Jeonghyeon Woo

Cpts 415

HW 5

1. a.

   Data mining has following steps: selection, processing, transformation, data mining, and interpretation.

   Selection is about what to study. It is specifying the type of data. Collecting stones is an example.

   Processing is storing and organizing of raw data. Putting the stones in boxes can be an example.

   Transformation is about making the organized data accessible. Naming the stone boxes and organizing the boxes can be an example.

   Data mining is about making data model and finding patterns. It is processed by mining functionalities. Making a manual to find diamond box from the stone boxes is an example.

   Interpretation is getting valuable data. Getting diamond from the diamond box is an example.

   b.

   Classification is classifying the type of data. Classifying the types of stones in the stone boxes is an example.

   Regression is mapping the tendency of data. Recording the weights and sizes of the stones is an example.

   Clustering is grouping similar data together into clusters. Putting jewel boxes together is an example.

   Summarization is mapping data into subsets of attributes. Making small diamond box for small size diamond and small diamond box for bigger size diamond is an example.

   Link analysis is defining the relationship among data. Making a manual to find a right stone is an example of link analysis.

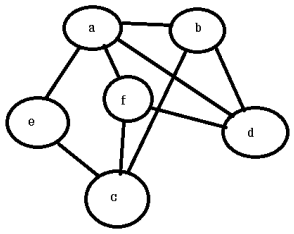
   c.

   Scalability is of the data is a challenge since it is not possible to speed-up by using the properties of small data. Dimensionality is another challenge. Complexity rapidly increases because of the scale.

2. a.

   Graph pattern mining is finding meaningful patterns from a given dataset. With some function and number k, measure how good is each pattern, return a set of patterns that are measured as good.

   b.



   It is revealed that it is the only pattern found when Apriori-based search is done. The graph is closed, which means no edge is opened, so it is not possible to join the graphs in the other ways.

   c.

   Assume a graph G is given.

   When |G|=1. n of subgraph is 1, which is itself.

   When |G|+1

   n subgraph (|G|+1) = n of subgraph G + n(G)

   = n of subgraph G + 1

   is surely frequent since n of subgraph G + 1 > P

   Thus, it is true by induction

3. a.

   There are two steps in classification: model construction and model usage.

   Model construction is classifying stored data. Initial class is set in this step. Model usage is using the existing class to classify a new data piece. New classes may be created in this step.

   For example, in the decision tree model of b, model construction is determining if the weather on the list is good to play golf and model usage is determining if a new name on the list is good to play golf.

   b.

   E(humidity)=-9/14ln(9/14)/ln(2)-5/14ln(5/14)/ln(2) = 0.94

E(wind)=-9/14ln(9/14)/ln(2)-5/14ln(5/14)/ln(2) = 0.94

E(high)=0.985

E(normal)=0.592

E(strong)=1

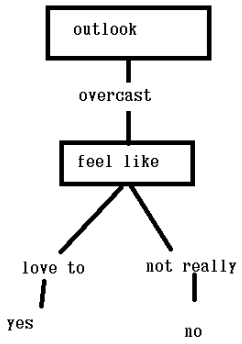
E(weak)=0.811

G(humidity)=0.151

G(wind)=0.048

The table used in this problem has the same information Gains as the one is used in the class. In the class, we concluded that humidity is better, so it is not possible to show that wind is a better choice.

c. new tuple: (overcast, hot, high, strong, not really)



4. a.

K=3

1) A(2,10) C(2,5) B(8,4) B(5,8) B(7,5) B(6, 4) C(1, 2) B(4, 9)

Acent=(2, 10) Bcent=(6, 6) Ccent(1.5, 3.5)

2) A1(2,10) B1(8,4) B2(5,8) B3(7,5) B4(6, 4) B5(4, 9) C1(1, 2) C2(2,5)

b.

K mean has both the strength and weakness of mean while k medoid has both the strength and weakness of median. Mean is sensitive to outliers while median is not distorted by outliers. On the other hand, mean can tolerate data addition while median can not.

c.

DBC needs two parameters: eps(maximum radius of neighborhood) and minpts(minimum number of points in the neighborhood). When reachability becomes

bigger, the shape becomes a plane, and when reachability becomes smaller, the shape becomes a point. DBC is sensitive to parameters. It may not work well when density is too high or too low, or the density is uniform.