Assignment 8 (Sol.)

Introduction to Data Analytics

Prof. Nandan Sudarsanam & Prof. B. Ravindran

1. For students in a school, if we want to partition the students based on the different extra-curricular activities that they participate in, which clustering approach do you think will be most suitable for this task?

   (a) hierarchical

   (b) overlapping

   (c) partitional

   (d) partial

   Sol. (b)
   As students can be expected to participate in multiple activities, it would make sense to use an overlapping clustering approach, with one cluster for each activity.

2. Given some data set, you are interested in identifying outliers. If you were to use clustering for this task, which among the following approaches would you prefer?

   (a) hierarchical

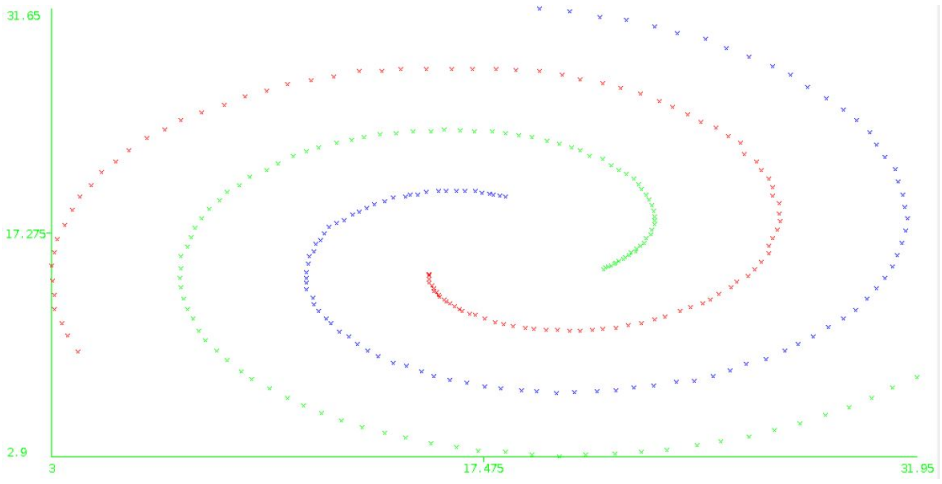
   (b) overlapping

   (c) partitional

   (d) partial

   Sol. (d)
   For this problem, a partial clustering approach would be most suitable, so that all data points which are not outliers are grouped into different clusters (based on their attributes), whereas outlier points are not assigned to any cluster.

3. Consider the following image showing data points belonging to three different clusters (indicated by the colours of the points).



If we run the k-means clustering algorithm with k = 3, do you think the algorithm will be able to correctly cluster the data points belonging to the three clusters?

(a) no

(b) yes

Sol. (a)
Clusters obtained by the k-means algorithm are spherical in shape. It would not be possible to correctly cluster the points shown in the above diagram using the k-means algorithm with the value of k set to 3.

4. Suppose that for the same data set as in the previous question, we use hierarchical clustering. Which approach, single-link, or complete-link would you expect to do better in correctly clustering the data points?

   (a) single-link

   (b) complete-link

   Sol. (a)
   Single-link hierarchical clustering has the ability to cluster data points lying in a chain when the separation between the chains (i.e., between the data points of two separate clusters) is more than the separation between the data points in the chain, as in the above example. The same would not be possible with complete-link clustering.

5. In designing an experiment, we choose to make use of the take-the-best heuristic. However, once we start conducting experiments, we observe that there is significant noise in the output. Can we counter this by sticking with the same heuristic but increasing the number of experiments we conduct for each treatment?

   (a) no

   (b) yes

Sol. (a)
Increasing the number of experiments for each treatment will not allow us to overcome issues with noisy observations, since we still consider as best, the treatment resulting in the best performance, regardless of whether the output of this treatment was affected by noise or not.

6. Given a two-class training data set with 100 unlabelled data points, suppose we randomly select 10 data points and query for their labels. We supply these 10 labelled data points to a SVM, and obtain a decision boundary. Assuming a limit on the number of additional points that we can select to improve this classifier, in general, would you prefer to query the labels of points lying close to the decision surface or those that are far from the decision surface?

   (a) close to the decision surface

   (b) far from the decision surface

   Sol. (a)
   This is an example of the active learning approach. Given what we know about SVMs, we would, in general, prefer to query points lying close to the decision surface, as taking such points into consideration can lead to a change in the decision surface.

7. Would you characterise multi-arm bandit problems under the supervised learning or unsupervised learning category of problems?

   (a) supervised learning

   (b) unsupervised learning

   Sol. (a)
   While we study multi-arm bandit problems under the reinforcement learning paradigm, it should be clear that these are supervised learning problems since supervision in terms of rewards are available to the learning agent.

8. Suppose you used an algorithm based on the PAC framework (with parameters ϵ, δ) to solve a given bandit problem. In the next iteration of the bandit problem, you pick the arm suggested by the algorithm. With what probability is this arm the optimal arm (assuming that the given bandit problem has one optimal action)?

   (a) $P(optimal\ arm) < \epsilon$

   (b) $P(optimal\ arm) < 1 - \epsilon$

   (c) $P(optimal\ arm) < \delta$

   (d) $P(optimal\ arm) < 1 - \delta$

   Sol. (d)
   Recall that a (ϵ, δ)-PAC algorithm will return a solution that is ϵ-close to the optimal solution with probability greater than or equal to $(1 - \delta)$. Hence, the probability that an optimal solution is returned will be less than this probability value.

9. Suppose we are trying to solve a multi-arm bandit problem where there is one optimal arm. We apply the median elimination algorithm to solve this problem. Is it possible that the optimal arm is eliminated in the first round?

(a) no

(b) yes

Sol. (b)
Recall that in the median elimination algorithm, in each iteration we eliminate half of the arms having low estimated value. Given that rewards can be stochastic, it is possible that the observed rewards of the optimal arm in the first round lead to an estimated value falling in the bottom half, resulting in the optimal arm being eliminated.

10. (2) After 12 iterations of the UCB algorithm applied on a 4-arm bandit problem, we have $n_1 = 3$, $n_2 = 4$, $n_3 = 3$, $n_4 = 2$ and $\bar{x}_1 = 0.55$, $\bar{x}_2 = 0.63$, $\bar{x}_3 = 0.61$, $\bar{x}_4 = 0.40$. Which arm should be played next?

(a) 1

(b) 2

(c) 3

(d) 4

Sol. (d)
The next action, A_{13}, will be the action with the maximum upper confidence bound among the four arms. Calculating these values, we have

$$\bar{x}_1 + \sqrt{\frac{2ln12}{n_1}} = 0.55 + \sqrt{\frac{2ln12}{3}} = 1.837$$

$$\bar{x}_2 + \sqrt{\frac{2ln12}{n_2}} = 0.63 + \sqrt{\frac{2ln12}{4}} = 1.745$$

$$\bar{x}_3 + \sqrt{\frac{2ln12}{n_3}} = 0.61 + \sqrt{\frac{2ln12}{3}} = 1.897$$

$$\bar{x}_4 + \sqrt{\frac{2ln12}{n_4}} = 0.40 + \sqrt{\frac{2ln12}{2}} = 1.976$$

Clearly, arm 4 has the highest upper confidence bound and hence will be selected by the UCB 1 algorithm.