CPTS 415 Big Data

Assignment 5

Instructor: Yinghui Wu

---

1. [Data mining functionalities] [15] In the class we have introduced fundamental data mining tasks. Answer the following questions with your own opinions.

   a. Describe the steps involved in data mining viewed as a process of knowledge discovery. Give an example using real-world application to illustrate each step.

   b. Define each of the following data mining functionalities: classification, regression, clustering, association rules, and link analysis. Give examples of each data mining functionality using a real-life data set that you are familiar with (different from what we introduced in the class).

   c. What are the major challenges of mining big data in comparison with mining small-scale dataset (e.g., data set of a few hundred tuples)?

2. [Graph mining] [25] We have introduced graph (pattern) mining as a fundamental mining problem in graph data. Answer the following questions.

   a. Give the definition of graph pattern mining problem.

   b. Consider a graph database in Fig. 1. Draw a closed frequent pattern, when the support threshold =3. Describe how it can be discovered following Apriori-based search.

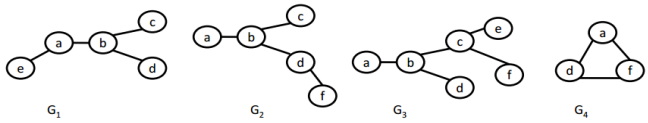
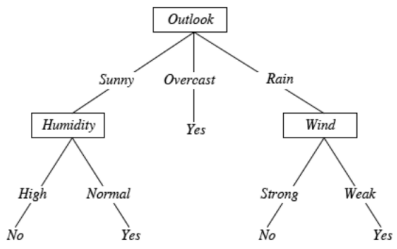


Figure 1: a graph database

   c. Given a graph database D as a set of small graphs, the *Apriori* property of support means if a pattern is frequent in D, then all of its subgraphs are frequent. Consider a graph mining problem defined over a single graph G, where the support of a pattern P refers to the number of all the subgraphs in G that are isomorphic to P. Does the *Apriori* property still holds? If yes, give a proof. If not, give a counter example.

3. [Classification – decision trees] [30]
   a. Briefly outline the major steps of classification algorithm using decision tree model.
   b. Consider the following decision tree learned from a corresponding data set. The decision tree predicates the value of the Boolean attribute PlayGolf. Show that the Wind attribute at the second level of the tree is a better choice than other attributes. [*hint: show its information gain is superior to other choices*].



Day	Outlook	Temperature	Humidity	Wind	PlayGolf
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

c. Add one more tuple to the table, so the tree will contain additional nodes.

4. [Clustering] [30]
   a. We have introduced K-means as a commonly used method for clustering. Suppose the clustering task is to cluster points (with (x,y) representing locations) into three clusters, where the points are: A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9). Define the distance function as Euclidean distance. Suppose initially we assign A1, B1 and C1 as the center of each cluster. Use the k-means clustering to show: (1) the three cluster centers after the first round of execution, and (2) the final three clusters.

   b. Both k-means and k-medoids algorithms can perform effective clustering. Illustrate the strength and weakness of k-means in comparison with k-medoids.

   c. Describe the DBSCAN algorithm, a density-based clustering method, by specifying the following criteria: shapes of clusters that can be determined, input parameters that must be specified, and possible limitations.

   d. * (bonus) A relation R on a set X is a set of pairs (x, y). It is an equivalence relation if for any element a, b, and c in X, (1) (a,a) is in R, (2) if (a,b) in R then (b,a) in R, and (3) if (a,b) in R and (b,c) in R, then (a, c) in R. Review the following two definitions of DBSCAN.
   ci) *In DBSCAN, an object p in a data set D is density-reachable from q w.r.t Eps and MinPts, if there is a chain of objects $p_1,..p_n$, s.t $p_1=q$, $p_n=p$, and p_{i+1} is directly density-reachable from p_i w.r.t Eps and MinPts, for 1<=i<=n, pi in data set D.*
   cii) *Two objects p1, p2 from D are density-connected w.r.t Eps and MinPts, if there is an object q in D s.t. both p1 and p2 are density-reachable from q w.r.t Eps and MinPts*.

   Define the density-connected relation R as a relation where two objects a, b is in R if a is density-connected with b. w.r.t Eps and MinPts. Show that R is an equivalence relation.