CS 429 - Information Retrieval
Homework 6
Shiva Sankar Modala(A20517528)

1 Recitation Exercises

These exercises are to be found in: Introduction to Information Retrieval 1 st Edition (Online) by Christopher Manning, Prabhakar Raghavan, Hinrich Sch¨utze.

1.1 Chapter 16

Exercise 16.3

Replace every point d in Figure 16.4 with two identical copies of d in the same class. (i) Is it less difficult, equally difficult or more difficult to cluster this set of 34 points as opposed to the 17 points in Figure 16.4? (ii) Compute purity, NMI, RI, and F5 for the clustering with 34 points. Which measures increase and which stay the same after doubling the number of points? (iii) Given your assessment in (i) and the results in (ii), which measures are best suited to compare the quality of the two clusterings?

Ans:

(i) It doesn't make it more or less difficult. I'm interpreting this question as the points being equal copies placed one on top of the other. The clusters should remain the same in that case. Different shapes and arrangements of points can impact clustering difficulty but this situation wouldn't seem to change much.

(ii)

Purity = (1/34) * ((5*2) + (4*2) + (3*2)) = 0.70588

Cluster 1 -> 12 points, 10 x's, 2 o's

Cluster 2 -> 12 points, 2 x's, 8 o's, 2 diamonds

Cluster 3 -> 10 points, 4 x's, 6 diamonds

NMI = I(Ω; C) / ([H(Ω) + H(C)]/2) =

I(Ω; C) -> summation k, summation j (probability of doc being in cluster k and class j) / number of docs * log (probability of doc being in cluster k and class j) / (number of docs in cluster k, number of docs in class j)

Summation of:

Cluster 1 x -> ([(12/34)*(10/18)] / 34) * (log ([(12/34)*(10/18)] / (12*16))

Cluster 1 o -> ([(12/34)*(2/10)] / 34) * (log ([(12/34)*(2/10)] / (12*10))

Cluster 1 diamond -> ([(12/34)*(0/8)] / 34) * (log ([(12/34)*(0/8)] / (12*8))

Cluster 2 x -> ([(12/34)*(2/18)] / 34) * (log ([(12/34)*(2/18)] / (12*16))

Cluster 2 o -> ([(12/34)*(8/10)] / 34) * (log ([(12/34)*(8/10)] / (12*10))

Cluster 2 diamond -> ([(12/34)*(2/8)] / 34) * (log ([(12/34)*(2/8)] / (12*8))

Cluster 3 x -> ([(10/34)*(4/18)] / 34) * (log ([(12/34)*(4/18)] / (12*16))

Cluster 3 o -> ([(10/34)*(0/10)] / 34) * (log ([(10/34)*(0/10)] / (12*16))

Cluster 3 diamond -> ([(10/34)*(6/8)] / 34) * (log ([(10/34)*(6/8)] / (12*16))

H(Ω) -> - (12/34log(12/34) + 12/34log(12/34) + 10/34log(10/34))

H(C) -> -(16/34log(16/34) + 10/34log(10/34) + 8/34log(8/34))

RI = (TP + TN) / (TP + FP + FN + TN)

TP + FP = (12 choose 2) + (12 choose 2) + (10 choose 2) = 177

TP = (10 choose 2) + (2 choose 2) + (8 choose 2) + (2 choose 2) + (2 choose 2) + (6 choose 2) + (4 choose 2) = 97
FP = 117 - 97 = 80
Now take pairs in diff clusters that should have been in same
FN = (10*2) + (10*4) + (4*2) + (2*8) + (2*6) = 96
FN+TN = (34 choose 2) - (TP+FP) = 384
TN = 348 - 96 = 288
RI = (97 + TN) / (97 + 80 + FN + TN)
RI = (97 + 288) / (97 + 80 + 96 + 288) = 0.6862
F5 = ((26)*PR) / (25P + R)
P = TP / (TP + FP) = 97/(97+80) = 0.82906
R = TP / (TP + FN) = 80/(80+96) = 0.45454
F5 = ((26)*(0.82906)(0.45454)) / (25(0.82906) + (0.45454)) = 0.463
Stays same -> Purity
NMI
Increase -> RI, F5

(iii) Use the ones that stay the same, purity and NMI

Exercise 16.4
Why are documents that do not use the same term for the concept car likely to end up in the same cluster in K-means clustering?
Ans: Documents that refer to similar concepts such as "car" but use different terms tend to utilize related vocabulary (e.g., "vehicle," "wheels," "speed"), resulting in K-means clustering. This clustering technique groups documents based on semantic similarities while minimizing Euclidean distances. Despite the differences in words, documents share contextual cues that cause them to cluster together. K-means prefers similarity based on word usage rather than precise words, allowing for more efficient clustering of conceptually related documents.

Exercise 16.5
Two of the possible termination conditions for K-means were (1) assignment does not change, (2) centroids do not change (page 361). Do these two conditions imply each other?
Ans: Indeed, the conditions indicate one another. The points will remain in their current clusters and their assignments won't change if the centroids remain unchanged. If no cluster assignments change, the centroids do not need to be updated as the values remain constant.

Exercise 16.9
Mutual information is symmetric in the sense that its value does not change if the roles of clusters and classes are switched: $I(\Omega; C) = I(C; \Omega)$. Which of the other three evaluation measures are symmetric in this sense?
Ans: Normalized mutual information (NMI) is the only one of the three common clustering assessment methods that demonstrates symmetricity. The values of Purity and the Rand Index are not symmetric; they can change based on the definition or switching of clusters and classes.

Exercise 16.15
In the last iteration in Table 16.3, document 6 is in cluster 2 even though it was the initial seed for cluster 1. Why does the document change membership?
Ans: At start -> 6 is centroid for cluster 1, 7 is centroid for 2 Documents 1, 2, 3, 4, and 5 start moving towards cluster 1, and then 10 and 11 move to cluster 2 since they are not similar to the first 5. Document 6 is more similar docs in cluster 2 at this point, so it moves there.

1.2 Chapter 17
Exercise 17.3
For a fixed set of N documents there are up to N2 distinct similarities between clusters in single-link and complete-link clustering. How many distinct cluster similarities are there in GAAC and centroid clustering?
Ans: For GAAC, avg similarity of all document pairs including those from same cluster, exclude self-similarities, so: $2\text{^}N * 2\text{^}N$

Exercise 17.7
a. Consider running 2-means clustering on a collection with documents from two different languages. What result would you expect?
Ans: I would assume this result in two clusters, each corresponding to the two different languages.

b. Would you expect the same result when running an HAC algorithm?
Ans: Single-link: Might mix the two kinds of docs (chaining)
Complete link: One big cluster and a smaller outlier cluster
Centroid: Would expect two clusters, one per language
GAAC: Would expect two clusters, one per language

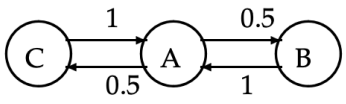
Exercise 17.9
Suppose a run of HAC finds the clustering with K = 7 to have the highest value on some prechosen goodness measure of clustering. Have we found the highest-value clustering among all clusterings with K = 7?
Ans: If it is a single link, then yes. Clustering with K=N clusters has combination similarity 1.0, the largest possible values. A single link clustering with k clusters is optimal (according to ch 17 of textbook, proof given).

Exercise 17.12
For N points, there are $\leq N^K$ different flat clusterings into K clusters (Section 16.2, page 356). What is the number of different hierarchical clusterings (or dendrograms) of N documents? Are there more flat clusterings or more hierarchical clusterings for given K and N?
Ans: $N!$ $(N-1)!$ / $2\text{^}N-1$ there are (n choose 2) ways to choose the first cluster, (n-1 choose 2) ways to choose the second, etc until there are (2 choose 2) at the last choice. Recurrence relation -> a(n) = (n choose 2)*a(n-1), a(1) = 1
For example, if there are 5 points you are clustering,
Reference: https://core.ac.uk/download/pdf/82235912.pdf

Higher number for hierarchical

1.3 Chapter 18
Exercises: N/A
1.4 Chapter 19
Exercises: N/A
1.5 Chapter 20
Exercises: N/A
1.6 Chapter 21
Exercises (Optional): 21.1,21.5,21.6,21.7,21.8,21.19
Exercise 21.1
Is it always possible to follow directed edges (hyperlinks) in the web graph from any node (web page) to any other? Why or why not?
Ans: No. Webpages with no outlinks are the examples of nodes which make the given statement false. Some of the webpages which fall into this category are blogs, new webpages and personal pages.

Exercise 21.5
Write down the transition probability matrix for the example in Figure 21.2.



► **Figure 21.2** A simple Markov chain with three states; the numbers on the links indicate the transition probabilities.

Ans:
Transition probability matrix =

$$\begin{matrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{matrix}$$

Rows 1, 2, 3 and col 1, 2, 3 represent A, B, C

Exercise 21.6
Consider a web graph with three nodes 1, 2 and 3. The links are as follows: 1 → 2, 3 → 2, 2 → 1, 2 → 3. Write down the transition probability matrices for the surfer's walk with teleporting, for the following three values of the teleport probability: (a) $\alpha = 0$; (b) $\alpha = 0.5$ and (c) $\alpha = 1$.
Ans:

a) $P = \begin{matrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \end{matrix}$

b) $P = \begin{matrix} 1/6 & 2/3 & 1/6 \\ 5/12 & 1/6 & 5/12 \\ 1/6 & 2/3 & 1/6 \end{matrix}$

c) P = $\begin{matrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{matrix}$

Exercise 21.7

A user of a browser can, in addition to clicking a hyperlink on the page x he is currently browsing, use the back button to go back to the page from which he arrived at x. Can such a user of back buttons be modeled as a Markov chain? How would we model repeated invocations of the back button?

Ans: No, the user can't be modeled as Markov's Chain. This is because the user may hit the back-button multiply and the semantics should be you unwind the path up to that point - this is not Markovian.

Exercise 21.8

Consider a Markov chain with three states A, B and C, and transition probabilities as follows. From state A, the next state is B with probability 1. From B, the next state is either A with probability p_A, or state C with probability $1 - p_A$ From C the next state is A with probability 1. For what values of $p_A \in [0, 1]$ is this Markov chain ergodic?

Ans: $p_A \in (0, 1)$

Exercise 21.19

If all the hub and authority scores are initialized to 1, what is the hub/authority score of a node after one iteration?

Ans: number of outlinks/inlinks