Data Mining

Data

Data Preprocessing

Dimensionality Reduction

Fourier Transform

Wavelet Transform

Haar Wavelet Transform

$$\psi(t) = \begin{cases} 1 & 0 \leq t < 1/2, \\ -1 & 1/2 \leq t < 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$\phi(t) = \begin{cases} 1 & 0 \leq t < 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$h[n] = \begin{cases} \frac{1}{\sqrt{2}} & \text{if n} = 0,1 \\ 0 & \text{otherwise} \end{cases}$$

PCA - Principle component analysis

PCA just project raw data to another space, using eigvectors as basis vectors. It reduces some dimensions of the data while remaining most of the information in the data.

1. Calculate covariance matrix cov_matrix for each feature.
$$\sum_i \frac{(a_i - E_a)(b_i - E_b)}{n-1}$$
2. Calculate Eigenvalues $\lambda_1 \cdots \lambda_n$ and Eigenvectors $\epsilon_1 \cdots \epsilon_n$ of cov_matrix
3. Order $\lambda_1 \cdots \lambda_n$ and choose the top-k λs and related ϵs
4. The size of origin data is $m \times n$, the transform matrix (size $n \times k$) consists of k vectors $[\epsilon_1 \cdots \epsilon_k]$.
5. Transformed data is $origin \times trans$, and its size is $m \times k$
6. To reconstruct the origin data, just use $transformed \times trans^T$

Pattern Mining

Basic Pattern Mining

Apriori

Just one principle:

> $sup(S_1) >= sup(S_2)$ when $S_1 \subseteq S_2$

So we can do pruning using this principle.

FP-Growth

1. Order the supports of frequent items, and order items in transactions. Record the items (if on conditional FP-tree, also record the condition part).
2. Build FP-tree
3. Build conditional FP-tree, repeat this process on conditional FP-tree

Advanced Pattern Mining

Sequential Pattern Mining

GSP - Generalized Sequential Pattern

Just an algorithm to generate candidate (k+1)-sequences from k-sequences.

$S_1 + S_2 \rightarrow S_3$ and $S_3 = S_1 + S_2[-1]$ iff $S_1[2:] = S_2[:-1]$

Stream Data Pattern Mining

Total size N

Error rate σ

Support rate s

Use buckets to process data, and after each bucket, decrease count by 1.

Pattern with support over $(s - \sigma)N$ are found