T-122.103 EXERCISE 4/2003

In T4 on 24 October 2003 at 12:15–14 o'clock.

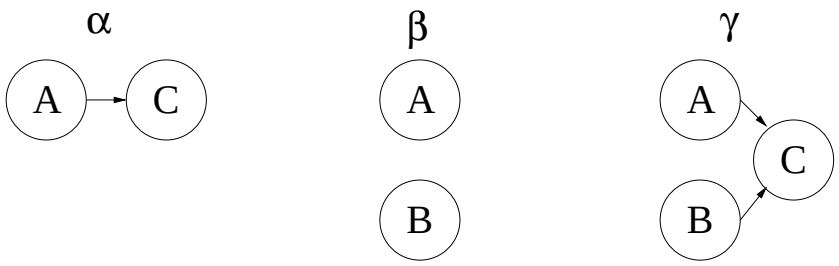

Figure 1: Some episodes.

1. Consider episodes α, β and γ (figure 1) in an event sequence $s = \langle(A,1),(B,2),(C,4),(A,9),(A,13),(B,20),(A,22),(C,23),(A,28),(B,32),(C,35)\rangle$.

   (a) Find frequencies of the episodes (e.g. $fr(\alpha, s, W)$) for some suitable time windows W.

   (b) What can be said of the frequencies of the episodes α, β and γ in an arbitrary event sequence?

   (c) Find the sets of minimal occurences, $mo(\alpha)$, $mo(\beta)$ and $mo(\gamma)$, and the corresponding supports.

   (d) What is the confidence of a rule $\beta[3] \Rightarrow \gamma[5]$?

2. Consider a very long event sequence s containing events of types A and C distributed uniformly at random over the time interval. Let ν_A and ν_C be the expected number of events A and C occuring in a time window W, respectively. What is the expectation of the frequency $fr(\alpha, s, W)$, where the episode α is given by figure 1?

3. When the error parameter ϵ of the error-tolerant frequent itemset (ETI) definition is increased or the support parameter κ is decreased, the number of ETIs in any data will increase, even if there is no structure in the data. The objective of this exercise is to estimate what kind of ETIs would be expected in completely random data.

   Assume that a binary $n \times d$ matrix is generated at random, so that each position in the matrix has a 1 with probability p independently of all

the other positions. Find an upper bound for the probability that the data has at least one ETI of size r with parameters ϵ and κ. If $n = 10^6$, $d = 500$, $p = 0.15$, $\epsilon = 0.2$, and $\kappa = 0.01$, what size ETIs would lead you to suspect that there is indeed some structure in the data?

4. Fix the support parameter κ, and consider the error rates of weak ETIs. Denote by $\delta(E)$ the smallest number such that E is a weak ETI with parameters κ and $\epsilon = \delta(E)$. Show that if $E = F \cup G$ with $F \cap G = \emptyset$,

$$\delta(E) \cdot |E| \geq \delta(F) \cdot |F| + \delta(G) \cdot |G|.$$

5. Prove the results about "lift" between two attributes in a topic model:

   (a) If A is a core attribute of topic T_j, i.e., only T_j can ever generate A,

$$\text{Lift}(A, B) = \frac{P(T_i \mid B)}{P(T_i)}$$

   for any other (core or not) attribute B.

   (b) If A and B are core attributes of the same topic T_j,

$$\text{Lift}(A, B) = P(T_j)^{-1}.$$

   (c) If A and B are core attributes of two different topics,

$$\text{Lift}(A, B) = 1.$$