National University of Singapore
School of Computing
CS3244 Machine Learning

Tutorial 5: Clustering

Issue: March 24, 2022 Due: March 28, 2022

Important Instructions:

- *Your solutions for this tutorial must be TYPE-WRITTEN.*
- *Make TWO copies of your solutions: one for you and one to be SUBMITTED TO THE TUTOR IN CLASS. Your submission in your respective tutorial class will be used to indicate your CLASS ATTENDANCE. Late submission will NOT be entertained.*
- *Indicate your NAME, STUDENT NUMBER, and TUTORIAL GROUP in your submitted solution.*
- *YOUR SOLUTION TO QUESTION* **CL 1(a)(b)** *will be GRADED for this tutorial.*
- *You may discuss the content of the questions with your classmates. But everyone should work out and write up ALL the solutions by yourself.*

CL 1 Consider the following unsupervised dataset. Assume that $0 \leq x_1, x_2 \leq 10$.

ID	(x_1, x_2)
X_1	(1, 2)
X_2	(2, 5)
X_3	(2, 10)
X_4	(4, 9)

ID	(x_1, x_2)
X_5	(5, 8)
X_6	(6, 4)
X_7	(7, 5)
X_8	(8, 4)

(a) Apply the k-means algorithm (using Euclidean distance) on the above dataset using $k = 2$. Use the initial centroids C_1 at $(2, 7)$ and C_2 at $(8, 2)$. In your answer, for each iteration of k-means, list (i) the cluster memberships for each instance, and (ii) the new coordinates of the centroids.

(b) Calculate the total SSE of the resultant clusters formed in Part (a).

(c) Repeat Parts (a) and (b), but this time using $k = 3$, and the initial centroids C_1 at $(2, 7)$, C_2 at $(8, 2)$, and C_3 at $(2, 2)$.

CL 2 Given k equally sized clusters, the probability that a randomly chosen initial centroid will come from any given cluster is $1/k$, but the probability that each cluster will have exactly one initial centroid is much lower. (It should be clear that having one initial centroid in each cluster is a good starting situation for k-means.) In general, if there are k clusters and each cluster has n points, then the probability, p, of selecting in a sample of size k one initial centroid from each cluster is given by:

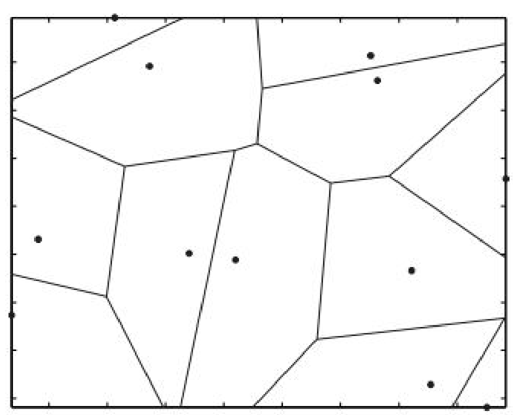
$$p = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } k \text{ centroids}} = \frac{k!n^k}{(kn)^k} = \frac{k!}{k^k}$$

(This assumes sampling with replacement.) From this formula we can calculate, for example, that the chance of having one initial centroid from each of four clusters is $4!/4^4 = 0.0938$.

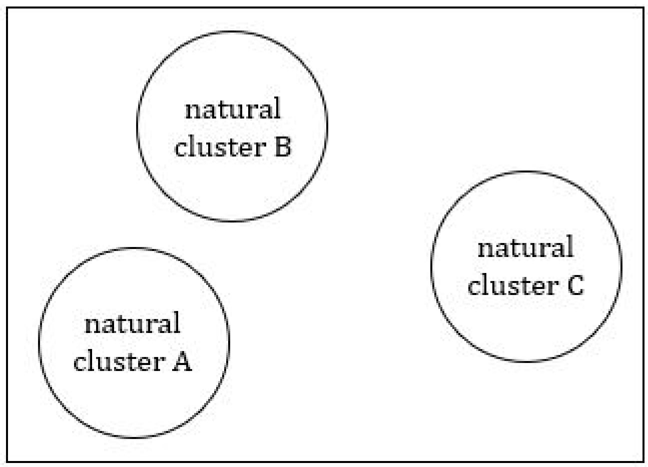
(a) Plot the probability of obtaining one point from each cluster in a sample of size k for values of k between 2 and 100.

(b) For k clusters, $k = 10$, 100, and 1000, estimate the probability that a sample of size $2k$ contains at least one point from each cluster. You can use either mathematical methods or statistical simulation to determine the answer.

CL 3 The Voronoi diagram for a set of k points in the plane is a partition of all the points of the plane into k regions, such that every point (of the plane) is assigned to the closest point among the k specified points. This is depicted in the figure below. What is the relationship between Voronoi diagrams and k-means clusters? What do Voronoi diagrams tell us about the possible shapes of k-means clusters?



CL 4 Suppose that you are given a dataset consisting of instances in a two-dimensional, continuous instance space. Let each instance in this dataset be a member of exactly one of three natural clusters, such that these natural clusters all have the same circular shape and size, and also all contain the same number and distribution of instances. An example of such a dataset is depicted below.



Reposition the natural clusters within the given instance space such that, almost always, (i) the k-means algorithm would find the correct clusters, but (ii) bisecting k-means would not. Also, assume that both algorithms are applied with $k = 3$.