**WEEK 2**

1. In the Video Lecture #1 this week, I asked a question: Given a dataset, under what conditions would you apply Data Mining algorithm(s) as a rule of thumb to solve the problem(s)? List three conditions. Can you think of any other additional conditions? If so, why?
2. In the Association Rule Mining, how can you justify a rule (e.g., Milk -> Apple) to be not trivial?
3. When do you apply the Clustering analysis? Can you solve a prediction problem using the Clustering? Justify your answer to either Yes or No.
4. Discuss two activities of your own for one supervised and another unsupervised data mining tasks. Each activity description doesn’t have to be extensive. Refer to the events listed in Exercise 1.7 of our textbook (Tan et al. 2nd Ed.) for example, when you write your own activities.

Rule of Thumb to apply Data Mining:  
1.If the given data set is huge and there are all sorts of information present in the given data.

2.Existence of hidden patterns in the data.

3. If there is no exact solution.

Additional Conditions:  
Competition:  
Lot of businesses are competing with each other to thrive and thus there is a race of who can provide a better customer experience to ensure their customers constantly choose them for their services.  
  
Market penetration:  
To make a name for yourself in the present market with already established businesses .  
Data mining can help provide the current marketing conditions , past/future patterns to help the businesses advertise themselves better and determine how they can :  
a.reach the maximum audience  
b.what can they do better so the audience seeks them out.  
c.what different ways can they become customer friendly.  
  
Also mining the data gives us an opportunity to compare the past and the present patterns in the business growth thus predicting the future patterns and develop the right Business strategy for maximum efficiency  
  
2.  
Association Rule Mining:  
This rule is also known as Market Basket Transactions. Association rule mining basically deals with associating certain items together. What this means is that most of the customers who buy Milk along with other items also buy Apple. This discovery among the data suggests a strong relationship between Milk and Apple  
To justify the rule as being non-trivial we need to find the strength of the rule which can be computed by evaluating the degree of of their relationship.  
The formula: A ∩ B = Φ  
The measurement is observed in terms of Support: How frequent the rule is applicable in the datasets  
and  
Confidence : The number of times item B appears when item A is purchased

The value of Support determines the strength of the rule , hence if this value is low it points to a weak association and the value of Confidence determines how dependable this conclusion is. Thus higher the value of the Confidence , the more it states the presence of item B alongside item A in datasets.   
  
3.  
Clustering analysis:  
Clustering analysis is useful in discovering structures among the data. It helps us segment similar cases together to form groups if unknown initially.  
Yes we can solve a prediction problem using Clustering Analysis. For example using clustering analysis we can predict the disorders/diseases that are caused by certain genes  
How this works ?  
The genes responsible for causing disorder/disease are observed to have an accelerated protien interaction using PPIs(protein-protein interaction)  
.Depending upon the type of clustering algorithm used to decompose the PPIs, the calculated value with a higher score indicates the presence of the disease in the cluster.  
Thus if a candidate wants to know which health problems he/she can have in future so that he starts taking certain precautions in his present,the clustering analysis can find the disease-related gene cluster to predict the likelihood of the disease.

4.  
Unsupervised learning  
Basic gene clustering can be identified under the unsupervised learning technique. Thousands of correlated gene clusters can be clustered in a hierarchical fashion exhibiting similar behavior. Lets say we need to classify tumor tissue. To achieve this we need to identify the functional group present in these genes to correspond to the tissue.  
However this hierarchical structure poses a difficulty in identifying the functional groups present because the genes were clustered together by similarity among them and the variables were not mapped according to the output we wanted.  
  
Supervised Learning  
The above mentioned shortcoming can be overcome by using a supervised learning method where the input variable X can be used to determine the output variable Y i.e the tumor type tissue.  
Y=f(X)  
Where X= mean expression profile of each cluster  
and thus the group of genes containing the most useful information about the functional groups present to discriminate the tumor type tissue is hence identified.

less