Data Mining - The Diary

Rodion "rodde" Efremov

April 21, 2015

Introduction

This document is my learning diary written on behalf of Data Mining course led at spring term 2015 at University of Helsinki.

1 Week 1

The **support count** $\sigma(X)$ of an item set X is the amount of transactions containing X ($X \subset t_i$). Basically, we were computing support counts for various itemsets with the exception of applying additional constraints to the queries (such as particular grade range).

The **support** of an item set X is $\sigma(X)/N$, where N is the amount of all transactions. Support of X may be thought of as a classical probability of a random transaction containing X.

An **association rule** is an implication of the form $X \to Y$, where X and Y are itemsets having no items in common. The interpretation of an association rule is that if a transaction contains X, it "tends" to contain Y as well. Note that "tends" depends on parameters we specify to a data mining system. **Support** of an association rule $X \to Y$ is

$$s(X \to Y) = \frac{\sigma(X \cup Y)}{N}.$$

Support of the rule R may be thought of as a classical probability of R appearing in a random transaction. **Rule confidence** gives the probability of Y appearing in the same transactions with set X and is defined as

$$c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}.$$

1.1 Reflection

Getting the data from a file to internal representation was pretty challenging: the data seems a little bit "dirty" and I am sure there is room for improvement. What comes to accessing data, I have made an effort to make sure that it runs fast. Basically I have three model classes:

`Course` holds the course name, the course code, grading mode and the amount of credits awarded, `Student` holds only a unique student ID and enrollment year,

`CourseAttendanceEntry` holds a course C, a student S, the year and month S attended C, and the grade S received. Basically, these entries implement a many-to-many relationship between courses and students.

2 Week 2

Task 5

The supports are as follow:

E	0.684
O	0.632
P	0.526
W	0.158
EO	0.474
EP	0.316
EW	0.053
OP	0.263
OW	0.053
PW	0.105
EOP	0.221
EOW	0.053
EPW	0
OPW	0
EOPW	0

The only observation that I was able to come up with is that if $s(X)$ is support of an itemset X, then

$$s(X) \leq \min_{A \subsetneq X} s(A).$$

Task 10

We have around 23 million (N) different paperback books and we want to generate all 10-combinations of those. Suppose we are given an index tuple $t = (t_1, t_2, \dots, t_{10}) = (1, 2, \dots, 10)$. Next generate a combination of books indexed by t and increment t_{10}. When $t_{10} = N + 1$, increment t_9 and set $t_{10} = t_9 + 1$. After $t_9 = N - 1$ (and thus $t_{10} = N$) has been generated, increase t_8 and set $t_9 = t_8 + 1, t_{10} = t_8 + 2$. Continue this routine until $t_1 = N - 9, t_2 = N - 8, \dots, t_9 = N - 1, t_{10} = N$.

Task 15

In this task we are supposed to measure time of generating k-combinations of courses for $k \in \{2, 3, 5\}$. The results are summarized in the following table:

k	t
2	4 ms
3	40 ms
5	291 ms

Increasing k from 2 to 3 increases the running time by a factor of 10; increasing k from 3 to 5 increases the running time by a factor of 7,3. Since $n = 213$,

$$\begin{aligned}\binom{n}{3}\binom{n}{2}^{-1} &= \frac{n!2!(n-2)!}{n!3!(n-3)!)} \\ &= \frac{(n-2)!}{3(n-3)!} \\ &= \frac{n-2}{3} \\ &\approx 70,\end{aligned}$$

and

$$\begin{aligned}\binom{n}{5}\binom{n}{3}^{-1} &= \frac{n!3!(n-3)!}{n!5!(n-5)!)} \\ &= \frac{(n-3)!}{20(n-5)!} \\ &= \frac{(n-4)(n-3)}{20} \\ &\approx 2100,\end{aligned}$$

which does not quite go hand in hand with the measurements.

Task 19

The objective of this task is to compare brute-force and Apriori algorithms for frequent itemset generation.

k	support	Brute-force (ms)	Apriori (ms)
2	0.3	379	154
3	0.175	9389	774
4	0.1	N/A	1845
5	0.1	N/A	1637

After Arto's counsel, I was able to speedup generation of 3-combinations by a factor of 20, but I was not able to make 4-combination generation feasible.

Task 21

The largest size of itemsets with support at least 0.05 seems to be 11. I got 19 of such itemsets; one of them is

- TVT-ajokortti
- Ohjelmoinnin perusteet

- Opiskelutekniikka
- Tietokantojen perusteet
- Ohjelmoinnin jatkokurssi
- Tietoliikenteen perusteet
- Tietorakenteet ja algoritmit
- Johdatus tietojenkäsittelytieteeseen
- Tietokone työvälineenä
- Ohjelmistotekniikan menetelmät
- Aineopintojen harjoitustyö: Tietokantasovellus

3 Week 3

Task 10

Given a set of events $E = \{e_1, e_2, \ldots, e_d\}$, a sequence s over E is $\langle S_1, S_2, \ldots, S_n \rangle$, where $\varnothing \neq S_i \subseteq I$ for all i. The sequence $t = \langle t_1, \ldots, t_k \rangle$ is said to be a *subsequence* of s if there exist integers $1 \leq j_1 < j_2 < \cdots < j_k \leq n$ such that $t_i \subseteq S_{j_i}$ for all $i = 1, 2, \ldots, k$.

Task 11

We are given events A, B and C. All possible 1-sequences are:

1. $\langle \{A\} \rangle$
2. $\langle \{B\} \rangle$
3. $\langle \{C\} \rangle$

All possible 2-sequences are:

1. $\langle \{A, B\} \rangle$
2. $\langle \{A, C\} \rangle$
3. $\langle \{B, C\} \rangle$
4. $\langle \{A\} \{A\} \rangle$
5. $\langle \{A\} \{B\} \rangle$
6. $\langle \{A\} \{C\} \rangle$
7. $\langle \{B\} \{A\} \rangle$
8. $\langle \{B\} \{B\} \rangle$
9. $\langle \{B\} \{C\} \rangle$

10. $\langle \{C\}\,\{A\} \rangle$
11. $\langle \{C\}\,\{B\} \rangle$
12. $\langle \{C\}\,\{C\} \rangle$

Above we have $d = 3$, which produces 12 2-sequences. For general $d > 1$, there would be $\binom{d}{2} + d^2$ 2-sequences.

Task 13

We are given the following sequences:

- $\langle \{B\}\,\{C\}\,\{H\} \rangle$
- $\langle \{B, P\}\,\{C\} \rangle$
- $\langle \{C\}\,\{H\}\,\{P\} \rangle$
- $\langle \{P\}\,\{C, H\} \rangle$
- $\langle \{T\}\,\{B\}\,\{C\} \rangle$
- $\langle \{T\}\,\{B, P\} \rangle$
- $\langle \{T\}\,\{P\}\,\{C\} \rangle$

where B is for bathroom, C is for computer, H is for homework, P is for phone and T is for TV.

Now the possible 4-candidates are:

- $\langle \{B\}\,\{C\}\,\{H\}\,\{P\} \rangle$
- $\langle \{T\}\,\{B\}\,\{C\}\,\{H\} \rangle$
- $\langle \{B, P\}\,\{C, H\} \rangle$
- $\langle \{T\}\,\{B, P\}\,\{C\} \rangle$
- $\langle \{T\}\,\{P\}\,\{C, H\} \rangle$
- $\langle \{P\}\,\{C, H\}\,\{P\} \rangle$

Task 15

The supports for *maxspan* of 1 is as follows:

Sequence	Support
⟨{courses} {courses}⟩	0.2
⟨{courses} {dm}⟩	0.6
⟨{dm} {courses}⟩	0.2
⟨{index} {courses}⟩	0.0
⟨{teaching} {dm}⟩	0.0

The *maxspan* is a pruning parameter: let t_1 be the moment at which the very first event begins and t_2 the moment at which the very last event ends, then the sequence is pruned if $t_2 - t_1 > \textit{maxspan}$.

Task 16

The top five 2-sequences are:

Sequence	Support
Lineaarialgebra ja matriisilaskenta I, Lineaarialgebra ja matriisilaskenta II	0.326
Lineaarialgebra ja matriisilaskenta I, Analyysi I	0.317
Turvallinen työskentely laboratoriassa, Yleinen kemia I	0.259
Analyysi I, Analyysi II	0.257
Yleinen kemia I, Yleinen kemia II	0.236

Task 18

The top 5 8-sequences are:

1. Turvallinen työskentely laboratoriossa
2. Yleinen kemia I
3. Kemian orientoivat opinnot
4. Yleinen kemia II
5. Orgaanisen kemian perustyöt I
6. Liouskemian perusteet
7. Atomien ja molekyylien rakenne
8. Kemian tietolähteet

with support 0.0211,

1. Turvallinen työskentely laboratoriossa
2. Yleinen kemia I
3. Kemian orientoivat opinnot
4. Yleinen kemia II
5. Orgaanisen kemian perustyöt I
6. Matematiikkaa kemisteille
7. Atomien ja molekyylien rakenne
8. Kemian tietolähteet

with support 0.0204,

1. Turvallinen työskentely laboratoriossa
2. Yleinen kemia I
3. Kemian orientoivat opinnot

4. Yleinen kemia II
5. Orgaanisen kemian perustyöt I
6. Liouskemian perusteet
7. Orgaanisten yhdisteiden rakenteiden selvittäminen
8. Integroidut TVT-opinnot

with support 0.0204,

1. Turvallinen työskentely laboratoriossa
2. Yleinen kemia I
3. Kemian orientoivat opinnot
4. Yleinen kemia II
5. Orgaanisen kemian perustyöt I
6. Liouskemian perusteet
7. Orgaanisten yhdisteiden rakenteiden selvittäminen
8. Kemian tietolähteet

with support 0.0204,

1. Turvallinen työskentely laboratoriossa
2. Yleinen kemia I
3. Kemian orientoivat opinnot
4. Yleinen kemia II
5. Orgaanisen kemian perustyöt I
6. Liouskemian perusteet
7. Matematiikkaa kemisteille
8. Atomien ja molekyylien rakenne

with support 0.0204.

It is obvious that the above five sequences are very alike. Actually the four last sequences have exactly the same support.

Task 19

What comes to the results in Task 18, doing the same with *maxspan* produces a result with “less variation”. This can be explained by assuming that those students that “fit in“ *maxspan* of 36 months, tend to perform the same course permutation. On behalf of Task 17, applying the *maxspan* of 36 months produces the same sequences (with slightly smaller supports each). This can be explained by assuming that 36 months is enough for any student in the data to score 5 courses.

4 Week 4

Task 2

A frequent itemset is *maximal* if adding any item to it makes it infrequent. By Apriori principle, any subset of a frequent maximal itemset is also frequent. (It is hard to avoid rephrasing the definition in the course book.)

Task 5

A closed itemset X is an itemset for which all of its supersets have support less than the support of X.

Task 7

A closed frequent itemset X is a closed itemset whose support is at least *minsup* (in which case, X is called "frequent").

Task 10

Any itemset having support no less than *minsup* is considered to be "interesting" due to the fact that it occures in the database "frequently". Once a maximal frequent itemset X is found, we know that all its subsets $A \subseteq X$ are frequent too, and all supersets of X will be non-frequent. The purpose of a closedness of an itemset is as follow: if X is closed and non-frequent, there is no way any its superset A can be frequent.

Task 11

The set EOW is frequent because its support is 0.053 (see Week 2, Task 5). It is also closed because all of its supersets (only EOPW in this case; has support 0) has same support as EOW.

Also, E, O, P, W, EO, EP, OP, PW, EOP are closed frequent itemsets. EW, OW, EPW, OPW are not closed. EOPW is not frequent.

Task 14

We used the support of 0.15 in order to "get" to rules in which the Introduction to programming course (fin. Ohjelmoinnin perusteet) is in consequent. The following table lists five rules with maximum confidence (which is 1 one for all five rules).

Rule	Support	Confidence
{ JTKT } → { OHPE, OHJA, TITY }	0.162	1.0
{ OHJA } →{ OHPE, TITY, OHME }	0.162	1.0
{ JTKT } →{ OHPE, TITY, OHME }	0.162	1.0
{ OHJA } →{ JTKT, OHPE, TITY }	0.162	1.0
{ OHME } →{ JTKT, OHPE, TITY }	0.162	1.0

Above, we used the following abbreviations:

JTKT Johdatus tietojenkäsittelytieteeseen

OHPE Ohjelmoinnin peruskurssi

OHJA Ohjelmoinnin jatkokurssi

OHME Ohjelmistotekniikan menetelmät

TITY Tietokone työvälineenä

Above, one can see that the five courses are related to each other, since the set of all five courses are "popular" among computer science students.
The following table lists five rules with "Introduction to programming" in the antecedent and with low confidence:

Rule	Support	Confidence
{ TITY, OHME } → { OHPE, OHJA }	0.161	0.344
{ TITY, JTKT } →{ OHPE, OHJA }	0.161	0.344
{ TIKAPE, OHME } →{ OHPE, OHJA }	0.161	0.344
{ TIKAPE } →{ OHPE, OHJA }	0.161	0.344
{ TITY } →{ OHPE, OHJA }	0.161	0.344

Above, TIKAPE stands for "Tietokantojen perusteet" (engl. Introduction to Databases). Since OHPE is in consequent, I think that the above rules apply to students whose major is not computer science. Also, there seems no strong relation between those courses.

Task 15

The *lift* of an association rule $X \to Y$ is defined as

$$Lift = \frac{c(X \to Y)}{s(Y)} = \frac{\sigma(X \cup Y)/\sigma(X)}{\sigma(Y)/N} = \frac{N\sigma(X \cup Y)}{\sigma(X)\sigma(Y)},$$

or namely as a ratio of the rule's confidence and the support of its consequent. Since $s(Y) \leq 1$, $Lift \geq c(X \to Y)$. I am not quite positive, but it appears to me that *lift* communicates "coolness" of an association rule. Now, what comes to lifts for the rules in Task 14, the low confidence rules have *lift* of around 1.2 and the *lift* values for high-confidence rules are almost up to 6.0.

Task 16

The IS measure is defined as

$$\begin{aligned} \text{IS}(A, B) &= \sqrt{c(A \to B) \cdot c(B \to A)} \\ &= \sqrt{\frac{\sigma(A \cup B)\sigma(B \cup A)}{\sigma(A)\sigma(B)}} \\ &= \frac{\sigma(A \cup B)}{\sqrt{\sigma(A)\sigma(B)}}. \end{aligned}$$

The IS measures for rules from the Task 14 do not seem to variate too much (approximately within the range $[0.7, 0.95]$),

5 Week 5

Task 6

It seems reasonable to code the grades by means of seven items: PASS, FAIL, 1, 2, 3, 4, 5.

Task 7

The amount of credits for a course and students' enrollment years.

Task 11

	FAIL	1-3	4-5
Introduction to programming	81	168	393
Advanced programming	73	154	295
Both	154	322	688

Task 13

I am pretty confident that grade is a categorical attribute.

Task 16

The mean grade in question is ≈ 2.6.

Task 17

The mean grade in question is ≈ 2.8.

Task 18

The mean grade in question is ≈ 3.4.