Assignment 8

Introduction to Data Analytics

Prof. Nandan Sudarsanam & Prof. B. Ravindran

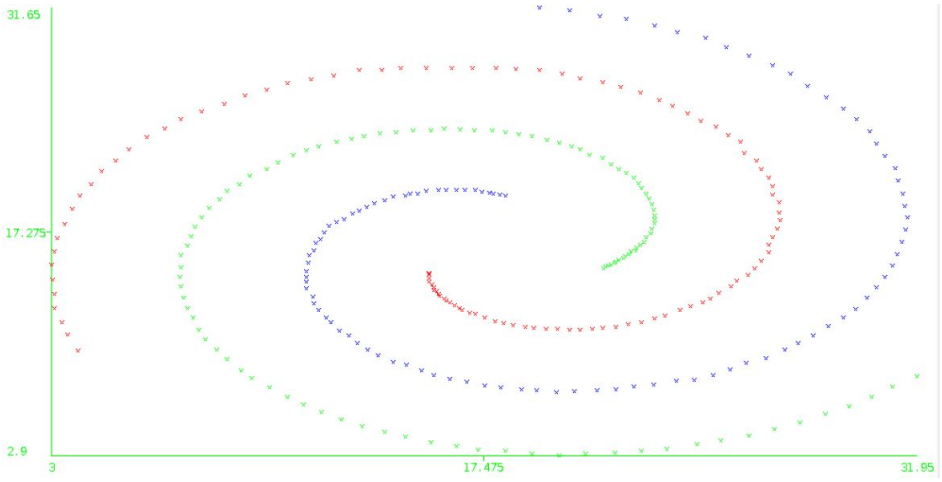
1. For students in a school, if we want to partition the students based on the different extra-curricular activities that they participate in, which clustering approach do you think will be most suitable for this task?

   (a) hierarchical
   (b) overlapping
   (c) partitional
   (d) partial

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

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

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

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

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

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

   (a) supervised learning

   (b) unsupervised learning

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

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

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