Association analysis

Lectures based on this

Association Rules

- Assume access to some transaction data.
- We are interested in finding rules on the form $i_1, ..., i_m -> i_{m+1}, ...i_n$
  - i.e given previous items bought we want to find items that are likely to be bought in the future.
- Application: market basket analysis to support business decisions.

Example: $milk, eggs -> bread, butter$

$Smoking -> Cancer$ (higher probability) $Cancer -> Smoking$ (Not causal, cancer does not result in smoking)

- We are interested in finding rules on the form described above with user defined minimum-support and confidence, where
  - Support = fraction for the transactions which contain X and Y ($p(X, Y)$)
  - Support = how general the rule is
  - Confidence = fraction of the transactions that contain X which also contains Y ($p(Y|X)$).
  - Confidence = how accurate the rule is
  - Confidence = support(X,Y) / support(X).

Example:

```
Transaction id  Items bought
+---------------+-----------+
|1              |a,b,c      |
+---------------+-----------+
|2              |a,c,d      |
+---------------+-----------+
|3              |a,d,e      |
+---------------+-----------+
|4              |b,e,f      |
+---------------+-----------+
|5              |b,c,d,e,f  |
+---------------+-----------+
```

$a -> d$ has support 0.6 and confidence 1 (3/5), (3/3) $d -> a$ has support 0.6 and confidence 0.75

- We are not interested in all rules, only rules with a minimum support and minimum confidence

Frequent Itemsets

- Itemset is short form of "set of item"
- Frequent/Large itemset ({A,D}) is a frequent item set in previous example when min support = 0.5
- Find the the desired rules in two steps
  - Find all frequent itemsets via the apriori or FP grow algorithm
    - Every subset of a frequent itemset is frequent alternatively: every superset of an infrequent itemset is infrequent (below minimum support)
  - Generate all the rules within min confidence from the frequent itemsets.

Apriori Algorithm

Exercise in slide 9

	A	B	C	D	E
1	1	1	1	0	0
2	1	1	1	1	1
3	1	0	1	1	0
4	1	0	1	1	1
5	1	1	1	1	0

Count each column:

```
A 5
B 3
C 5
D 4
E 2
```

Combine the ones with same prefix (empty)

```
AB c
AC c
AD c
AE c
bC c
BD c
BE c
CD c
CE c
DE c
```

For each transaction, count the columns

```
AB 3
AC 5
AD 4
AE 2
bC 3
BD 2
BE 1    not frequent!
CD 4
CE 2
DE 2
```

Combine the ones with same prefix (all possible combinations)

```
ABC c
ABD c
ABE not frequent!
ACD c
ACE c
ADE c
BCD c
CDE c
```

For each transaction, increment the counter

```
ABC 3
ABD 2
ACD 4
ACE 2
ADE 2
BCD 2
CDE 2
```

Check for candidates

```
ABCD c
ACDE c
```

Count occurrences

```
ABCD 2
ACDE 2
```

Not sharing prefixes, meaning end of algorithm. The algorithm is pruning the itemsets which are not frequent.

Now we want to generate rules with minimum confidence, use apriori property: if X does not result in a rule with minimum confidence for L, neither does any subset of X.

Do not check brute-force matching, we use the property of above.

```
L = {A,B,C}
```

```
c(AB -> C) < min confidence
c(A->BC) <= c(AB -> C)
```

$L_k = A, B, C$ min conf $= 0.8$

```
A = {AB,AC,BC}
 take c(AB -> C) = sup(ABC) / sup(AB) = 3/3 = 1, output
        A = {A,B} take c(A -> B) = sup(ABC)/sup(A) = 3/5 = 0.6, not output
 take c(B -> AC) = sup(ABC)/sup(B) = 1, output
 take c(AC -> B) = sup(ABC)/sup(AC) = 0.6, not output
 ...
```