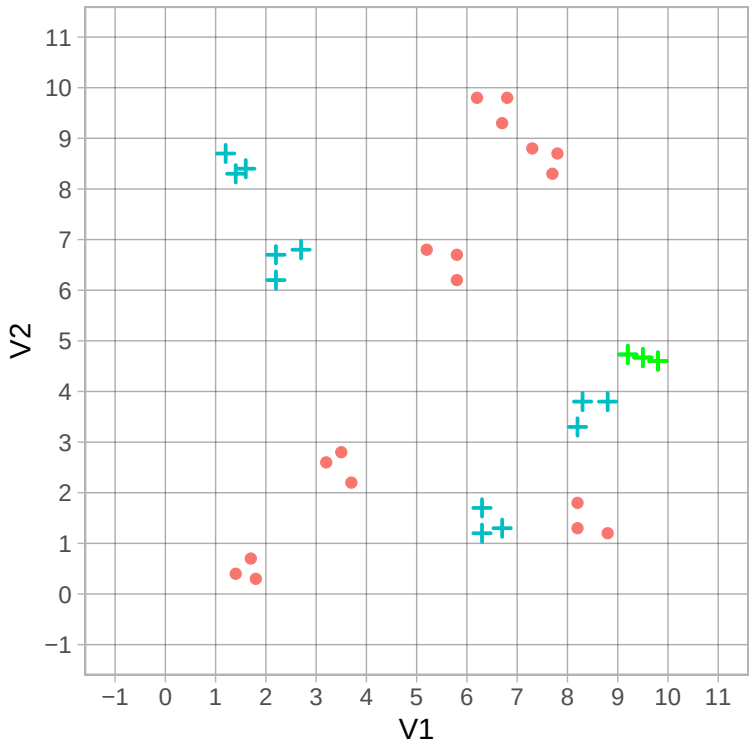
CS3440 Exam - Reference answers

Felipe Campelo

```
#> CANDIDATE ID: 9 8 7 6 5 4
#>               A B C D E F
```

Question 1

```
#> New points
#>   ID  V1   V2    Class
#> 1  K 9.2 4.73 positive
#> 2  K 9.5 4.67 positive
#> 3  K 9.8 4.60 positive
#> Updated plot (new points highlighted in green)
```



Question 1-a

- **[2 marks]** if student added new points to both table and plot (1 mark for each).

Question 1-b

The student is expected to split the data into approximately 2/3 for training, 1/3 for testing. Distribute points according to:

- **[1 mark]** if student split data into two sets.
- **[3 marks]** if student ensured both sets have both classes represented (1 for correct split, 2 for clear explanation)
- **[3 marks]** if student split the classes by the grouping variable ID. This will result in a training set with 24 observations (8 IDs) and a test set with 9. (1 for correct split, 2 for clear explanation).
- **[1 mark]** for correctly indicating test points on the table.

Question 1-c

The student is expected to build a decision tree based only on the *training set*. Distribute marks according to:

- **[3 marks]** for a clear explanation of the rationale behind the choices at each split (only 1 mark if student used full data instead of only training).
- **[1 mark]** for student only using integer thresholds for the splits.
- **[2 marks]** for a tree of *maximum* depth 3, i.e., maximum 3 splits from root to leaf. (zero if student tries to force precisely 3 levels by somehow splitting a pure node)
- **[4 marks]** for a correct calculation of relevant entropies and information gain at each split. (1.5 for correct methodology, 1.5 for correct results. Proportional marks if some calculations are incorrect).
- **[5 marks]** for correctly drawn DT highlighting:
  - the split criteria used at each non-leaf node (e.g., "V1 > 3") **[1 mark]**
  - the Information Gain of each split (e.g., "IG = 0.23") **[2 marks]**
  - the proportion of training examples of each class in the leaf nodes (e.g., "(6+/0-)") **[2 marks]**.

Question 1-d

The student is expected to calculate performance based only on the *test set*. Distribute points according to:

- **[1 marks]** Student uses only the test set for performance calculation.
- **[2 marks]** Student draws **correct** confusion matrix.
- **[2 marks]** Student correctly calculates F1 score (1 mark for correct formula, 1 for correct result).

Question 1-e

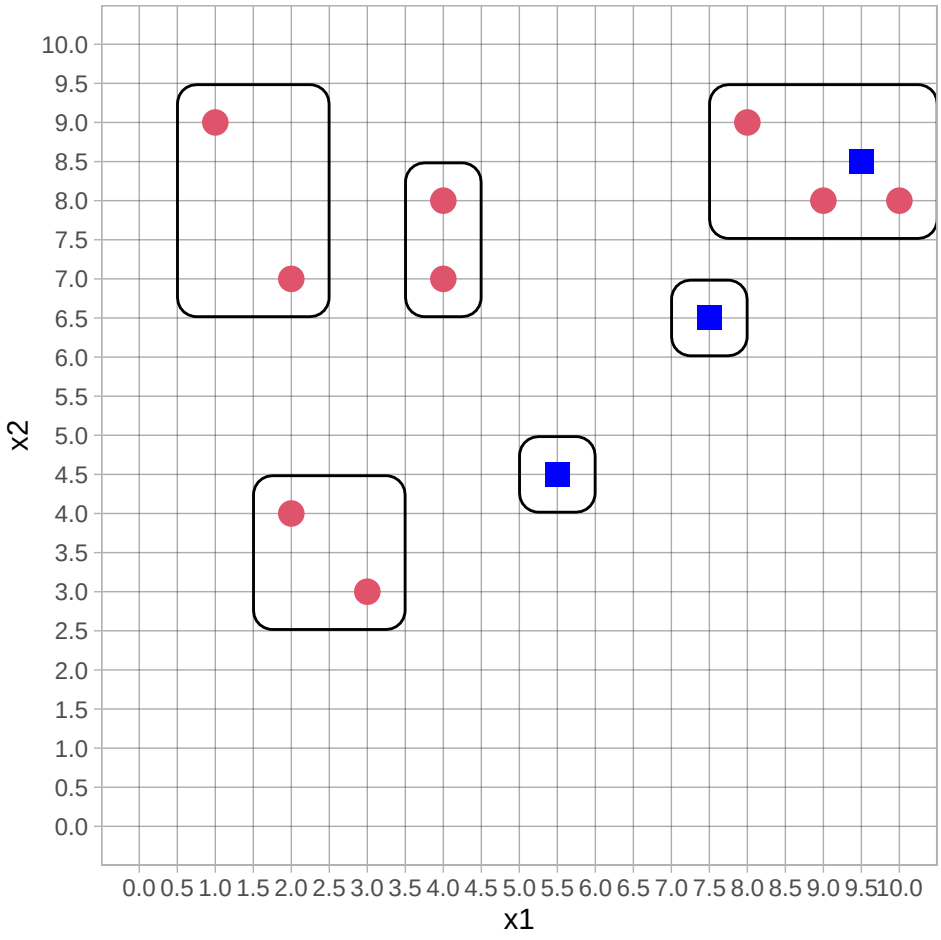
[5 marks] The student is expected to explain two performance indices, their mathematical definition and what they quantify. 2.5 marks for each correctly defined/discussed metric.

Question 2

```
#> New points
```

```
#>   ID  x1  x2
#> 1  J 9.5 8.5
#> 2  K 7.5 6.5
#> 3  L 5.5 4.5
```

```
#> Updated plot (new points highlighted in blue, groups indicate solution of item 2c)
```



Question 2-a

- **[2 marks]** if student added new points to both table and plot (1 mark for each).

Question 2b

The student is expected to detail the calculations related to each merge as she/he iteratively merges the points/clusters. Attribute marks as follows:

- **[5 marks]** if the student correctly calculates dissimilarities to their points using the Manhattan/L1 distance. (only 3 marks if calculations are incorrect, or if another distance is used)

```
#> Updated dissimilarity matrix (only student-filled rows shown)
```

	A	B	C	D	E	F	G	H	I	J	K	L
J	12	12	9	9	6	7	2	1	1	0	X	X
K	8	8	9	6	5	4	3	4	3	4	0	X
L	4	4	9	6	5	4	7	8	7	8	4	0

- **[10 marks]** if the student correctly merges the points / clusters using complete (maximum) linkage, and clearly indicate their steps. (deduct 30% of marks earned if calculations are incorrect. Deduct 30% of marks earned if another linkage is used. Zero marks if answer is provided without explicitly showing calculations or justifications.)

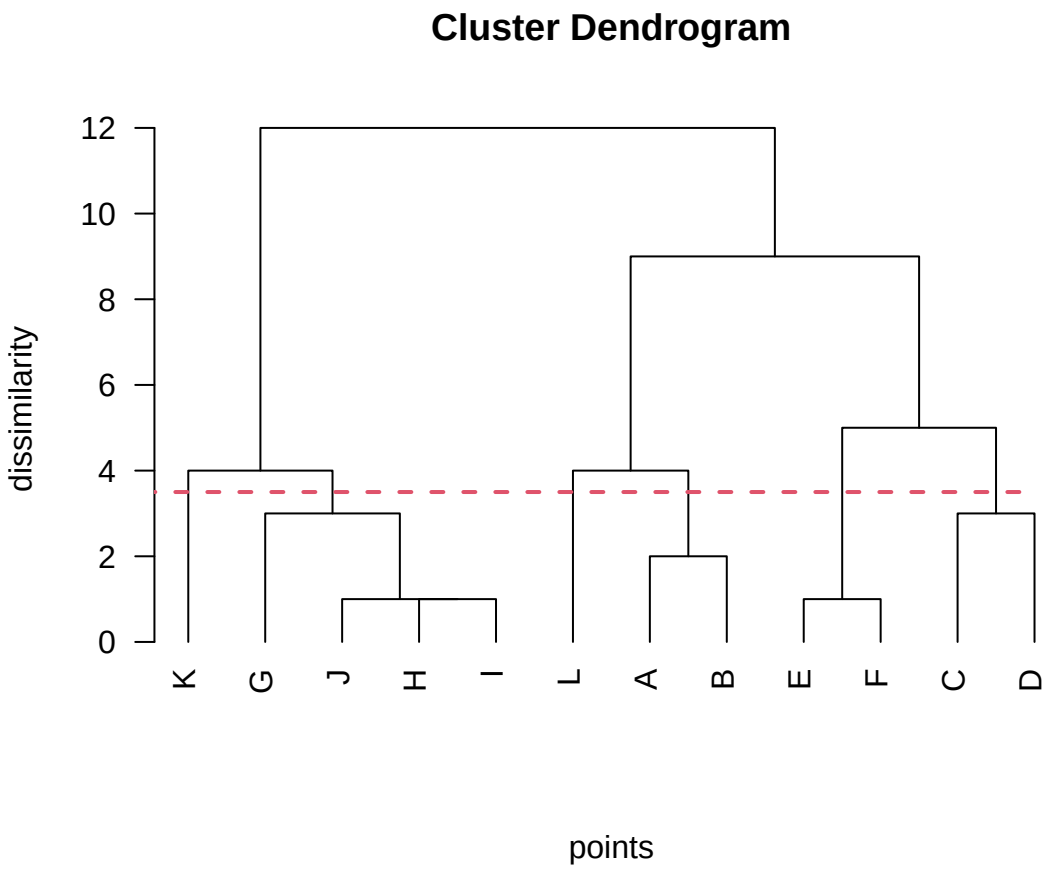
As an example, the explanation of merging steps should look like this:

- *Merge 1: smallest distance (dist = 3) is between points F, K. Merged into cluster {F,K}.*
- *Merge 2: smallest distances (dist = 4) are between point G and cluster {F,K} and between points B and C. Selected merged the first into cluster {F,K,G}*
- *Merge 3: . . .*

. . . until all points are joined into a single cluster.

- **[5 marks]** for a correctly built dendrogram of the resulting clustering, including clearly labeled points on the x-axis and clearly indicated levels in the y-axis (only 2.5 for partially correct or unlabeled labeled axes). Note that the dendrogram may vary slightly due to student decisions when there are ties.

```
#> Resulting dendrogram (red dashed line is part of 2c)
```



Question 2c

[3 marks] The student is expected to add a cut line to the dendrogram produces in the previous item (see red dashed line in the figure above) and indicate the clusters in the original figure (see figure at the start of this section for the ellipses indicating the clusters). (1.5 mark for correct level on dendrogram, 1.5 for correct ellipses in the plot).

Question 2d

The student is expected to show her/his work on two iterations of k-means starting from given centroids.

- [**1 mark**] for correct attribution of points and centroids
- [**2 marks**] for using correct method
- [**1 mark**] for correct first iteration
- [**1 mark**] for correct second iteration
- [**1 mark**] for explicitly stating the final clusters

```
#> Points:

#> [1]  9  8  7  6 12 23 29 37

#> Initial centroids:

#> [1]  9 19

#> First iteration:

#> Distance from each point to Centroids:
```

Point	9	8	7	6	12	23	29	37
Dist C1	0	1	2	3	3	14	20	28
Dist C2	10	11	12	13	7	4	10	18

```
#> Centroid attribution:

#> Pts:  9  8   7   6   12  23  29  37

#> Cen:  1  1   1   1   1   2   2   2

#> New centroids (mean of corresponging points):

#> [1]  8.4 29.7

#> Second iteration:

#> Distance from each point to Centroids:
```

Point	9.0	8.0	7.0	6.0	12.0	23.0	29.0	37.0
Dist C1	0.6	0.4	1.4	2.4	3.6	14.6	20.6	28.6
Dist C2	20.7	21.7	22.7	23.7	17.7	6.7	0.7	7.3

```
#> Centroid attribution:

#> Pts:  9  8   7   6   12  23  29  37

#> Cen:  1  1   1   1   1   2   2   2

#> New centroids (mean of corresponding points):

#> [1]  8.4 29.7

#> Final clusters

#> Cluster 1:  9   8   7   6   12

#> Cluster 2:  23   29  37
```

Question 2e

The student is expected to list and **discuss** at least two of each list below.

- Advantages:
  - Relatively simple to implement.
  - Scales to large data sets.
  - Guarantees convergence.
  - Can warm-start the positions of centroids.
  - Easily adapts to new examples.
  - Generalizes to clusters of different shapes and sizes, such as elliptical clusters.
- Disadvantages / limitations:
  - Need to choose k
  - Being dependent on initial positions of centroids
  - Has trouble clustering data where clusters are of varying sizes and density.
  - Centroids can be dragged by outliers, or outliers might get their own cluster instead of being ignored.
  - Distance-based similarity measures have problems in very high-dimensions (particularly if feature space is sparse)

[**4 marks**] for correct definitions and explanations (1 for each advantage/disadvantage correctly defined/explained. Zero marks if answer is provided without any explanations.)