Data Mining, Week 4, Individual assignments

This week, you'll be doing reading, small pen-and-paper assignments as well as a small bit of programming. You will familiarize yourself with compact representations of itemsets, generation of association rules, and a few objective measures of interestingness. Note that these assignments are due by April 14th due to the easter holiday. Guidance will be provided in B221 on Friday the 10th of April at 14-16hrs, and on Monday the 13th of April at 10-12hrs. If additional guidance is needed, just ask!

Note! Due to the easter holiday, the assignments are due by the lecture on 15th.

For the first assignments, we use the data from the individual assignments in Week 2, replicated the bottom of this document.

1. Consider a dataset that has items A, B, C, D.
    a. Assume ACD is frequent. Which itemsets must then be frequent?
    b. Assume further that no superset of ACD is frequent. Which itemsets must then be infrequent?
    c. Assume further that B is frequent and that no superset of B is frequent. List all frequent itemsets, all infrequent itemsets, and all itemsets that remain undecided.

* 2. Study the concept of *maximal frequent itemsets* from the coursebook (p. 354->), and define it in your own words in the learning diary.

3. See the data and figure at the last page of this document, which we have also visited during the individual assignments from the second week. Define a minsup-value that allows you to illustrate maximal frequent itemsets using the picture. What is your minsup-value, and what are the maximal frequent itemsets for that value?

4. Consider a dataset that has items A, B, C, D.
    a. Assume that conf(A -> C) = 1, i.e., that C always occurs when A occurs, and that supp(A)=0.4, supp(AB)=0.2, supp(AD)=0.1. What are supp(AC), supp(ABC), and supp(ACD)? (They can all be inferred from the information given.)
    b. Assume that supp(A)=0.3, supp(B)=0.4, supp(C)=0.5, supp(AB)=0.3, supp(AC)=0.3, supp(BC)=0.35, supp(BD)=0.4. What are supp(ABC), supp(ABD), supp(BCD), and supp(ABCD)? (Hint: which rules have confidence 1? Can they be used the same way as in a) above?)

* 5. Study the concept of *closed itemsets* from the coursebook (p. 355->), and define it in your own words in the learning diary.

6. See the data and figure at the last page of this document again. Define a minsup-value that allows you to illustrate closed itemsets using the picture. What minsup-value did you choose, and what are the closed itemsets for that value?

* 7. Study the concept of *closed frequent itemset* from the coursebook (p. 356->), and define it in your own words in the learning diary.

8. See the data and figure at the last page of this document again. Define a minsup-value that allows you to illustrate closed frequent itemsets using the picture. What minsup-value did you choose, and what are the closed frequent itemsets for that value?

9. Assume that supp(B)=0.7, supp(BC)=0.5, supp(ABD)=0.6, and supp(ABCD)=0.3 are the supports of the frequent closed itemsets in a dataset over items A, B, C, D, E. What do you know about the supports of the following itemsets: A, BD, E, ADE?

* 10. Reflect in your learning diary about the purpose of the concepts from 2, 5 and 7. Why would someone use them, and why they could be useful? (Note, references updated)

* 11. In the course data, what are the closed frequent itemsets with support over 0.05? Note, if your implementation cannot handle 0.05, pick a higher support. As there might be quite a few of them, just write down the ballpark -estimate of the amount, and your observations.

Next, let us delve back into programming, just for a moment, and remind ourselves again about association rules. Read the course book on association rules (p. 329->) and rule generation (p. 349->).

12. (Exercise 17 from coursebook, p. 413).

13. Using the answer from exercise 8, what rules can you identify from the closed frequent itemsets, and what are their confidences? Seek to list at least a few.

* 14. Implement the rule generation procedure as described in pages 351 and 352. List (1) five rules that contain the course "Introduction to programming" as a consequent (Ohjelmoinnin perusteet) that have high confidence, and (2) five rules that contain the course "Introduction to programming" as a consequent (Ohjelmoinnin perusteet) but have low confidence. Write your observations into the learning diary.

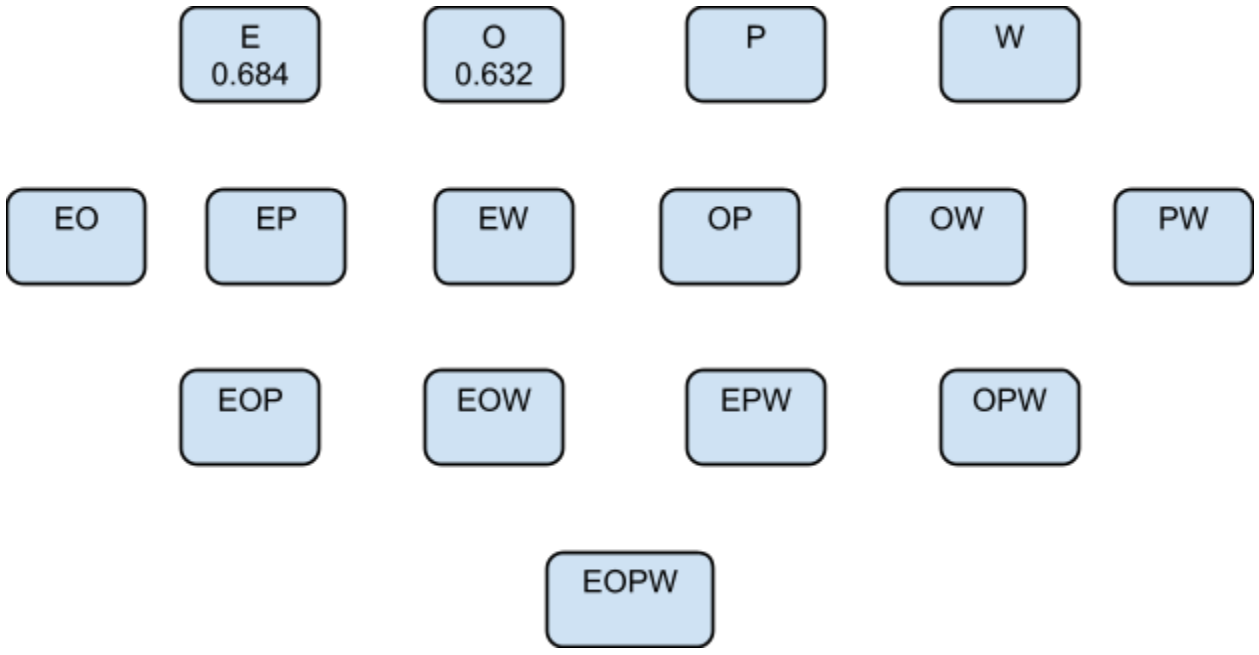
Next, let us consider approaches for measuring interestingness of a set of features. For this, read the chapter 6.7. from the coursebook (p. 370->).

* 15. Study and then implement the concept of *lift* into your rule generation procedure. Describe lift in your learning diary, and explain what the lift values for the rules that you extracted in ex. 14. are. Reflect on your findings.

* 16. Implement the calculation of *IS* into your rule generation procedure. Describe IS in your learning diary, and explain what the IS values for the rules that you extracted in ex. 14. are. Reflect on your findings.

17. Finally, read about the Simpson's paradox, p. 384. How could the paradox influence our interpretations from the course data?

Figure:



Book orders:

T O I
T I E T O
H E T I
W E T O
P I H W I
T I I P I I
P O E T
E H T O
T E H O
H I I H T O
P I H W I T
P E T I
H E P
P E T O
P I T O
H E P O
O T E
P E I P P O
T I E