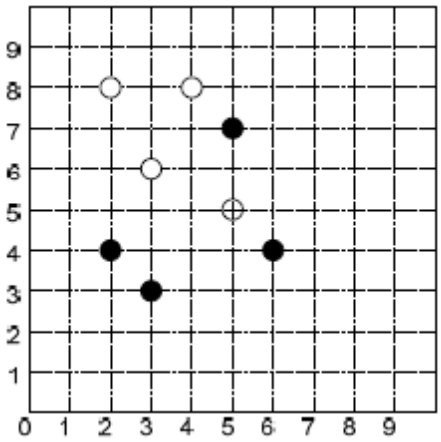
Assignment 6

Note: Show all your work.

Problem 1. The k-means algorithm is being run on a small dataset and. After a certain number of iterations, we have two clusters as shown in the figure. Here, clear circles are Cluster1 objects and filled circles are Cluster2 objects.



Cluster1: (2,8) (3,6) (4,8) (5,5)
Cluster2: (2,4) (3,3) (5,7) (6,4)

Run two more iterations of the k-Means clustering algorithm and show the two clusters at the end of each iteration. You don't need to draw figures like above. It is sufficient that you indicate which objects belong to each cluster at the end of each iteration. Again, show all your work and use Manhattan distance when calculating distances. Note that this is not the beginning of the running of k-means. You are in the middle of the running of kmeans. So, the first thing you need to do is to compute new centroids of two clusters.

Answer:

Cluster0: A(2,8), B(3,6), C(4,8), D(5,5) Cluster1: E(2,4), F(3,3), G(5,7), H(6,4)
We assume C_0 = A(2,8), C_1 = F(3,3)

The distance between object B and C_0 is 1 + 2 = 3 and B and C_1 is 0 + 3 = 3 then B belong C_0

The distance between object C and C_0 is 2 + 0 = 2 and C and C_1 is 1 + 5 = 6 then C belong C_0

The distance between object D and C_0 is 3 + 3 = 6 and D and C_1 is 2 + 2 = 4 then D belong C_1

The distance between object E and C_0 is 0 + 4 = 4 and E and C_1 is 1 + 1 = 2 then E belong C_1

The distance between object G and C_0 is 3 + 1 = 4 and G and C_1 is 2 + 4 = 6 then G belong C_0

The distance between object H and C_0 is 4 + 4 = 8 and H and C_1 is 3 + 1 = 4 then H belong C_1

Now we have cluster C_0 = {B, C, G} and cluster C_1 = {D, E, H}

Next, a new centroid is computed for each cluster.

For Cluster-0, *x* component of C_0 = (3+4+5)/3 = 4 and, *y* component of C_0 = (6+8+7)/3 = 7

Cluster-1, *x* component of C_1 = (5+2+6)/3 = 4.34 and, *y* component of C_1 = (5+4+4)/3 = 4.34

New C_0 = (4, 7) and C_1 = (4.34, 4.34) Then, the same process is repeated for all objects.

A(2,8), B(3,6), C(4,8), D(5,5), E(2,4), F(3,3), G(5,7), H(6,4)

The distance between object A and C0 is 2 + 1 = 3 and A and C1 is 2.34 + 3.66 = 6 then A belong C0

The distance between object B and C0 is 1 + 1 = 2 and B and C1 is 1.34 + 1.66 = 3 then B belong C0

The distance between object C and C0 is 0 + 1 = 1 and C and C1 is 0.34 + 3.66 = 4 then C belong C0

The distance between object D and C0 is 1 + 2 = 3 and D and C1 is 0.66 + 0.66 = 1.32 then D belong C1

The distance between object E and C0 is 2 + 3 = 5 and E and C1 is 2.34 + 0.34 = 2.68 then E belong C1

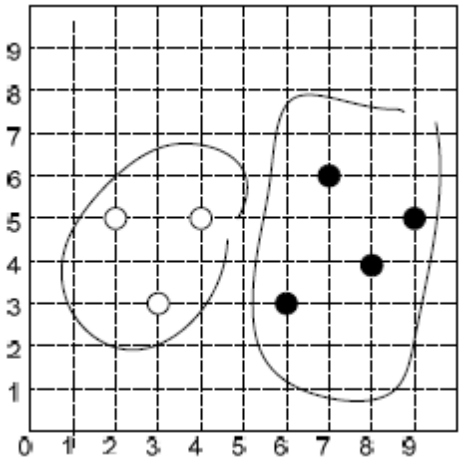
The distance between object F and C0 is 1 + 4 = 5 and F and C1 is 1.34 + 1.34 = 2.68 then F belong C1

The distance between object G and C0 is 1 + 0 = 1 and G and C1 is 0.66 + 2.66 = 3.32 then G belong C0

The distance between object H and C0 is 2 + 3 = 5 and H and C1 is 1.66+ 0.34 = 2 then H belong C1

Finally we have cluster C_0 = {A, B, C, G} and cluster C_1 = {D, E, F, H}

Problem 2. Consider the following two clusters:



Compute the distance between the two clusters (1) using minimum distance and (2) using

average distance. These distance measures are defined in page 461 of the textbook.

Answer:

C_0: a(2,5), b(3,3), c(4,5) C_1: d(6,3), e(7,6), f(8,4), g(9,5)

d(a,d) = 6, d(a,e) = 6, d(a,f) = 7, d(a,g) = 7

d(b,d) = 3, d(b,e) = 7, d(b,f) = 6, d(b,g) = 8

d(c,d) = 4, d(c,e) = 4, d(c,f) = 5, d(c,g) = 5

(1) Minimum distance: d(b,d) = 3

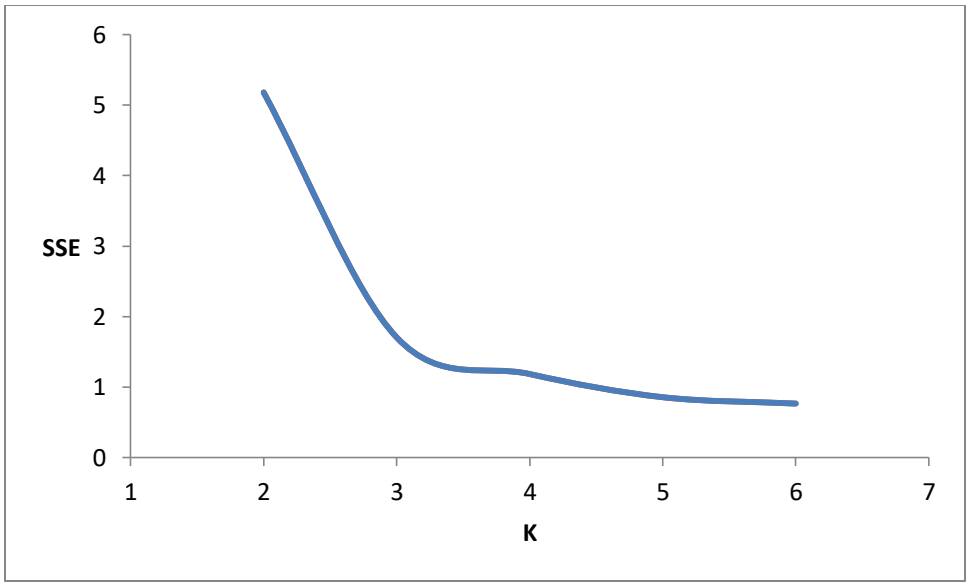
(2) Average distance: (6+6+7+7+3+7+6+8+4+4+5)/12 = 5.25

Problem 3 (25 points). Use the provided *a6-p3.arff* dataset for this problem. It has 150 instances and 2 attributes. For this problem, you need to learn how to use the *SimpleKMeans* algorithm on Weka.

Problem 3-1 Run the *SimpleKMeans* algorithm on Weka on this dataset with *k* = 2, 3, 4, 5, and 6. For each *k*, record the value of *within cluster sum of squared errors* (which you can find in Weka's cluster output window) and plot a graph where the x-axis is *k* and yaxis is *within cluster sum of squared errors*. Then, determine an optimal number of clusters using the *elbow method* which is described in page 486 of the textbook (it is also described in Lecture 6 slides).

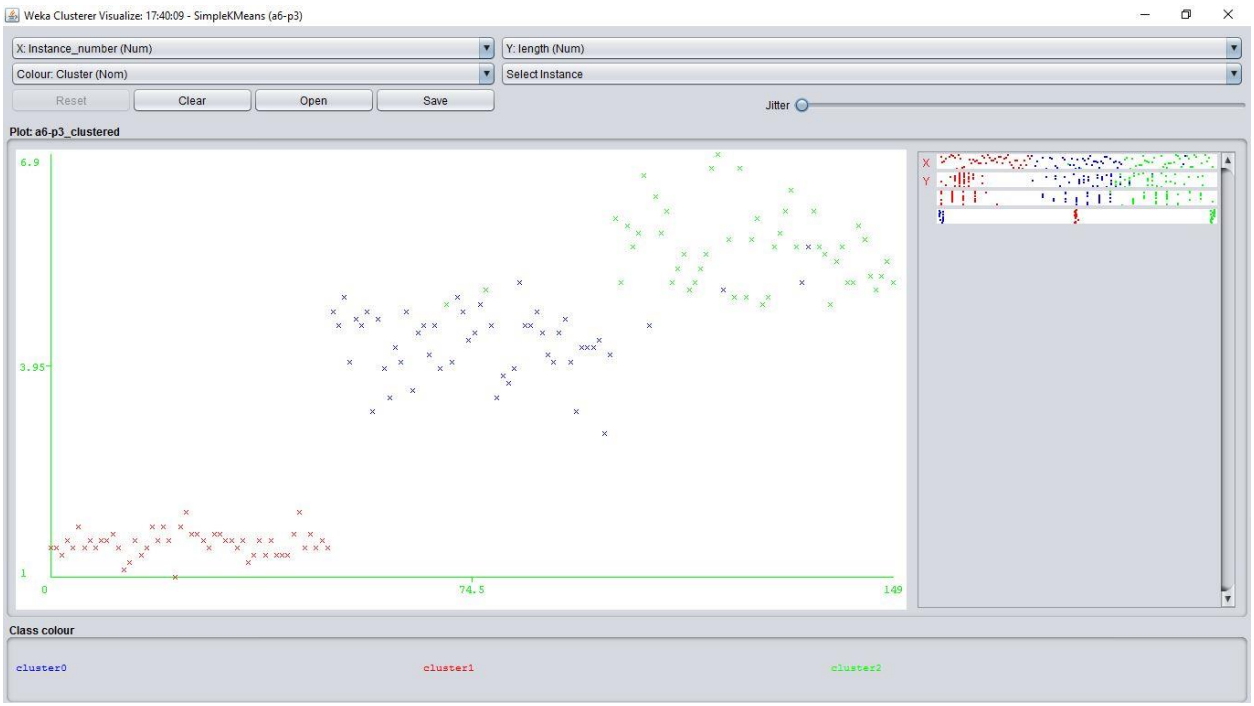
Answer:

K	SSE
2	5.179687
3	1.705098
4	1.185911
5	0.857314
6	0.766619



By considering above chart, we can conclude that an optimal number of clusters is K = 3

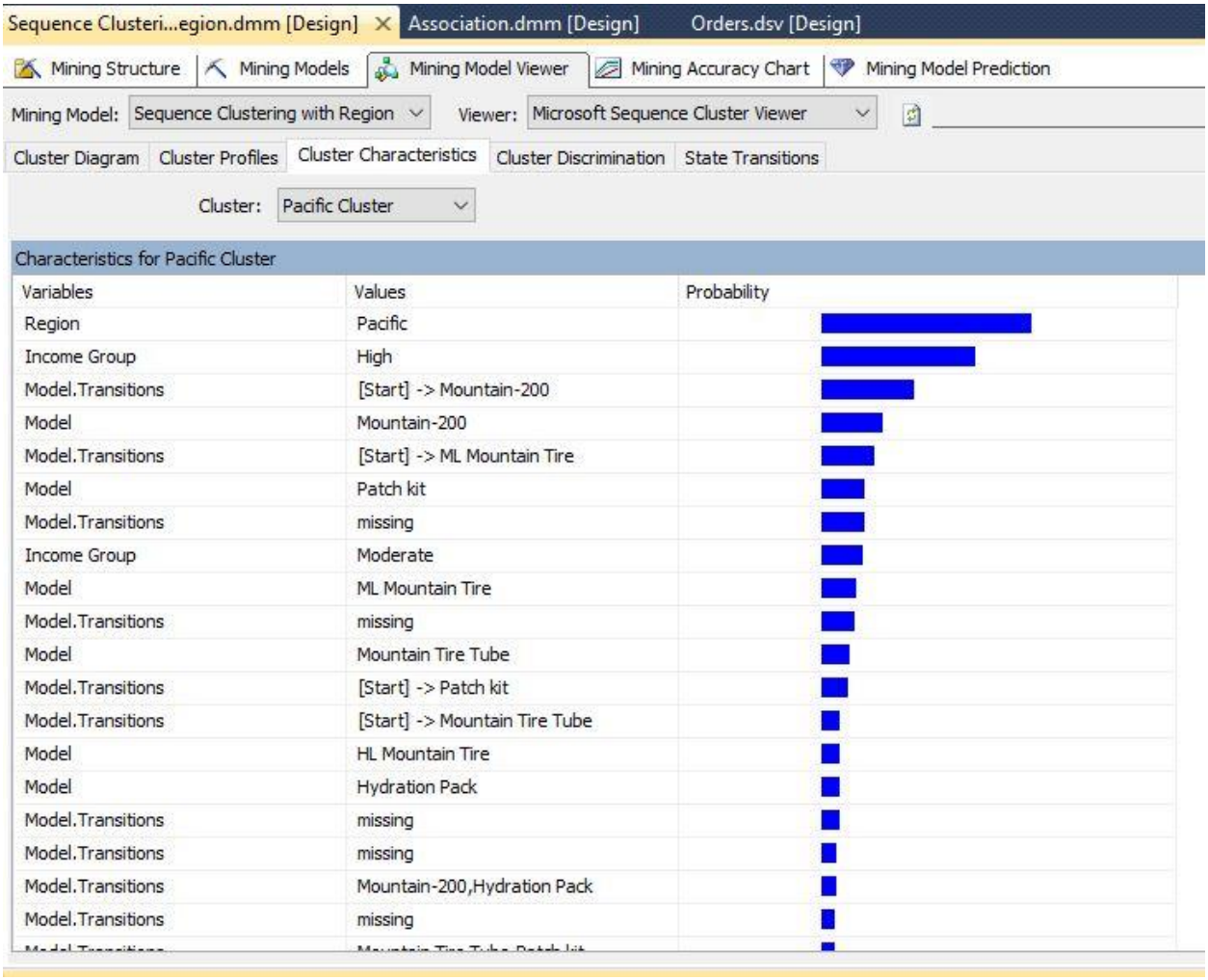
Problem 3-2 Using the optimal number of clusters which you determined in Problem 3-1, run *SimpleKMeans* again and characterize the generated clusters using the two attribute values. For example, if two attributes were age and income, characterization of clusters would look like:
Cluster 0: Mostly younger than 21 and income between 15K and 35K
Cluster 1: Mostly ages between 21 and 45 and income between 35K and 90K
…
After you run a clustering algorithm, you can save the clustering result as follows:
1. Right-click the entry of your clustering in the "Result list."
2. Select "Visualize cluster assignments."
3. Click the "save" button.



Problem 4-1 (SQL Server).
This problem is a practice of *sequence clustering* using SQL Server 2012. Lesson 4 of *Intermediate Data Mining tutorial* illustrates how to build and analyze a sequence clustering model. The link is:
http://msdn.microsoft.com/en-us/library/ms167594(v=sql.110).aspx
It has five tasks. They are:
1. Creating a Sequence Clustering Mining Model Structure
2. Processing the Sequence Clustering Model
3. Exploring the Sequence Clustering Model
4. Creating a Related Sequence Clustering Model.
5. Creating Predictions on a Sequence Clustering Model.

Among these, you are required to perform the first four tasks, except the section titled *Generic Content Tree Viewer* which is in the 3rd task.
The tutorial explains the basic concept of Microsoft sequence clustering. If you need more information about their sequence clustering, refer to appropriate Microsoft documentation (links to some documents are included in the tutorial).
Requirements: At the beginning of the 3rd task (*Exploring the Sequence Clustering Model*), you will rename two clusters as *Pacific Cluster* and *Largest Cluster*. After that, you will explore your model using various tabs under *Mining Model Viewer*.
Problem 4-1-1 After you explore your mining model in the 3rd task (*Exploring the Sequence Clustering Model*), click *Cluster Characteristics* tab and choose *Pacific Cluster* from *Cluster* drop down menu.

(1) Capture the screen and paste it onto your submission.



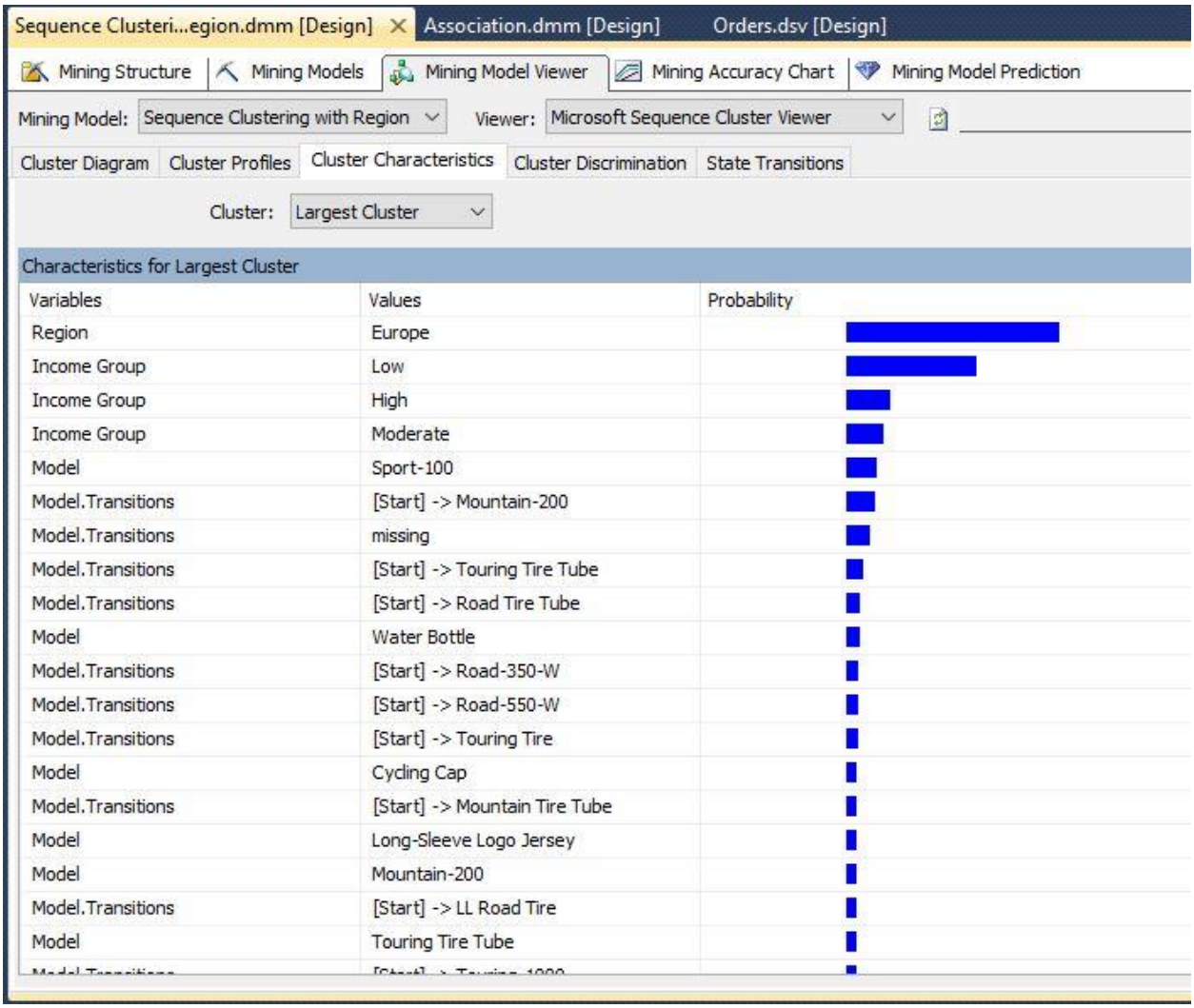
(2) List the top three models that are put in customer's shopping basket as the first model (or item) most frequently in the decreasing order of probability.

Mountain-200, Patch Kit, ML Mountain Tire

(3) Choose the 2-model sequence that is most frequently put in customer's shopping basket.

Mountain-200, Hydration Pack

Problem 4-1-2 This time select *Largest Cluster* from *Cluster* drop down menu.
(1) Capture the screen and paste it onto your submission.



(2) List the top three models that are put in customer's shopping basket as the first model (or item) most frequently in the decreasing order of probability.

Sport-100, water bottle, cycling cap

(3) Choose the 2-model sequence that is most frequently put in customer's shopping basket.

Water bottle, Mountain bottle cage

Problem 4-1-3 After you create a *related sequence clustering model* in the 4th task

(*Creating a Related Sequence Clustering Model*) in the *Mining Model Viewer*, click *State Transitions* tab. To the left of the window, there is a slider. Raise the slider all the way up to *All Links*. Capture the screen and paste it onto your submission