video

YouTube video
The freehand method
- Fit the curve by looking at the graph
- Costly and barely reliable for large-scaled data mining

The least-square method
- Find the curve minimizing the sum of the squares of the deviation of points on the curve from the corresponding data points

The moving-average method
Moving Average

Moving average of order $n$

\[
\frac{y_1 + y_2 + \cdots + y_n}{n}, \quad \frac{y_2 + y_3 + \cdots + y_{n+1}}{n}, \quad \frac{y_3 + y_4 + \cdots + y_{n+2}}{n}, \quad \cdots
\]

- Smoothes the data
- Eliminates cyclic, seasonal and irregular movements
- Loses the data at the beginning or end of a series
- Sensitive to outliers (can be reduced by weighted moving average)
Seasonal index

- Set of numbers showing the relative values of a variable during the months of the year
- E.g., if the sales during October, November, and December are 80%, 120%, and 140% of the average monthly sales for the whole year, respectively, then 80, 120, and 140 are seasonal index numbers for these months

Deseasonalized data

- Data adjusted for seasonal variations for better trend and cyclic analysis
- Divide the original monthly data by the seasonal index numbers for the corresponding months
Seasonal Index

Monthly Sales of Tyres

- Blue: Sales Data
- Green: Deseasonalized Data
- Red: Linear Regression Function

sales vs month

0 10 20 30 40 50 60

0 10 20 30 40 50 60
Similarity Search in Time-Series Analysis

Normal database query finds exact match
Similarity search finds data sequences that differ only slightly from the given query sequence

Two categories of similarity queries
- Whole matching: find a sequence that is similar to the query sequence
- Subsequence matching: find all pairs of similar sequences

Typical Applications
- Financial market
- Market basket data analysis
- Scientific databases
- Medical diagnosis
Can be day, week, month, year, and so on

Better fitting time windows could be found by periodicity
- do a Fourier analysis of the time sequence
- find significant frequencies = intrinsic periodicity in the periodogram

Example (see next page)
- world-wide daily page views of a web site over a 2070-day period (about 5.5 years).
- observe a strong sharp peak at 7 days, corresponding to the expected workday/weekend cycle
- smaller peak at 365 days (corresponding to a sharp dip each year during the winter holidays)
- smaller peak at 182 days (roughly a half-year), probably caused by increased use in the two-per-year semester cycle at universities.
PERIODOGRAM EXAMPLE

Time

Periodogram
Once time windows are established one can do

- clustering
- classification
- correlation analysis
- causal analysis
- predictive analysis
- outlier (anomaly) detection
- and so on
The value of $y_t$ at time $t$ is defined as a linear combination of the values in the immediately preceding window of length $p$

$$y_t = \sum_{i=1}^{p} a_i \cdot y_{t-i} + c + \epsilon_t$$

The values of the regression coefficients $a_1 \ldots a_p$, $c$ need to be learned from the training data.

Can use it to
- predict (forecast) future time events (given the change is small)
- compare other time series by predicting it