12/5/04

supervised/unsupervised learning

sequence analysis ≡ t.s.a.

$\underset{\sim}{X} = \{x(1), \ldots, x(T)\}$

Retrieval - find subsequence best matching given sequence

Query

$\underset{\sim}{Q} = \{q(1), \ldots, q(m) \ldots\}$

distance $\| \underset{\sim}{X}_t - \underset{\sim}{Q} \|$

nearest neighbor

slide along

$\underset{\sim}{X}_t$: stretch of length m starting at t

Euclidian

$$[x(t) - q(1)]^2 + \ldots + [x(t+m) - q(m)]^2$$

City block

$$|x(t) - q(1)| + \ldots + |x(t+m) - q(m)|$$

Cosine

$$(\underset{\sim}{X}_t, \underset{\sim}{Q}) \Big/ \sqrt{(\underset{\sim}{X}_t, \underset{\sim}{X}_t)^2 (\underset{\sim}{Q}, \underset{\sim}{Q})^2}$$

similarity search

- trend
- seasonal
- repeating pattern

SARIMA(0,1,1), period 7.

(0,0,1) period 7

$$Y_t \sim (1-\Theta)\left[Y_{t-7} + \Theta Y_{t-14} + \Theta^2 Y_{t-21} + \dots\right] = Y_t(\Theta)$$

$$\hat{Y}_{n+1} = (1-\hat{\Theta})\left[Y_{n-6} + \hat{\Theta} Y_{n-13} + \dots\right]$$

$$SS(\Theta) = \sum_t \left[Y_t - Y_t(\Theta)\right]^2$$

Expression employed.

$$Y_t \sim (1-\Theta)\left[Y_{t-7} + \Theta Y_{t-14} + \Theta^2 Y_{t-21} + \dots\right]$$

$$+ \Phi Y_{t-1} - \Phi(1-\Theta)\left[Y_{t-8} + \Theta Y_{t-15} + \Theta^2 Y_{t-22} + \dots\right]$$

$$= Y_t(\Theta, \Phi)$$

$$SS(\Theta, \Phi) = \sum_t \left[Y_t - Y_t(\Theta, \Phi)\right]^2$$

Validation

12/5/04

Association rules/ Market basket analysis

co-occurrence of items

n x p matrix of 0's and 1's

dummy variables

association rule

expression

$$\{z_1, z_2, \ldots, z_q\} \Rightarrow \{z_r, \ldots, z_p\}$$

(A: $z_1, \ldots, z_q$; B: $z_r, \ldots, z_p$)

meaning

$z_1 = 1, \ldots, z_q = 1$ implies* $z_r = 1, \ldots z_p = 1$

---

*"good chance of finding"

12/5/04

Often care about rules involving frequent sets

support of $A \Rightarrow B$

$fr(A \cup B)$

confidence of $A \Rightarrow B$

$fr(A \cup B) / fr(A)$

lift of $A \Rightarrow B$

$fr(A \cup B) / \{ fr(A)\, fr(B) \}$

Find all cases with confidence > p_c and support > p_s

Data too large to fit in main memory

a priori trick

- C frequent, every subset of C frequent

Apriori algorithm

12/5/04

Safeway example

customer and product vectors

sought minima

support

confidence

lift

to create recommendations

" statisticians need to focus on interpretation of the rules that the algorithms produce, not the computational efficiency"

Hand, Mannila & Smyth (2001)