Announcements

- Final exam/presentation time: Tuesday, June 12th from 4-7pm
  - Who will be at Santa Cruz on that day?
Mining Time-Series, and Sequence Data
Mining Time-Series Data

- Time-series database
  - Consists of sequences of values or events changing with time
  - Data is recorded at regular intervals
  - Characteristic time-series components
    - Trend, cycle, seasonal, irregular
Price History - International Business Machines Corp... (3/17/97 - 3/20/98)

Compare with:
- IBM
- MSFT
- INTC

Adapt from ©2006 Jiawei Han and Micheline Kamber Eamonn Keogh, Gunopulos & Das
The Goal of Mining Time Series

- To predict future variables of interest
  - Stock price – almost classical
  - Trends in a product’s sales
    - Forecast future sales
    - Impact of a marketing campaign; Lenovo coupons
- Identify useful patterns:
  - Shift from classical forecasting to modern data mining
  - Discover stocks with similar movement in stock prices
  - Find cases in the past that resemble last month’s sales pattern of a product
  - Characterize a patient’s recovery based on similar patterns
  - Highly non-stationary and action dependent biological time series such as neurotrauma responses of head injury patients in Intensive Care Units (ICU)
- Label time sequences
  - DNA sequence analysis
  - Text annotation and information extraction
Categories of Time-Series Movements

- Categories of Time-Series Movements/Components
  - Long-term or trend movements (Trend curve)
  - Cyclic movements or cycle variations
  - Seasonal movements or seasonal variations
  - Irregular or random movements
Example: U.S.A. Population

- Ten-year intervals: 1790 to 1990.
Example: Accidental Deaths in U.S.A.

Example: International Airline Passengers

- Monthly totals: January 1949 to December 1960;
Example: Level of Lake Huron

- Annual levels (feet): 1875-1972.
- Linearly (?) decreasing trend.
Time Series Modeling

- Decomposing of a time series into these four basic movements/Components
  - Additive Modal: $TS = T + C + S + I$
  - Multiplicative Modal: $TS = T \times C \times S \times I$

Adapt from ©2006 Jiawei Han and Micheline Kamber Eamonn Keogh, Gunopulos & Das
Estimation of Trend Curve

- The moving-average method
- 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
Moving Average

- Moving average of order n

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

- 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)

Adapt from ©2006 Jiawei Han and Micheline Kamber Eamonn Keogh, Gunopulos & Das
Estimation of Seasonal Variations

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

Estimation of Cyclic Variations

- If (approximate) periodicity of cycles occurs, cyclic index can be constructed in much the same manner as seasonal indexes.
Estimation of Irregular Variations

Adjust the data for trend, seasonal and cyclic variations
Time Series Analysis Steps

**Step 1:** Model any trend and/or seasonal components that may be present, and remove these from the data.

**Step 2:** Choose from among a family of probability models that *best* represents the residuals from step 1.

**Step 3:** Estimate the parameters of the chosen model.

**Step 4:** Check model for goodness of fit.

**Step 5:** Resulting model:
- provides a compact description of the data, and can be used to interpret features therein;
- can be used for *in-sample* inference e.g. confidence intervals and hypothesis tests on model parameters;
- can be used for *out-of-sample* inference i.e. forecasting.
Decomposition of a Time Series

- **Additive model:** \( X_t = m_t + s_t + Y_t \)
- **Multiplicative model:** \( X_t = m_t s_t Y_t \)
  - \( m_t \): trend component
  - \( s_t \): seasonal component
  - \( Y_t \): noise/residual component (random, stationary).

**Aim:** Extract components \( m_t \) and \( s_t \), and hope that \( Y_t \) will be stationary. Then focus on modeling \( Y_t \).

We may need to do preliminary transformations (e.g., log or square-root) if the noise or amplitude of the seasonal fluctuations appear to change over time.
Stationarity

- We define $X_t$ to be **stationary** if the following hold:
  (i) The mean $\mu(t) = E(X_t)$ does not depend on $t$, and
  (ii) The covariance $\gamma(t+h,t) = \text{Cov}(X_{t+h}, X_t)$ depends only on $h$, for any integer $h$.

  This means that for any $h$, the behavior of $\{X_t, t=\ldots, -1,0,1,\ldots\}$, is similar to that of the “time-shifted” series, $\{X_{t+h}, t=\ldots, -1,0,1,\ldots\}$.

- White Noise
Sample Autocorrelations

If $X_t$ is stationary, we define its **autocovariance function** (ACVF) at lag $h$ as:

$$\gamma(h) \equiv \gamma(h,0) = \gamma(t+h,t)$$

and its **autocorrelation function** (ACF) at lag $h$ as:

$$\rho(h) \equiv \gamma(h) / \gamma(0) = \text{Corr}(X_{t+h}, X_t).$$

For each $h$, $-1 \leq \rho(h) \leq 1$. Also note that $\rho(0)=1$. 
Autocorrelation Plots for Checking Randomness

- Vertical axis: Autocorrelation coefficient
- Horizontal axis: Time lag $h$ ($h = 1, 2, 3, ...$)

Adapt from ©2006 Jiawei Han and Micheline Kamber
Eamonn Keogh, Gunopulos & Das
Sample Autocorrelations(2)

- If the series is purely random ie IID, then the normalized estimated autocorrelations are asymptotically standard normally distributed. This means that approximately 95% of the sample AC’s should fall between the bounds: $\pm \frac{1.96}{\sqrt{n}}$

- Note: deterministic features in a time series like trends and seasonality, make the stationary assumption unlikely.
- There are no significant autocorrelations.
- The data are random.
Autoregressive processes

- AutoRegression of order p, or AR(p):
  \[ X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + Z_t, \quad \{Z_t\} \sim WN(0,\sigma^2). \]

- This is similar to linear regression, except we are regressing the series on past values \(X_{t-1}, \ldots, X_{t-p}\) of itself.
- Parameter estimation: least square fit ...
MA(q) regresses the current value \{X_t\} against the white noise or random shocks of prior values of the series:

\[ X_t = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}, \quad \{Z_t\} \sim WN(0, \sigma^2). \]
ARMA Processes

- ARMA(p,q) process:
  \[ X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}, \]
  where \( \{Z_t\} \sim WN(0,\sigma^2) \)

- AR & MA are special cases: an AR(p)=ARMA(p,0), and an MA(q)=ARMA(0,q)

- ACF & PACF both decay exponentially
Building a Model

- Model identification
- Model estimation: Estimation of $\mu, \sigma^2, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$.
  - non-linear least squares
  - maximum likelihood estimation
  - Using statistical tools
- Model Validation: (Testing goodness-of-fit) Checking whiteness of residuals.
Model Validation

- Checking whiteness of residuals. Examine the residuals $R_t, \ t = 1, \ ..., \ n$
  - Time plot of residuals
  - Autocorrelation and partial autocorrelation plots of residuals
  - IID tests.
  - Normal probability plot
- Cross-validation
## Sample ACF for Model Identification

<table>
<thead>
<tr>
<th>SHAPE</th>
<th>INDICATED MODEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential, decaying to zero</td>
<td>Autoregressive model. Use the partial autocorrelation plot to identify the order of the autoregressive model.</td>
</tr>
<tr>
<td>Alternating positive and negative, decaying to zero</td>
<td>Autoregressive model. Use the partial autocorrelation plot to help identify the order.</td>
</tr>
<tr>
<td>One or more spikes, rest are essentially zero</td>
<td>Moving average model, order identified by where plot becomes zero.</td>
</tr>
<tr>
<td>Decay, starting after a few lags</td>
<td>Mixed autoregressive and moving average model.</td>
</tr>
<tr>
<td>All zero or close to zero</td>
<td>Data is essentially random.</td>
</tr>
<tr>
<td>High values at fixed intervals</td>
<td>Include seasonal autoregressive term.</td>
</tr>
<tr>
<td>No decay to zero</td>
<td>Series is not stationary.</td>
</tr>
</tbody>
</table>
Example 1

- Time plot indicates stationarity
- Doesn’t show any significant seasonality
A mixture of exponentially decaying and damped sinusoidal components $\Rightarrow$ autoregressive model.
AR(?) Model Might be Appropriate?

- => an AR(2) model might be appropriate.
Example 2

- Time plot indicates a rising trend.
  - A linear fit may remove this upward trend.
Time plot of residual after removing the trend indicates constant location and variance, which implies stationarity.

However, the plot does show seasonality. We generate an autocorrelation plot to help determine the period of seasonality.
AutoCorrelation

CO2 Concentrations for Mauna Loa Observatory

95% Confidence Band
99% Confidence Band

Autocorrelation

Lag

Adapt from ©2006 Jiawei Han and Micheline KamberEamonn Keogh, Gunopulos & Das
Seasonal SubSeries Plot Shows the Seasonal Patterns
AutoCorrelation
(Seasonally Differenced)
In summary, our initial attempt would be to fit an AR(2) model with a seasonal AR(12) term on the data with a linear trend line removed.
Non-stationary models

- **ARIMA (AutoRegressive Integrated Moving Average)**
  - Autoregressive parameters
  - Moving average parameters
  - Differencing (to remove trend or seasonality)
  - Slowly decaying ACF (difference until it decays rapidly)
Statistical Tools

- SemStat
- DataPlot
- XLMiner
- Matlab
- Minitab
- R, etc
With the systematic analysis of the trend, cyclic, seasonal, and irregular components, it is possible to make long- or short-term predictions with reasonable quality.
Time-Series & Sequential Pattern Mining

- Regression and trend analysis—A statistical approach
- Similarity search in time-series analysis
- Hidden Markov Model
Similarity Search in Time-Series Analysis

- Similarity search finds data sequences that differ only slightly from the given query sequence
- Typical Applications
  - Financial market
  - Market basket data analysis
  - Scientific databases
  - Medical diagnosis
Data Transformation

- Many techniques for signal analysis require the data to be in the frequency domain
- Usually data-independent transformations are used
  - The transformation matrix is determined a priori
    - discrete Fourier transform (DFT)
    - discrete wavelet transform (DWT)
- The distance between two signals in the time domain is the same as their Euclidean distance in the frequency domain
Discrete Fourier Transform

from $\vec{x} = [x_t], t = 0, \ldots, n - 1$ to $\vec{X} = [X_f], f = 0, \ldots, n - 1$:

$$X_f = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_t \exp(-j2\pi f t/n), \ f = 0, 1, \ldots, n - 1$$

- DFT does a good job of concentrating energy in the first few few coefficients
DFT (continued)

- Parseval’s Theorem

\[
\sum_{t=0}^{n-1} |x_t|^2 = \sum_{f=0}^{n-1} |X_f|^2
\]

- Keep the first few (say, 3) coefficients underestimates the distance and there will be no false dismissals!

\[
\sum_{t=0}^{n} |S[t] - Q[t]|^2 \leq \varepsilon \implies \sum_{f=0}^{3} |F(S)[f] - F(Q)[f]|^2 \leq \varepsilon
\]
Multidimensional Indexing in Time-Series

- Multidimensional index for efficient access
  - Using the first few Fourier coefficients
- Similarity search
  - Use the index to retrieve the sequences that are at most a certain small distance away from the query sequence
  - Post-processing: computing the actual distance between sequences in the time domain and discard any false matches
Subsequence Matching

- Break each sequence into a set of pieces of window with length $w$
- Extract the features of the subsequence inside the window
- Map each sequence to a “trail” in the feature space
- Divide the trail of each sequence into “subtrails” and represent each of them with minimum bounding rectangle
- Use a multi-piece assembly algorithm to search for longer sequence matches
Enhanced Similarity Search Methods

- **Allow for gaps** within a sequence or differences in offsets or amplitudes
- **Normalize** sequences with **amplitude scaling** and **offset translation**
- Two subsequences are considered **similar** if
  - one lies **within an envelope of \( \varepsilon \) width** around the other, ignoring outliers; or
  - they have enough **non-overlapping time-ordered pairs of similar subsequences**
- **Parameters** specified by a user or expert: **sliding window size**, **width of an envelope for similarity**, **maximum gap**, and **matching fraction**
Analysis of Similar Time Series

1. Original Sequences
2. Removing Gap
3. Offset Translation
4. Amplitude Scaling
5. Subsequence Matching

Adapted from ©2006 Jiawei Han and Micheline Kamber, Eamonn Keogh, Gunopulos & Das
Steps for Performing a Similarity Search

- Atomic matching
  - Find all pairs of gap-free windows of a small length that are similar
- Window stitching
  - Stitch similar windows to form pairs of large similar subsequences allowing gaps between atomic matches
- Subsequence Ordering
  - Linearly order the subsequence matches to determine whether enough similar pieces exist
Similar Time Series Analysis

VanEck International Fund

Fidelity Selective Precious Metal and Mineral Fund

Two similar mutual funds in the different fund group

Adapt from ©2006 Jiawei Han and Micheline Kamber Eamonn Keogh, Gunopulos & Das
Other State-Of-the-Art Methods

- Wavelets
- Change-point detection
- Kalman filters
- Hidden Markov Models, Conditional Random Field
  - Biological data, speech, text, etc.
- Dynamic Bayesian Networks