CS570 Introduction to Data Mining

Data Exploration and Data Preprocessing

Li Xiong
Data Exploration and Data Preprocessing

- Data pre-processing
  - Data cleaning
  - Data integration
  - Data transformation

- Data reduction
  - Instance reduction
  - Dimension reduction
    - Feature selection
    - Feature extraction/creation
Feature Extraction

- Create new features (attributes) by combining/mapping existing ones
- Methods
  - Principle Component Analysis
  - Fourier transform
  - Discrete Wavelet Transform
Principal Component Analysis (PCA)

- Principle component analysis: find the dimensions that capture the most variance
  - A linear mapping of the data to a new coordinate system such that the greatest variance lies on the first coordinate (the first principal component), the second greatest variance on the second coordinate, and so on.

- Steps
  - Normalize input data: each attribute falls within the same range
  - Compute $k$ orthonormal (unit) vectors, i.e., principal components - each input data (vector) is a linear combination of the $k$ principal component vectors
  - The principal components are sorted in order of decreasing “significance”
  - Weak components can be eliminated, i.e., those with low variance
**Dimensionality Reduction: PCA**

- **Goal** is to find a projection that captures the largest amount of variation in data.

- **Mathematically**
  - Compute the covariance matrix
    \[ \text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]. \]
  - Find the eigenvectors of the covariance matrix corresponding to large eigenvalues
    \[ Av = \lambda v. \]
Fourier transform

**Fourier transform: time-domain to frequency-domain**

\[ F(\nu) = \int_{-\infty}^{\infty} f(t) \cdot e^{-i2\pi \nu t} \, dt. \]

Two time series

Two time series + Noise

Frequency
Wavelet Transformation

- A common data compression technique (lossy)
- Discrete wavelet transform (DWT): linear signal processing technique divides signal into different frequency components
- Data compression/reduction: store only a small fraction of the strongest of the wavelet coefficients
- Discrete wavelet functions
  - Haar wavelet
  - Daubechies wavelets
DWT Algorithm

- Pyramid algorithm - averaging and differencing method
  - Input data of length L (an integer power of 2)
  - Each transform has 2 functions: smoothing, differencing
  - Applies to pairs of data, resulting in two set of data of length L/2
  - Applies two functions recursively, until reaches the desired length
  - Select coefficients by threshold

- Haar Wavelet Transform
  - Haar matrix (sum and difference) $H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$
    - Given $2^n$ sequence, resulting in $2^n - 1$ differences and one final sum
  - Example: (4,6,10,8,1,9,5,3)

- Filtering of data
  - Low frequency filter (averaging); High frequency filter (differencing)
  - Different features of a signal (background, edges, etc.) correspond to different frequencies

- Advantage over Fourier Transform?
Example of DWT Based Image Compression

DWT compression for test image Lenna (threshold = 1)
Summary

- Data Exploration and Data Preprocessing
  - Data and Attributes
  - Data exploration
    - Descriptive statistics
    - Data visualization
  - Data pre-processing
    - Data cleaning
    - Data integration
    - Data transformation
    - Data reduction