Exam Data Mining 7-11-2018

Short answer indications

Q1: 1.A 2.A 3.A 4.B 5.B 6.A 7.B 8.B 9.B 10.A

Q2: (a) A,B,C,D,AB,AC,AD
    (b) C, CD, and CDA
    (c) 0.01

Q3: Level 1: A:7,B:5,C:3,D:2,E:2, all of them frequent
    Level 2: AB:4,AC:2,AD:2,AE:2,BC:2,BD:1,BE:1,CD:0,CE:0,DE:0, first 5 are frequent
    Level 3: ABC:1, not frequent. Pruned pre-candidates: ABD,ABE,ACD,ACE,ADE

Q4: (a) No closed form solutions because of chordless 4-cycle A-B-C-D-A (and B-C-D-E-B)
        Number of parameters: 1+5+7+2=15
    (b) \hat{n}(A,B,C,D) = (n(A,B,D)*n(B,C,D))/n(B,D)
        Number of parameters: 1+4+5+2=12
    (c) \hat{n}(A,B,C,D) = (n(A,B)*n(B,D)*n(C))/(N*n(B))
        Number of parameters: 1+4+2=7

Q5: (a) Score of node "treatment" in initial model: 350 * log(350/700) + 350 * log(350/700) = -485.203
        Score of node "treatment" after adding edge from "size" to "treatment":
        87 * log(87/357) + 270 * log(270/357) + 263 * log(263/343) + 80 * log(80/343) = -384.547
        Change in score: -384.547 + 485.203 = 100.656

    (b) add(outcome --> treatment)

    (c) Treatment A is more effective. The severe cases (large kidney stones) are more likely to
        receive treatment A, and are at the same time less likely to recover. To determine the effect
        of treatment on outcome we must therefore control for size.
        Hence, the size-specific success percentages give the correct result.