Lesson 10

Clustering

Clustering of a collection

- A process (method) of grouping a collection of objects into subsets, called clusters, according to their distinct characteristics in the group
- Forms one or more clusters, such that objects within one cluster similar to each other while the objects belonging to different clusters are dissimilar .

Clustering

- Discovers a large number of close-by points which form a distinct set in a collection
- How much large and how much close that depends on the chosen criterion function.

Clustering

- Refers to a segmentation of a population (sets of data) into a number of subgroups (subsets) using unsupervised techniques of data mining
- Sets of data examples is set of Computer Courses $\mathscr{C}$, set of semesters $\mathscr{S}$, set of students datasets, $\mathscr{D}$

Clustering Example

- Consider plot of the P_GPAs and T_GPAs, GPAs in practical subjects as independent variable along the x axis and the GPAs in theory subjects as dependent variable along the y axis.

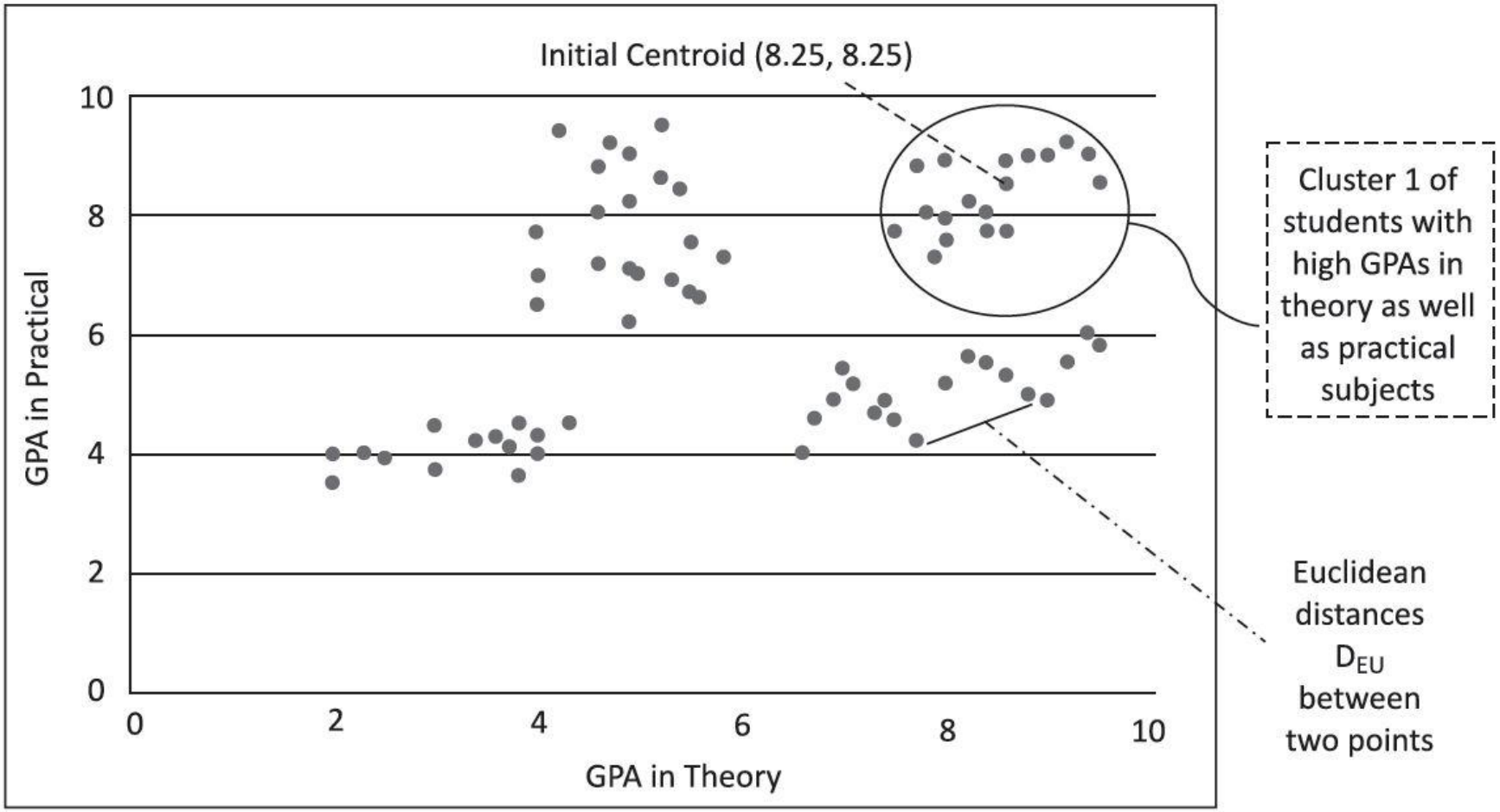
Clustering Example

- Figure 6.9 shows cluster 1 and the results of cluster analysis of students with criterion 1

Figure 6.9

- Shows a cluster consists of students with high GPAs in theory $\mathcal{T}$ as well as practical $\mathcal{P}$ subjects in a number of university courses $\mathcal{C}$ in a semester $\mathcal{S}$
- Each dot in Figure 6.9 corresponds to a distinct enrolled student in university computer courses (set $\mathcal{C}$) for semester examination (set $\mathcal{S}$).

Figure 6.9 cluster 1 and the results of cluster analysis of students with criterion 1



Centroid (s)

- Figure 6.9 (Example 6.11) shows a centroid (a central point of a cluster) GPAs of practical subjects (P_GPAs) and GPAs (T_GPAs) for data points of in subset S of set of students C.

Centroid(s)

- Example 6.11 considers centroid as point for the student sub-groups where means of both the GPAs (P_GPAs) and GPAs (T_GPAs) are high (near 8.0).

Input Vectors

- Column vectors of each data point in the figure have elements in the metric and non metric space
- Example 6.11, the elements of column vector for each student data-point are Y (Year), CC (Course-code), ID (Student-ID), SC (Semester-code), P_T [subject type (practical or theory), SubjID (Subject code) and GP (grade point).

Output Vectors

- Column vectors T_GPA (Grade point average of theory subjects), P_GPA (Grade point average of practical subjects) are the output column vectors (responses) in metric space for input column vectors

Target Variables

- Cluster centroid and criterion for inclusion in a cluster
- Multiple cluster case, there will be multiple centroids and corresponding criteria for each one.

Unsupervised Learning

- Refers to a process in which an ML algorithm does not use known target variable-outputs for the selected inputs for taking decisions or making predictions or finding the target (Cluster data-points)

Unsupervised Learning

- Cluster computations use input vectors, and the closeness and distinctiveness criteria only
- Does not use a training dataset consisting of output target vectors for selected input vectors, which also means (response variables for explanatory variables)

Clustering Criterion

- A similarity criterion defines the condition of inclusion of data points from a set into the cluster
- For example, The similarity criterion function is that GPAs in theory and practical (GPA_T, GPA_P), both are high for a subset of students in Courses $\mathcal{C}$, semesters $\mathcal{S}$ in the students' datasets, $\mathcal{D}$

Clustering Criterion

- Assume that (GPA_T, GPA_P) are within = (8.25 ± 1.75, 8.25± 1.75)
- The number criterion function can be that when 8% and above students in the set $\mathcal{C}$, then a cluster of high GPA_T with high GPA_P cluster 1 exists.

Clustering

1. An algorithm first finds the distances from a centroid in v-dimensional space
2. Does similarity (closeness) and dissimilarity (distinctiveness) analysis using a criterion, for example, criterion that points within = (8.25 ± 1.75, 8.25± 1.75) are similar and outside dissimilar

Clustering

3. Stopping condition, for example, students ☺ beyond the data points not included between (GPA_T, GPA_P) = (10.0, 6.5) and (6.5, 10.0), which means beyond the radius of 1.75 with respect to Centroid (8.25, 8.25) [Refer circle in the Figure] and Clustered data-points > Threshold % stop further iterations

Partitions in Two Dimensional Space

- Figure showed that data-points of the students can be partitioned into four regions in two dimensional space

Multiple Clustering

- Example 6.11 considers cluster 1 of High GPA_T and high GPA_P
- Consider other three options be as follows:
- Cluster 2— GPA in practical subjects between 4.5 and 6.0 and theory GPA between 6.5 and 10.0

Multiple Clustering

Cluster 3 ─ GPA practical subjects between 6.5 and 10.0 and theory GPA between 4.5 and 6.0

Cluster 4─── GPA practical subjects below 4.5 and theory GPA below 4.5.

Partitions/centroid based Clustering Algorithms

- K-means, K-medoids, Fuzzy k-means, Mean-shift clustering and other related methods K-Means Clustering

Clustering in v-dimensional space with metrics and non-metrics

- Examples of Feature Variables
- Apple feature: red, pink, maroon, yellowish, yellowish green and green
- Generally represented by text characters.
- Numbers also represent features: red with 1, orange with 2, yellow with 3, …. .

Recall Categorical variable

- A variable representing a category, a unit of observation to a particular group . For example:
- Car, tractor and truck belong to the same category, i.e., a four-wheeler automobile.
- Generally represented by text characters
- Set or Group with limited, and/or fixed number of possible values, features, …

Partitions in v-dimensional space

- Data-points can be in general metric as well non-metric (feature, category,..)
- For example, m-dimensions of metric and n dimensions of non-metrics, total v-dimensional space, $v = m + n$
- Partitioning of v-dimensional space and define accordingly the centroids and criteria

Data Points Example in v-dimensional Space

- Assume m = 2 and n = 4, m +n = v = 6: [Category, colour, shape feature, size feature, size (cm) and weight (gm.)]
- Fruit Category Space Variable:

(Apple, Orange, Pear, Pomegranate)

Data Points Example in v-dimensional Space

- Assume m = 2 and n = 4, m +n = v = 6: [Category, colour, shape feature, size feature, size (cm) and weight (gm.)]
- Fruit Category Space Variable:

(Apple, Orange, Pear, Pomegranate)

Input Vectors Example in v-dimensional Space (v=4)

- $\mathbf{I1}_{apple}$ = [Apple, red, round, 7.8, 205]
- $\mathbf{I2}_{apple}$ = [Apple, red, round, 6.2, 127],
- $\mathbf{I3}_{apple}$ = [Apple, yellow, Oblong, 5.1, 92],
- $\mathbf{I1}_{orange}$ = [Orange, yellow, oblate, 78] …

Feature space Variables

- Apple colours: red, pink, maroon, yellowish, yellowish green and green, tri-coloured, ….
- Shape feature: round, conic, oblate, oblique, oblong, and ovate

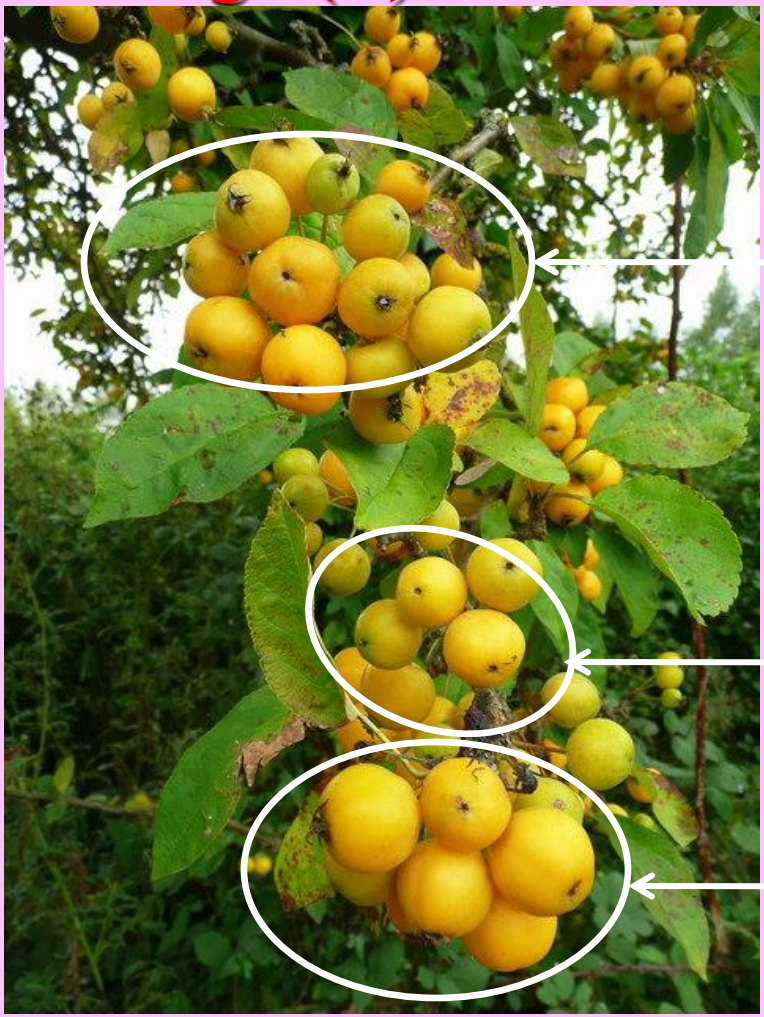
Size Feature Variable

- Size feature: above 7.0 cm (medium), above 8.0 cm (large), below 6 cm (small), or small (S), medium (M) and large (L) size features …

Metric Space Variables

- Size (cm)
- Weight (gm)

Three Clusters of Apples in small (S), medium (M) and large (L) size features



Summary

We learnt:

- Clustering
- Partitions, Centroids and Criterion
- Input Vectors
- Unsupervised Learning
- Clustering in v-dimensional space with metrics and non-metrics variables

End of Lesson 10 on

Clustering