Final *Spring 2020*: INFO 437 Data Mining (English)

Duration: 1.5h   Documents not authorized (calculator authorized)

Exercise I (Decision Tree) (20 pts)

A candy manufacturer interviews a customer on his willingness to eat a candy of a particular color or flavor. The following table shows the collected responses:

Color	Flavor	Edibility
Red	Grape	Yes
Red	Cherry	Yes
Green	Grape	Yes
Green	Cherry	No
Blue	Grape	No
Blue	Cherry	No

1. Calculate the Entropy of the dataset given in the table above

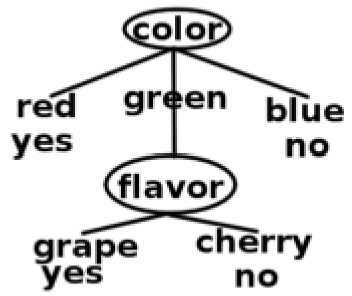
   $-\left(\frac{1}{2}log_2\frac{1}{2}+\frac{1}{2}log_2\frac{1}{2}\right)=1$

2. Which Attribute (Color or Flavor) would you select as the root of the tree using information gain. Show your calculation.

   $-\left(\frac{1}{3}(1log_2 1+0log_2 0)+\frac{1}{3}(\frac{1}{2}log_2\frac{1}{2}+\frac{1}{2}log_2\frac{1}{2})+\frac{1}{3}(1log_2 1+0log_2 0)\right)=\frac{1}{3}$

   `>From inspection, H(edibility | color) < H(edibility | flavor), so color has the larger mutual information with edibility.`

3. Draw the full decision tree for predicting edibility that maximizes the information gain.



4. Using your decision tree, what would you predict for the edibility of a blue, blueberry- flavored candy?

   Edibility: No

Exercise II (Hierarchical Agglomerative Clustering) (20 pts)

I. For the next ***four*** questions, consider a dataset containing six one-dimensional points: {2, 4, 7, 8, 12, 14}. After three iterations of **Hierarchical Agglomerative Clustering** using Euclidean distance between points, we get the 3 clusters: C_1 = {2, 4}, C_2 = {7, 8} and C_3 = {12, 14}.

1. Calculate the distances between clusters C_1 and C_2, C_2 and C_3 using **Single Linkage**?
   `d({2,4}, {7,8}) = 7 - 4 = 3 d({7,8},{12,14}) =12 - 8 = 4`
2. Calculate the distances between clusters C_1 and C_2, C_2 and C_3 using **Complete Linkage**?
   `d({2,4},{7,8}) = 8 - 2 = 6  d({7,8},{12,14}) =14 - 7 = 7`

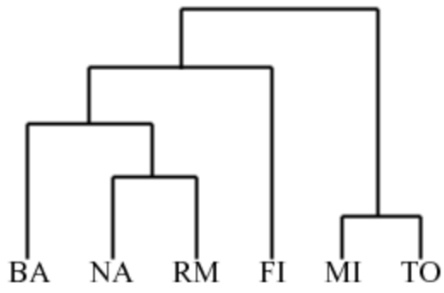
3. What clusters are merged at the next iteration using **Single Linkage**?

d({2,4}, {7,8}) = 7 - 4 = 3, d({2,4}, {12,14}) = 12 - 4 =8, and
d({7,8},{12,14}) = 12 - 8 = 4, so clusters C_1 and C_2 are the closest pair.

4. Draw the dendrogram for the final clusters using **Single Linkage**

II. Consider the **dendrogram**:

Using this dendrogram to create 3 clusters, what would the clusters be? Explain



a. {BA, NA}, {RM, FI}, {MI, TO}
b. {NA, RM}, {BA, FI}, {MI, TO}
c. {BA, NA, RM, FI}, {MI}, {TO}
d. {BA, NA, RM}, {FI}, {MI, TO}
e. None of these

Exercise III (K Means Clustering) (20 pts)

1. You want to cluster 7 points into 3 clusters using the ***k*-Means Clustering** algorithm. Suppose after the first iteration, clusters C1, C2 and C3 contain the following two-dimensional points:

   C1 contains the 2 points: {(0,6), (6,0)} (3,3)
   C2 contains the 3 points: {(2,2), (4,4), (6,6)} (4,4)
   C3 contains the 2 points: {(5,5), (7,7)} (6,6)

   (i) What are the **cluster centers** computed for these 3 clusters?

2. Consider performing *K-Means Clustering* on a one-dimensional dataset containing four data points:
   {5, 7, 10, 12} using k = 2, Euclidean distance, and the initial cluster centers are $c1$ = 3.0 and $c2$ = 13.0.
   (i) What are the initial cluster assignments? (That is, which examples are in cluster $c1$ and which examples are in cluster $c2$?)

Distance	5	7	10	12
c_1 = 3	2	4	7	9
c_2 = 13	8	6	3	1

So, the initial clusters are c_1 = {5, 7} and c_2 = {10, 12}

   (ii) What is the value of *SSE (Sum of Squared Error)* for the clusters computed in (i)?
   Distortion = $2^2 + 4^2 + 3^2 + 1^2 = 30$

   (iii) What are the final cluster centers after running the k-Means Clustering Algorithm? Show your work.
   c_1 = (5 + 7)/2 = 6 and c_2 = (10 + 12)/2 = 11

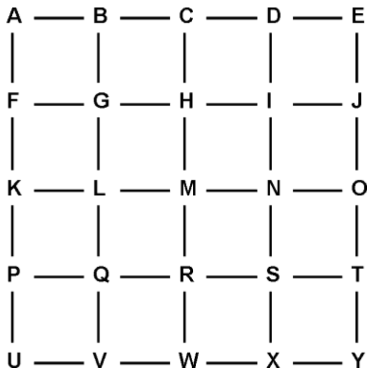
Exercise IV (K Nearest Neighbor) (20 pts)

I. Consider a set of five training examples given as ((x_i, y_i), c_i) values, where x_i and y_i are the two attribute values (positive integers) and c_i is the binary class label:

{((1, 1), −1), ((1, 7), +1), ((3, 3), +1), ((5, 4), −1), ((2, 5), −1)}.
Classify a test example at coordinates (3, 6) using a *k-NN classifier* with k = 3 and using Manhattan distance. Your answer should be either +1 or -1.

+1 because the Manhattan distances from (3,6) to each training example are: (1,1) is distance 2+5=7; (1,7) is distance 2+1=3; (3,3) is distance 0+3=3; (5,4) is distance 2+2=4; and (2,5) is distance 1+1=2. So, points ((2,5), -1), ((1,7), +1) and ((3,3), +1) are the 3 nearest neighbors and their majority class is +1, so classify (3,6) as class +1

II. In the following diagram, *A* through *Y* are all data points. Each data point has features (x,y) corresponding to its coordinate in the grid.



What are the *k* = 5 nearest neighbors of data point *M* using Euclidean distance? Break any ties with alphabetical ordering.
G,H,L,N,R

1. If data points *A* through *L* belong to class 1 and data points *N* through *Y* belong to class 2, what is the classification of *M* when using the *k* = 5 nearest neighbors?

Class 1

Exercise V (Association Rules) (20 pts)

Consider a database as shown below. Suppose each transaction is considered as a set of items; that is, there is no precedence relation imposed on the items in each transaction. Let minimum support be 2

ID	Items
t_1	A,B,C,D
t_2	A,B,C,F
t_3	A,C F
t_4	B,C,D

1. Trace the frequent-set mining process using the *Apriori* algorithm. Show your work.
   A:3 B:3 C:4 D:2 F: 2
   1 item: all
   2items: {A,B}:2 {A,C}:3 {A,D}:1 {A,F}:2 {B,C}:3 {B,D}:2 {B,F}:1 {C,D}:2 {C,F}:2 {D,F}:0
   3items: {A,B,C}:2 {A,B,F}:1 {A,B,D}:1 {A,C,D}:1 {A,C,F}:2 {B,C,D}:2 {B,C,F}:1 {C,D,F}: 0
   4items: {A,B,C,D}:1 {A,B,C,F}:1
2. List all **maximum frequent** sets.
   an itemset is maximal frequent if none of its immediate supersets is frequent
   Closed itemset: none of its supersets have the same support but there does exist a frequent superset hence it's not maximal.
   Max itemsets: {A,B,C}:2 {A,C,F}:2 {B,C,D}:2
3. List one association rule with the highest confidence level.

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{o(X)}$$

A → C conf =3/3=1