CS412: Introduction to Data Mining **Summer 2016**

Assignment 5

Due: 8/4/2016 11:59pm

General Instruction

- Errata: After the assignment is released, any further corrections of errors or clarifications will be posted at the Errata page at Piazza. Please watch it.
- Feel free to talk to other members of the class while doing the homework. We are more concerned that you learn how to solve the problem than that you solve it entirely on your own. You should, however, write the solution yourself.
- Please use Piazza first if you have questions about the homework. Also feel free to send us e-mails and come to office hours.
- For each question, you should show the necessary calculation steps and reasoning—not only final results. Keep the solution brief and clear.
- For a good balance of cognitive activities, we label each question with an activity type:
    - **L1 (Knowledge)** Definitions, propositions, basic concepts.
    - **L2 (Practice)** Repeating and practicing algorithms/procedures.
    - **L3 (Application)** Critical thinking to apply, analyze, and assess.

Assignment Submission

- Please submit your work before the due time. **We do NOT accept late submission!**
- Please submit your answers electronically via Compass. Contact CITES/TAs if you have technical difficulties in submitting the assignment.
- Please **type** your answers in an **Answer Document**, and submit it in PDF. **Handwritten answers or hand-drawn pictures are not acceptable**.
- Please **DO NOT** zip the Answer Document (PDF) so that the graders can read it directly on Compass. Compress other files into a single zip file.

1 Conceptual Questions (5 points)

a. (L1, 2) Do you think k-means clustering is sensitive to noise and outliers? If so why?

b. (L1, 1) Is single-link clustering a greedy approach?

c. (L1, 2) List down two pros of using hierarchical clustering over k-means?

2 Advanced Classification: Perceptron (5 points)

Purpose

- Understand how classification using Perceptron method works.

Requirements

- Show your calculations for all the questions.

Consider the data given in table 1

x_1	x_2	**y**
0	0	+
0	1	+
1	0	+
1	1	-

Table 1: Data points with class labels

(L2, 5) Show the steps for the iterations before you halt taking the initial values as $w_0 = 0.25$, $w_1 = 0.25$, $w_2 = 0.25$, and $\eta = 0.5$.

3 Hierarchical Agglomerative Clustering and B-Cubed Evaluation (8 points)

Purpose

- Understand how AGNES and B-Cubed work.

Requirements

- In sub-question a, only draw the dendrogram.
- In sub-question b, only list the members of each cluster.
- In sub-question c, show detailed calculations of B-Cubed precision and recall.

Suppose we have 7 two dimensional data points as listed in table 2. The ground truth (the correct clustering) is also provided.

Point	**x**	**y**	**Ground Truth**
P1	1	1	C1
P2	1	2	C1
P3	2	1	C1
P4	5	1	C2
P5	3	2	C1
P6	5	2	C2
P7	3	3	C1

Table 2: Data Points

a. (L2, 4) Draw the dendrogram using AGNES. Please use *single link* and *Euclidean distance* as the dissimilarity measure.

b. (L3, 2) If we want to cluster the dataset into 3 groups based on the dendrogram, what are the members of each of the 3 groups?

c. (L2, 2) Based on the given ground truth, what are the *B-Cubed* precision and recall of the output of part (b)?

4 K-Means (10 points)

Purpose

- Understand how k-Means works.

Requirements

- In sub-question a and b, for each iteration of k-Means:
  - Show the coordinates of the cluster centers in each iteration with the help of calculation.
- In sub-question c, any reasonable method with brief explanation is acceptable.

We will use the same data as table 2 from the previous question.

a. (L2, 4) Perform k-means using Euclidean distance with $k = 2$ and the initial cluster centers $P1$ and $P3$.

b. (L2, 4) Perform k-means using Euclidean distance with $k = 2$ and the initial cluster centers $P2$ and $P6$.

c. (L3, 3) As you can see in sub-questions a and b, choices of initial cluster centers can affect the speed of the clustering process. Assume $k = 2$ suggest a general and reasonable way to select initial cluster centers to speed up the clustering process. Briefly justify your choice.

Machine Problems (50 points)

Requirements

- You can access the MPs for Assignment 5 using this link for Machine Problems.
- MP14 (50 points) is **required** to be solved by all students. The deadline for the required MP is same as the Assignment.
- MP13 and MP15 are *optional* and you will get extra credits for solving them.
- The deadline for optional MPs (MP2 and MP4) is also same as Assignment.
- Each MP webpage consists of instructions, input data and the expected output data.