Problem - 3

Question - 1

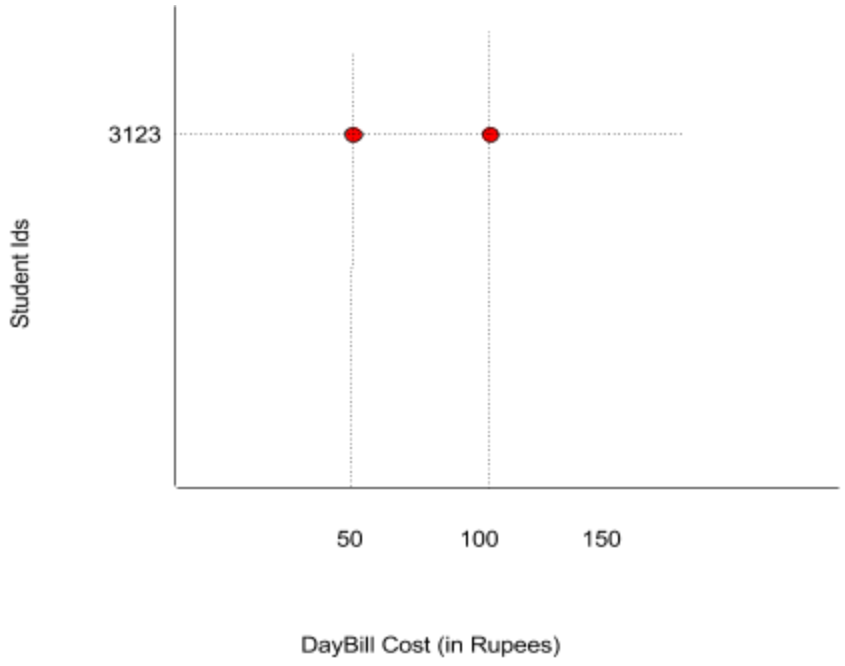
We propose to analyse the trends, across the years, of total ANC purchases by First Degree students. In other words, we propose to study how the total ANC purchase for a student increases as he/she progresses from the first to the fourth/fifth years.

Solution:

Our solution to the problem primarily employs clustering. The model is as follows:

As done in problem - 2, we obtain bills from the sales files, i.e. the contents of the slip obtained from the ANC counter for each transaction. For this problem, we go further and combine all transactions made by a student in a day into a single entity. For instance, if student A gets 2 slips: In the first one, A gets a maggi and a coke and in the second one, he/she gets a pasta. Then, we combine these two bills together and call the resulting entity a DayBill. Each DayBill has a cost which is the sum of prices of all items contained in it.

We have student Id numbers in the form “F1234”. In this problem, we are analysing only first-degree trends, hence our Id numbers essentially get reduced to 4-digit numbers in which the most significant digit represents the year. We plot these four digit numbers, each number representing a student, on the y-axis. On the x-axis, we plot prices, as depicted by the example in the figure. If a third year student eats for 2 days in ANC, worth rupees 50 and 100 respectively, then this leads to two points - (50, 3123) and (100, 3123).



We intend to perform clustering on data preprocessed in this manner to obtain trends in daily ANC spending as students progress from first to fourth/fifth years.

Through this analysis, we expect to obtain the following clusters:

1. A cluster at the bottom left, which would represent first year students and lower DayBills, since first year students frequent ANC less than students of higher years and are relatively controlled in their spending even when they do.
2. A large cluster in the center, corresponding to 2nd and 3rd year students and moderate-high DayBills.
3. A cluster in the top right corresponding to fourth and fifth years with DayBills driven up by treats and relatively unconstrained spending post placements season.

These are the trends we predict we would observe, and the results obtained after analysis as described below might be different.

Performance :

- Adjusted Rand Index - Given the knowledge of the ground truth class assignments and the clustering algorithm assignments, the adjusted rand index is a function that measures the similarity of the two assignments, ignoring permutations and with chance normalization. It has a bounded range of [-1,1] - negative values are bad (independent labelings), similar clusterings have a positive ARI with 1.0 as the perfect match score. It is better than raw Rand index as random or uniform label assignments have a ARI close to 0.0 and no assumption is made with regard to the clustering algorithm.
- V-measure : There are two objectives for any clustering assignment - Homogeneity and Completeness. Homogeneity requires each cluster to contain members of a single class only while Completeness requires all members of a given class to be assigned to the same cluster. The harmonic mean of these 2 measures is taken as the V-measure of the clustering assignment.
  No assumption is made with regard to the cluster structure. Moreover, V-measure has a very intuitive interpretation - the clustering can be qualitatively analyzed in terms of homogeneity and completeness.

One major requirement for both these measures is the presence of ground truth class assignments, which is not an issue here.

$$\mathrm{ARI} = \frac{\mathrm{RI} - E[\mathrm{RI}]}{\max(\mathrm{RI}) - E[\mathrm{RI}]}$$

$$\mathrm{RI} = \frac{a + b}{C_2^{n_{samples}}}$$

a = number of pairs of elements that have the same class and are in the same cluster

b = number of pairs of elements that have different class values and are present in different clusters

$C_2^{n_{samples}}$ = total number of possible pairs in the dataset without ordering

$E[RI]$ = Expected value of RI

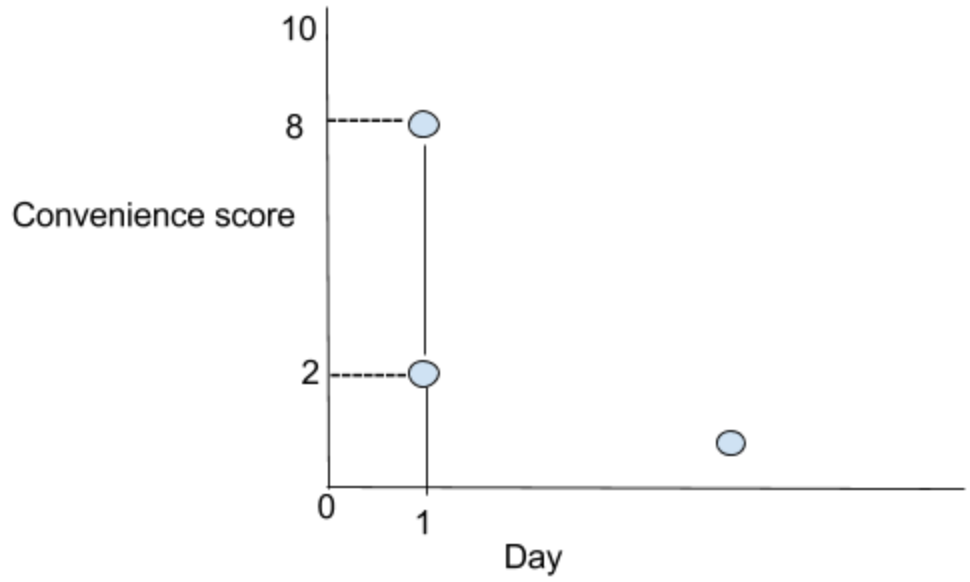
Question - 2

Find out trends in purchase of items on the basis of convenience of consumption across different parts of the semester.

Solution

First, a *convenience score* has to be assigned to each food item indicating the degree of convenience associated with the consumption of that item. For instance, an item like Chicken Nuggets or Momos has a higher degree of convenience in consumption as compared to an item like Paneer-Naan. The convenience score can be assigned to the items by taking feedback from the students and asking them to rate the item from 1 to 10 with 1 being the least convenient.

For this problem too, we work on bills. We get a bill convenience score by averaging the convenience of all the items present in a bill. On x-axis, we plot the day and on the y-axis, we plot the convenience score of each bill corresponding to that day.



From the given graph, it can be seen that on Day 1, 2 bills have been plotted - one having average convenience score 2 while the other has average convenience score 8.

Density-based clustering can be employed on this data to examine the purchase trends of items distinguished by their convenience scores in different parts of the semester. The resultant clusters can enable a third person to identify those days when people are busy and prefer to purchase items that can be easily consumed even while walking (examination times when most of the students are busy preparing for exams). There would be a different cluster which can have points for which the convenience score is very low corresponding to those times when the students are relatively free and tend to spend a lot of time in ANC, consuming mainly items which can not be consumed while commuting such as Butter chicken and naan.
The cluster corresponding to the comprehensive examinations will be having a lot of high convenience value purchases with students mainly buying items which can be easily consumed while commuting from ANC to their hostel in order to save time (and their semester).
Hence, after applying clustering, we can expect to observe clusters corresponding to those parts of the semester when students are busy with academics (Midsem time during October, Compre time for the whole of December) and those parts of the semester when the academic burden is less and students tend to while away a lot of their time in ANC (particularly during treats) purchasing items that have less convenience score.

The performance metrics specified in the first problem can be used for evaluating the clusters formed. In addition, another metric can be applied:-
Silhouette Coefficient - A higher Silhouette coefficient relates to a model with better defined clusters. It is composed of two scores:-
(i) Mean distance between sample and all other points in the same class **(a)**
(ii) Mean distance between sample and all other points in the next nearest cluster **(b)**
The Silhouette Coefficient is given by:-

$$s = \frac{b - a}{max(a, b)}$$