**9.1 Data Mining Overview**

Use of algorithms to discover trends in the data. There are 4 steps in data mining:

1. Data preparation: It involves understanding, organizing and transforming the data according to the problem’s needs. It is the most time consuming step.
2. Exploratory data analysis: This is a key step which involves visual analysis to determine the relationships that might exist. The visual cues give insight about the existing relationships between the dependent and independent variable.
3. Model development: This is the most complex step in the process. It involves building the model necessary for determining the relationship.
4. Interpretation of results: This step makes sense of the results of the data model. It determines which data model provides us with actionable insights.

**9.2 Associative Rule Mining**

Associative rules are generally if/then statements. We understand the possibility of a particular situation happening if a given situation has already taken place. This tries to determine relationships between two seemingly unrelated situations.

Three criteria to evaluate the rule:

1. Support rule: It understands how often a rule occurs.  
   Support: total rows of A and B/ total rows
2. Confidence quantity: How frequently the second item occurs where the first item is already present. GIves an idea of how often the rule is correct.  
   Confidence: total rows with A and B/ total rows with A
3. Lift: Confidence/expected confidence.   
   Expected confidence: total rows with B/ total rows

apriori() command in R: Finding rules in the transaction data.

RHS is one item, LHS can be multiple items.

* apriori(dataframe, parameter=list(support= A, confidence= B))  
  This gives the number of rules in the dataframe with the particular support and confidence values. As we lower the support and confidence values more rules will be generated.
* ruleset <- apriori(dataframe, parameter=list(support= A, confidence= B))  
  inspect(ruleset)  
  This enumerates the rules and gives the LHS and RHS and support, confidence and lift of the rule.
* plot(ruleset)  
  Generates scatterplot of the rules in the ruleset.   
  The Y-axis is confidence, X-axis is support and the color gradient is lift.
* goodrules <- ruleset[quality(ruleset)$lift>3.5]  
  This creates a subset of the rules with desired parameter with a certain value.

**Questions from the videos**

* How is this different from the overall data science process?
* The overall data science process does not necessarily involve relationships stated in a definitive way. For example, with data science, images can be classified by a model, however, one picture of a particular object, a tree for instance, can not be definitively stated to have a relation with another image of a tree. The associative rule mining however is more definitive.
* What are some examples of association in rule mining?
* When purchasing tickets for a movie online, at the checkout the offers for popcorn and soda are displayed which pushes the customer in the direction of buying food items with the movie ticket. The association of movies with popcorn and soda is made in this case.
* What might be the possible issues with using this algorithm?
* The association could sometimes be happening in one data sample due to a particular reason or can be seasonal, it can be easily mistaken for a rule, whereas it might be because of chance.

**Questions**

Which parameter is used in different contexts to judge which rule is better? Is any parameter usually preferred over the other two?